Preamble

0.1 Goals

Example 1

1.1 Healing of surgical wounds

The data shown in Table 1.1 were obtained in an experiment by Dr. George Huang of the Department of Surgery at the University of Chicago, the purpose of which was to investigate the effect of hyperbaric O_2 treatment on the healing of surgical wounds in diabetic rats. (Diabetics, both human and animal, tend to have more complications following surgery than non-diabetics, and these rats made the ultimate murine sacrifice by serving as the surgical model for diabetic effects in humans.) Thirty rats were first given a drug that has the effect of destroying the pancreas, with the goal of making the rats diabetic. All the rats underwent surgery, during which an incision was made along the entire length of the back. This was immediately sewn up with surgical staples, and the rats were returned to their cages.

The treatment group of fifteen rats was subjected to hyperbaric O_2 treatment, i.e., a 100% O_2 environment at two atmospheres pressure, for ninety minutes per day following surgery. The control group also received O_2 treatment for 90 minutes daily, but at normal atmospheric pressure. Six rats had glucose levels that were deemed too low to be considered diabetic, and were excluded from the experiment. (You may assume initially that these exclusions are unrelated to the O_2 treatment.) After a 24 day recuperation period, the 24 rats still participating in the experiment were sacrificed, i.e., killed. Strips of skin were taken from five sites labelled A–E on each rat, each site crossing the surgical scar in a right angle. The strips were put on a tensiometer, stretched to the breaking point, and the energy required to break the specimen was recorded. Unfortunately some specimens slipped out of the clamps for reasons unconnected with the strength of the specimen, and in such cases no observation could be made: the unmeasured specimens are indicated by `--` in the table. Rats 1–14 received the hyperbaric treatment: rats 15–24 were the controls.

Handling by humans is known to be stressful for rats, and stress is associated with poor health and shorter lifetimes. The experiment was designed to ensure that treated and control rats were handled in a similar manner throughout the experiment, so that any observed differences between the groups could confidently be attributed to treatment rather than to differences in the way

Table 1.1: Strength of skin specimens 24 days after surgery

	Site on the back: anterior to caudal					
Rat	A	B	C	D	E	Mean
1†	3.8300	7.3788	44.353	19.555	--	18.779
2	27.861	29.974	15.470	23.455	--	24.190
3	56.996	60.960	20.306	--	28.123	41.596
4	--	38.043	68.080	42.425	30.335	44.721
5	16.276	--	59.033	73.891	--	49.733
6	38.267	33.702	35.558	44.598	32.678	39.961
7	9.0384	11.259	27.121	31.984	--	19.851
8	16.728	27.590	13.238	12.139	6.3865	15.216
9	11.866	27.983	26.226	15.594	19.225	20.179
10	23.352	34.790	27.556	35.883	22.848	28.888
11	16.444	31.928	21.495	15.590	7.0750	18.506
12	23.342	46.313	33.810	15.686	--	29.788
13	15.267	14.452	10.635	22.156	6.8062	13.863
14	21.732	20.746	12.293	17.295	10.301	16.473
15†	82.508	13.645	49.187	--	53.432	49.693
16	--	45.919	63.090	68.137	36.500	53.412
17	80.147	29.943	71.928	--	46.609	57.157
18	31.938	--	36.211	49.815	44.468	40.608
19	15.453	31.384	27.127	27.961	9.9035	22.366
20	21.183	27.429	20.058	--	--	22.890
21	20.445	12.532	15.661	28.694	--	19.333
22	16.928	59.579	29.407	18.626	8.8352	26.675
23	35.631	21.613	23.155	42.379	16.203	27.796
24	20.523	24.621	16.292	--	18.680	20.029
Mean	27.534	29.627	31.970	31.888	23.436	29.126

†Rats 1–14 are hyperbaric O_2-treated; 15–24 are the controls.

that the rats were handled. Hence, the control rats were inserted daily into the hyperbaric chamber so that they might experience the same stress levels as the treated rats.

The main objective is to determine whether or not hyperbaric O_2 treatment has an effect on the healing of surgical wounds, and if so, whether the effect depends on the position of the wound on the back. It was anticipated that the oxygen effect would be beneficial, or at least not detrimental, for healing, and that the site effects would be negligible. Confidence intervals or posterior distributions for the size of any such effects are a part of the answer.

1.2 An elementary analysis

It is unclear whether in fact the treatment was assigned to the rats by objective randomization, but it is reasonable to proceed as if this were the case. In principle, this is a completely randomized design with no blocking. Each rat-site pair is one observational unit, and each rat is one experimental unit, so each experimental unit consists of five observational units. The distinction between observational units and experimental units is crucial in the design, in the analysis and in the interpretation of results.

If there were no missing values, the analysis would be relatively straightforward, so we first illustrate the reasoning behind the simpler analysis. First, the observations are strictly positive strength measurements having a moderately large dynamic range from 3.8 to 82.5, so a log transformation is more or less automatic. Normality of residuals is important, but it is not nearly so important as additivity assumptions that are made in a typical linear model—in this case additivity of rat effects, site effects and treatment effects. So it is arguably misleading in most instances to point to either marginal histograms or residual plots as the principal reason for transformation.

To each experimental unit there corresponds an average response, giving 14 values for O_2-treated rats, and ten values for control rats. In the absence of missing components, the site effects contribute equally to rat averages, so the averages are not contaminated by additive differences that may be present among sites. The sample means for control and treated rats are 3.372 and 3.113 on the log scale, the sample variances are 0.174 and 0.200 respectively, and the pooled variance is

$$\frac{9 \times 0.174 + 13 \times 0.200}{22} = 0.189$$

on 22 degrees of freedom. This analysis, which is based on the rat averages, leads to an estimated treatment effect

$$\text{average for treated rats} - \text{average for control rats} = -0.259$$

with standard error $\sqrt{0.189(1/10 + 1/14)} = 0.180$. The estimated effect is only 1.4 standard deviations from the null value of zero, and the deviation is in the direction not anticipated. The conclusion from this analysis is that there is no evidence of a treatment effect—positive or negative.

This elementary arithmetic analysis is standard for a design that is complete with no missing observational units. It is open to criticism in this setting because the comparison may be unfair if site effects are appreciable and the pattern of missing treated units is substantially different from the pattern for controls. For example, site D is missing for 40% of the control rats, but only for 7% of treated rats. If the response at site D were appreciably different from the other sites, the pattern of missing units would create a bias in the treatment comparison. However, the site averages on the log scale

$$3.093,\ 3.263,\ 3.317,\ 3.326,\ 2.928,$$

show little indication of appreciable differences or a strong trend, so the preceding analysis appears reasonably sound. Nonetheless, it is only natural to ask for a more definitive analysis taking into account the possibility of additive site effects.

1.3 Two incorrect analyses

One way to adjust for site effects is to fit a simple linear Gaussian model in which site and treatment effects are additive on the log scale:

$$E(Y_{is}) = \beta_0 + \beta_{t(i)} + \beta_s; \quad \text{var}(Y_{is}) = \sigma^2, \tag{1.1}$$

where s is the site, and $t(i)$ is the treatment indicator for rat i. The least-squares treatment-effect estimate is -0.298 with standard error 0.119, which is computed from the residual sum of squares of 34.30 on 98 degrees of freedom. According to this analysis, the treatment estimate is 2.5 standard errors away from its null value, a magnitude that is sufficient to make a case for publication in certain scientific journals, even if its direction is opposite to expected.

Although the error in this analysis may seem obvious, the glib partial description in (1.1) is extremely common in the scientific literature. Very often, the model is stated additively in the form

$$Y_{is} = \beta_0 + \beta_{t(i)} + \beta_s + \varepsilon_{is}.$$

Implicitly or explicitly, the errors ε_{is} are assumed to be independent Gaussian with constant variance. Failure to account for correlations between different observations on the same experimental unit has little effect on the point estimate of the treatment effect, but it has a more substantial effect on the variance estimate.

It is good to bear in mind that there cannot be more degrees of freedom for the estimation of treatment contrasts than there are experimental units in the design. This design has 24 experimental units split into two subsets, so there cannot be more than 22 degrees of freedom for the estimation of inter-unit experimental variability. Thus, failure to mention covariances in the linear model specification, and the claim of 98 degrees of freedom are two red-flag indicators of gross statistical transgressions.

One way to adjust for rat-to-rat variability is to include an additive rat effect:

$$E(Y_{is}) = \beta_0 + \beta_{t(i)} + \beta_s + \gamma_i; \quad \text{var}(Y_{is}) = \sigma^2.$$

For example, this model can be fitted in R using the commands

```
fit <- lm(log(y)~site+treat+rat);  anova(fit)
```

The ANOVA function reports a very substantial F-ratio of 10.45 for treatment, with a p-value of 0.2%. However, the treatment effect estimate is only -0.600 ± 0.33, and the p-value is a more modest 7%. We will not attempt here to explain this apparent contradiction because the displayed code points to a serious lack of understanding of linear algebra, geometry and orthogonal projections. Neither part of the code or the computation is appropriate, and the fitted model is not suited for its intended purpose.

1.4 Model formulae

In discussions concerning least-squares coefficients and statistical model formulae, it is good to remember that each term in a linear-model formula is first and foremost a vector subspace of $\mathbb{R}^n$, where n is the number or set of observational units. In this context, the operator `+` denotes the span of subspaces, not a vector sum.

Each subspace associated with a factor has a natural basis consisting of one indicator vector for each factor level. However, certain statistical questions are concerned with the subspace, in which case statistical conclusions are, or should be, unaffected by the choice of basis: see Exercise 1.8. For example, `site`, `treat` and `rat` are subspaces of dimensions 5, 2 and 24 respectively, which is the number of levels of the factor that occur in the design. Every factor subspace includes the one-dimensional subspace $\mathbf{1}$ of constant vectors, which is the sum of the indicator vectors. In most situations, the intersection of a pair of factor subspaces such as `site` and `rat`, or `site` and `treat`, is precisely this one-dimensional subspace. However, the fact that treatment is assigned to rats means that `treat` is a subspace of `rat`, so `site+treat+rat = site+rat`. Treatment effects are said to be confounded with rat effects.

Numerical linear-algebra algorithms detect this confounding, and they resolve it by by picking the most convenient subset of the basis vectors on offer. This subset is invariably rather arbitrary, which explains part of the problem in the paragraph at the end of the preceding section. As a result, the numbers reported there are statistically uninteresting, and they are potentially misleading if the algebraic issues are not fully understood.

Each subspace associated with a *block factor* such as `rat` also has a natural indicator basis. Since each rat is one experimental unit, it is implicit that the associated effects are not entirely arbitrary, but are judged a priori to be statistically exchangeable. In a sense, randomization guarantees exchangeability of effects. The effects referred to in this setting are the coefficients of the basis

vectors, and specifically the coefficients of the indicator basis for the experimental units, so the indicator basis for the subspace `rat` is not on an equal footing with any other basis. The exchangeability argument does not apply with equal force to a classification factor such as `site`.

For the most part, these remarks are unaffected by replication or by the pattern of missing components in the design.

1.5 A more appropriate formal analysis

The default Gaussian model for the log-transformed measurements Y_{is} incorporates site and treatment effects additively as follows:

$$E(Y_{is}) = \beta_s + \beta_{t(i)}; \qquad \mathrm{cov}(Y_{is}, Y_{jt}) = \sigma_0^2 \delta_{ij} \delta_{st} + \sigma_1^2 \delta_{ij}, \tag{1.2}$$

where δ_{ij} is the Kronecker symbol for equality of subscripts. Computationally speaking, `treat` and `site` are two classification factors, which determine subspaces of dimensions two and five respectively, whereas `rat` is encoded as a block factor or symmetric indicator matrix $rat(is, jt) = \delta_{ij}$ with coefficient σ_1^2. The overall covariance matrix in (1.2) is invariant with respect to permutation of rats and permutation of sites, but, unlike (1.1), it is not invariant with respect to arbitrary permutation of observational units.

Expression (1.2) is equivalent to the vector statement $Y \sim N_n(\mu, \Sigma)$ in which μ belongs to the subspace `site+treat`, and Σ belongs to the convex cone spanned by the identity and `rat` as a block factor. The equivalent expression in terms of additive effects and random variables is

$$Y_{is} = \beta_s + \beta_{t(i)} + (\sigma_1 \epsilon_i + \sigma_0 \epsilon_{is})$$

in which all effects contributing to variances and covariances are in parentheses. Independence of components is not to be taken for granted, so is necessary to state explicitly that the rat effects $\epsilon_1, \ldots, \epsilon_{24}$ are independent and identically distributed, and are independent of the 120 standard Gaussian residual effects ϵ_{is}, which are also mutually independent.

All told, there are five site parameters, one treatment parameter, and two variance components whose estimates are $\hat{\sigma}_0^2 = 0.211$ and $\hat{\sigma}_1^2 = 0.148$. Observations on distinct rats are independent, but the covariance between observations at different sites on the same rat is σ_1^2, and the correlation is $\sigma_1^2/(\sigma_0^2 + \sigma_1^2)$, which is estimated as 0.41.

In standard software, the site and treatment parameters are estimated by weighted least squares using the inverse of the fitted covariance matrix as weights. The treatment effect estimate, which is automatically adjusted for additive site effects, is -0.294 with standard error 0.184. In the absence of missing values, the null distribution of the ratio is t_{22}, so the observed effect corresponds to a two-sided p-value of about 12%. The likelihood-ratio statistic of 2.51 on one degree of freedom gives an essentially identical conclusion. To be clear, this is the version recommended in section 17.3–17.4: see exercise 1.14.

The less recommended version using ordinary maximum likelihood is typically somewhat larger—2.62 in this instance.

The code shown in exercise 1.7 reports fitted anterior-to-caudal site effects

$$0.000, 0.158, 0.181, 0.260, -0.271.$$

The standard errors of pairwise site contrasts are 0.14–0.15, so it appears that skin from the caudal site is appreciably weaker than that from other sites. The REML log likelihood ratio statistic for site effects is 13.92, which is beyond the 99th percentile of the limiting null distribution, which is χ^2_4. Although they appear to be non-zero, the site effects are not sufficiently large to change appreciably the conclusions reached by the more elementary analysis based on rat averages.

It is mathematically possible that treatment could have an effect on either the mean or on the variance or on both, but the standard default formulation assumes that treatment affects only the mean of the distribution. Such an assumption can easily be checked by including two different variance components, one for treated rats and one for the controls. For these data, there is absolutely no evidence of an effect of treatment on variances.

1.6 Further issues for consideration

1.6.1 Exclusions

Six rats that were deemed non-diabetic on the basis of post-baseline glucose measurements were excluded from the main analysis. For the main goal of this study, this exclusion was judged to be scientifically reasonable on the basis of an argument that implies that the probability of exclusion is unrelated to treatment. However, the excluded rats consisted of five controls and only one treated rat. How extreme is that allocation relative to expectation? Does it suggest that treated rats are less likely to be excluded than the controls? If so, treatment may have an effect of an entirely different nature.

Given that six rats were excluded, the null distribution of the number of excluded controls is central hypergeometric

$$\binom{6}{y}\binom{24}{15-y} \Big/ \binom{30}{15};$$

the numerical values are $0.8, 7.6, 24.1, 34.9, 24.1, 7.6, 0.8$ in percentages for $y = 0, \ldots, 6$. The probability of an allocation at least as extreme as that observed is 8.4% in each tail. Exclusions and other departures from protocol must always be described and included as a part of the discussion. The imbalance in this study is greater than we might have wished for, but it is not sufficiently extreme to imply a systematic bias.

1.6.2 Missing components

What are the reasons for certain components to be missing?

1.6.3 Back-transformation

The analysis was done on the log scale. How should the conclusions be reported?

1.7 Exercises

1.1 To compute the control and treatment averages, the R command

```
tapply(log(y), treat, mean)
```

returns the pair of averages $3.360, 3.085$, which is not the pair reported in the text. The alternative log-scale computation

```
tapply(tapply(log(y), rat, mean), trt, mean)
tapply(tapply(log(y), rat, mean), trt, var)
```

returns the numbers $3.372, 3.113$ for means, and $0.174, 0.200$ for variances. Under what circumstances do the two mean calculations return the same pair of averages? Explain the difference between the factors `trt` and `treat`.

1.2 In the balanced case with no missing cells, the standard analysis first reduces the data to 24 rat averages $\bar{Y}_{i\bullet}$, the treatment and control averages $\bar{Y}_T, \bar{Y}_C$, and the overall average $\bar{Y}_{\bullet\bullet}$. The sum of squares for treatment effects is

$$\mathrm{SS}_T = 70\bar{Y}_T^2 + 50\bar{Y}_C^2 - 120\bar{Y}_{\bullet\bullet}^2 = (\bar{Y}_T - \bar{Y}_C)^2 \frac{5 \times 10 \times 14}{24}.$$

The total sum of squares for rats splits into two orthogonal parts

$$5\sum_i (\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet})^2 = \mathrm{SS}_T + \mathrm{SS}_R,$$

which are independent on one and 22 degrees of freedom respectively. If treatment effects are null, the mean squares $\mathrm{SS}_T/1$ and $\mathrm{SS}_R/22$ have the same expected value, and the mean-square ratio

$$F = \frac{\mathrm{SS}_T/1}{\mathrm{SS}_R/22}$$

is distributed as $F_{1,22}$. Simulate a complete design with additive effects, and check that the two terms shown above agree with parts of the decomposition reported by `anova(lm(y~site+treat+rat))`.

1.3 The F-ratio reported by `anova(...)` for treatment effects is not the ratio shown above. At least one is misleading for this design. Which one? Explain your reasoning.

1.4 For the model (1.2), verify that the R commands

```
reg_fit <- regress(log(y)~site+treat, ~rat)
lme_fit <- lmer(log(y)~site+treat+(1|rat))
```

return the same parameter estimates in a slightly different format.

1.5 The sub-model with zero rat variance can be fitted by omitting the relevant term from the model formula in either syntax. The conventional log likelihood ratio statistic is twice the increase in log likelihood for the larger model relative to the sub-model. Check that these functions report different numbers for the log likelihood, but they return the same log likelihood ratio. Report the value. (Recall that the log likelihood is defined up to an arbitrary additive constant, which may depend on the response or the design matrix.)

1.6 REML, or residual maximum likelihood, is the standard method for the estimation of variance components: see chapter 17 for details. Both `regress()` and `lmer()` allow other options, but both use REML as the default. However, `lmer()` constrains the coefficients to be positive, whereas `regress()` allows negative coefficients unless otherwise requested. If $\hat{\sigma}_1^2 > 0$, both functions should report the same values for all coefficients; otherwise if the unconstrained maximum occurs at a negative value, there will be differences both in the fitted variance components and in the regression coefficients.

For regular problems in which the null model is not a boundary subset, the null distribution of the conventional log likelihood ratio statistic is distributed asymptotically as χ_1^2. Assuming that the unconstrained version is regular with fitted coefficients approximately unbiased, what is the asymptotic distribution of the log likelihood ratio statistic for the constrained problem? Using this null distribution, report the tail p-value for the hypothesis of zero rat variance.

1.7 In the balanced case with no missing cells, show that the REML likelihood-ratio statistic for treatment effects is

$$\mathrm{LLR} = (n-1)\log\bigl(1 + F/(n-2)\bigr),$$

where $n = 24$ is the number of rats, and F is the treatment-to-rat mean-square ratio shown in exercise 1.2. Compute the F-value and the associated tail probability for LLR = 2.51 and LLR = 3.86. Comment briefly on the relevance of this calculation for the calibration of likelihood-ratio statistics in the present setting.

1.8 A quantitative factor `x` with four equally-spaced levels $0, 1, 2, 3$ may be coded using either the indicator basis e_0, e_1, e_2, e_3 (such that $e_r(i) = I(x_i = r)$) or the polynomial basis x^0, x^1, x^2, x^3 (with $x^0 = 1$, $x^1 = x$). Show that, if every level occurs with equal frequency in the design, the polynomials $1, z = 2x - 3, (z^2 - 5)/4, (5z^3 - 41z)/12$ are orthogonal with respect to the standard inner product in $\mathbb{R}^n$. Show that the components of the 4×4 transformation matrix that expresses this polynomial basis in terms of the indicator basis are all integers.

1.9 Use polynomials up to degree four to re-parameterize the site effects, and repeat the fitting procedure for (1.2) using the unnormalized orthogonal polynomial basis. Check that the treatment effect estimate and its standard error are unaffected by site re-parameterization. What effect does the change of basis have on the log likelihood?

1.10 How can we be assured that the log transformation is really needed or substantially beneficial? (Chapter 18).

1.11 Extend the model (1.2) so that it contains one variance component for treated rats and another for untreated rats. Show your code for fitting the extended model, report the two fitted variance components, and the REML likelihood ratio statistic for comparison with the simpler model.

1.12 Parameter estimates reported in section 1.5 were computed using the code in exercise 1.4. Following recommendations in section 17.5, the likelihood ratio statistic for treatment effects was computed using the code

```
K <- model.matrix(~site)
fit0 <- regress(log(y)~site, ~rat)
fit1 <- regress(log(y)~treat+site, ~rat, kernel=K)
llr <- 2*(fit1$llik - fit0$llik)
```

Modify this code to obtain the likelihood-ratio statistic for site effects.

1.13 It is mathematically possible that treatment could have a positive effect at some sites and a negative effect at other sites, so that the average over sites is negligible. Investigate this possibility by computing the appropriate likelihood ratio statistic.

1.14 Let Y be an $n \times m$ array of random variables with zero mean and covariance matrix

$$\mathrm{cov}(Y_{ir}, Y_{js}) = \sigma_0^2 \delta_{ij}\delta_{rs} + \sigma_1^2 \delta_{ij} + \sigma_2^2 \delta_{rs} + \sigma_3^2$$

for some non-negative coefficients $\sigma_0^2, \ldots, \sigma_3^2$. Show that the covariance matrix is invariant with respect to the product group consisting of $n!$ permutations applied to rows and $m!$ permutations applied to columns. In other words, show that the $n \times m$ matrix whose (i, s)-component is $Y_{\sigma(i),\tau(s)}$, has the same covariance matrix as Y.

1.15 For an $n \times m$ array, show that the four quadratic forms, $mn\bar{Y}_{\bullet\bullet}^2$,

$$\text{Row SS} : m \sum_i (\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet})^2;$$

$$\text{Col SS} : n \sum_r (\bar{Y}_{\bullet r} - \bar{Y}_{\bullet\bullet})^2;$$

$$\text{Resid SS} : \sum_{ir} (Y_{ir} - \bar{Y}_{i\bullet} - \bar{Y}_{\bullet r} + \bar{Y}_{\bullet\bullet})^2,$$

are invariant with respect to row and column permutations. Here, $Y_{i\bullet}$ is the ith row total, and $\bar{Y}_{i\bullet}$ is the row average.

1.16 Each of these quadratic forms is non-negative definite. In each case, the expected value is a non-negative linear combination of the four variance components, in which the coefficient of σ_0^2 is the rank of the quadratic form. Find the expected value of each quadratic form as a linear combination of the four variance components.

1.17 The set of linear functionals $\mathbb{R}^{nm} \to \mathbb{R}$ is called the dual vector space; it has dimension mn. Show that the column and row totals $Y \mapsto Y_{\bullet r}$ and $Y \mapsto Y_{i\bullet}$ are linear functionals, and that they are linearly independent. Show that the subspace spanned by $\{Y_{\bullet 1}, \ldots, Y_{\bullet m}\}$ is closed with respect to row and column permutations. What is its dimension? Show that the subspace spanned by $\bar{Y}_{\bullet\bullet}$, and the subspace spanned by $\{\bar{Y}_{\bullet 1} - \bar{Y}_{\bullet\bullet}, \ldots, \bar{Y}_{\bullet m} - \bar{Y}_{\bullet\bullet}\}$ are both closed with respect to row and column permutations. What are their dimensions?

1.18 The space of quadratic forms in the $n \times m$ array Y is a vector space of dimension $mn(mn-1)/2$. Exhibit a basis. Show that the four quadratic forms in Exercise 1.3 are invariant with respect to row and column permutations. Deduce that every invariant quadratic form is a linear combination of these four.

Example 2

2.1 Efficiency of chain saws

This example, taken from Bliss (1970, p. 440–441), is a description of an experiment by Zehnder, Weber and Linder (1951) which was designed to compare the performance of different brands of chain saw. The design is elegant and carefully controlled, but it is moderately complicated in structure, and it repays careful study.

The woodcutting efficiencies of three brands of saw were compared in a fractional factorial design using six cutting teams, three species of softwood (spruce, pine and larch) both with bark and without bark. The response variable is the time in minutes taken to complete a designated cutting task. The fractional factorial is embedded in a 6×6 Latin square whose columns correspond to six teams of workmen covering the range from experienced woodcutters to seasonal labourers. The letters correspond to six distinct saws, where A,D are duplicates of brand 1, B,E are duplicates of brand 2, and C,F are duplicates of brand 3. Table 2.1 shows the design and the response in standard readable format, which can be rearranged in spreadsheet format if needed.

To clarify matters for subsequent discussion, the design consists of 12 spruce logs, 12 pine logs, and 12 larch logs. All logs are presumed to be approximately equal in length and diameter, so the task demands a fixed number of cuts. Six spruce logs, six pine logs, and six larch logs were selected uniformly at random for de-barking. Each row of the table is one species/bark combination. The assignment of teams to logs is done uniformly at random subject to the condition that each team is required to cut one log of each species/bark combination. The assignment of saws to logs is done uniformly at random subject to the Latin-square condition that each team gets to use each saw exactly once, and each saw is used exactly once for each species/bark combination.

2.2 Covariate and treatment factors

Each observational unit is an ordered pair consisting of one log and one saw, so the units available at the outset may be arranged in a 36×6 array of (log, saw)-pairs. However, each measurement is destructive of the log, so it is necessary

		Team					
Species	Bark	I	II	III	IV	V	VI
spruce	no	6.4 F	10.9 E	9.8 D	7.5 B	4.6 A	4.9 C
pine	no	6.8 B	6.2 C	7.9 E	6.0 A	4.0 D	4.2 F
larch	no	12.7 E	13.4 A	12.5 B	7.3 C	6.1 F	7.4 D
spruce	yes	8.8 C	10.2 D	12.5 A	8.6 F	6.1 E	5.6 B
pine	yes	7.4 D	10.0 B	8.3 F	6.4 E	4.3 C	5.6 A
larch	yes	13.1 A	12.0 F	12.0 C	11.3 D	6.1 B	9.7 E

Table 2.1: Time in minutes taken by six teams to complete a woodcutting task using one of six available saws A–F. From Bliss (1970) with one correction in row 1, col 6.

to choose a subset or subsample of 36 observational units, one from each row. The Latin square design also calls for six units from each column or saw.

Despite the restrictions, the observational units are log-saw pairs arranged in a 36×6 array. Recall that a covariate is an intrinsic property of the observational units, as opposed to a treatment which is, or may be, assigned to the units. By definition, each marginal component *log* and *saw* is a covariate. In addition, *brand* is a covariate or classification factor, which is an property the saws, and *species* is also a covariate, which is a property of the logs; the design consists of two saws of each brand, and twelve logs of each species.

By contrast, *team* is a treatment that is assigned by the investigator to each of the 36 selected units only. In the description as given, *debarking* is also a treatment that is assigned to the logs. However, if the logs were initially segregated by bark status, it could plausibly be argued that no random assignment has occurred, in which case *bark* is a covariate or classification factor.

Regardless of its status, *bark* is a Boolean function $[36] \to \{0, 1\}$ on logs such that six logs of each species are debarked, and six are left intact. There are 924^3 functions of this type. If *bark* is a treatment factor assigned by randomization, it is a random variable selected according to some specified distribution from the indicated set of 924^3 functions. In most instances, the randomization distribution is uniform on functions having the desired balance.

Algebraically, *team* is a function from the 36 selected units into the set of teams; statistically, it is a random function chosen uniformly from a subset of such functions satisfying certain Latin-square constraints. Restricted randomization can be rather complicated, so this aspect is omitted from discussion here.

Each of the recorded factors

$$\textit{log}, \textit{species}, \textit{bark}, \quad \textit{saw}, \textit{brand}, \quad \textit{team},$$

is a function on the sample units. The number of levels is 36, 3, 2, 6, 3, 6 respectively. In general, the way in which a covariate such as *species*, *saw* or *brand* is accommodated in a statistical analysis or formal model is not the same

as the way in which a treatment is accommodated. Consider a particular unit u, a (log, saw) pair, which happens to be of the type (larch, brand 3). The model associates with u a probability distribution for the response-treatment pair; the treatment components are determined by randomization and are not independent. Hence, the model associates with u a conditional distribution $P_u(\cdot \mid T)$, one distribution on $\mathbb{R}$ for every treatment level that has positive probability of being assigned to u. It does not associate with u a [non-trivial] probability distribution for each species or for each brand because the species and the brand are both properties of u that are recorded at baseline.

2.3 Goals of statistical analysis

The chief purpose of the study is to compare the relative efficiencies of the three brands of saw, i.e., to compare one brand with another. There are 12 observations for each brand, and the sample averages are 8.78, 8.55 and 7.42 in minutes, or 2.10, 2.11 and 1.95 in log minutes, so brand 3 appears to be the most efficient. The main statistical challenge is to come up with a reasonable assessment of the standard error for brand contrasts. Is it better to do the analysis on the time scale or on the log scale? If we do the analysis on the log scale, how do we report effects on the time scale? Regardless of which scale is used, how do we calculate a standard error for brand effects?

In addition to brand effects, we can also investigate the effect of de-barking. The sample averages with and without bark are 8.80 and 7.70 minutes, or 2.13 and 1.97 on the log scale, so de-barking appears to reduce the cutting time by about 12–15%. How do we compute a standard error? Is the reduction approximately the same for each saw brand?

In addition to brand and de-barking effects, we can also investigate differences between the three species. There are 12 observations for each species, and the average cutting times for spruce, pine and larch are 8.03, 6.42 and 10.30 minutes respectively, or 2.04, 1.82 and 2.29 on the log scale. Larch, one of the few deciduous conifers, is evidently substantially harder or tougher than the other two. Regardless of whether we use the log scale or the time scale for averages, how do we calculate an honest standard error for species contrasts? Do we compute the standard error for each species contrast in the same way that we compute the standard error for brand contrasts?

Detailed answers to all of these questions are given in subsequent sections. At this stage, we provide brief answers to some of the questions without offering a detailed rationale.

First, the response is a time in minutes as measured by a stopwatch; ordinarily, the appropriate scale for analysis of temporal measurements by linear methods is the log scale. For some, this is obvious and needs no support; others may demand a formal justification (section 18.2). Research workers from an engineering background are accustomed to using logs to the base 10 without comment, but natural logs are used throughout these notes. On the log scale, the effects of bark, species, brand and team are additive, which implies that

they are multiplicative on the time scale. An analysis on the log scale does not imply that the conclusions must be reported on the same scale. Thus, in reporting point estimates of effects, we say that de-barking reduces the cutting time by 12–15%, not that de-barking reduces the cutting time by 1.1 minutes. For the particular task in the experiment, both statements are equally true, but, as an isolated statement, one is more sensible than the other. A more careful statement might emphasize that the de-barking reduction applies to the mean of the distribution. Likewise for species contrasts and brand contrasts.

Second, the estimated spruce versus larch contrast is a difference of average cutting times for two disjoint subsets of 12 observations each, the variance is $\sigma^2(1/12 + 1/12)$, and the standard error has the form

$$s\sqrt{1/12 + 1/12}$$

for some suitable estimator s^2 of σ^2. The estimated brand 3 versus brand 1 contrast is also a difference of averages of two disjoint subsets of 12 observations each, but the variance formula is entirely different and the estimated standard error is about 30% larger than that of the spruce/larch contrast. Why so? The reasons for this difference are subtle, but they are also fundamental and easily overlooked.

The difference is a consequence of the experimental design as described in the third paragraph of this section, rather than a consequence of any parametric or nonparametric model. The crux of the matter is that 36 logs are used in the design, but only six saws. It is one thing to make a statement about the relative efficiencies of two specific saws, C versus A; it is different matter to make a statement about the relative efficiencies of two brands, brand 3 versus brand 1. For a statement of the latter type, or a statement about spruce versus larch, the observed specimens must be typical for the brand or species. But the design includes 12 specimens of each species, and only two specimens of each brand.

The use of the two-sample variance formula $\sigma^2(1/12 + 1/12)$ for the spruce versus larch contrast does not imply that the set of cutting times for spruce and the set of cutting times for larch are assumed to be independent. They are not independent, and they are not assumed to be so, even conditionally on the design. Nonetheless, the two-sample formula makes good use of orthogonality, additivity and balance associated with the Latin square, so the analogous formula would not necessarily be appropriate in a less carefully designed experiment.

2.4 Formal models

Apart from the indicator for distinct logs, which is in 1–1 correspondence with sample units, the factors available in this design are as follows:

species, bark, saw.id, brand, team

In addition, *row* in the Latin square is equivalent to *species:bark*, and *col* is equivalent to *saw.id*. As always, each term that occurs in a linear model signi-

fies a vector subspace of $\mathbb{R}^n$, and the additive operator denotes the vector span, not the vector sum. Thus, the statement that *row* is equivalent to *species:bark* is intended as a statement about vector spaces, not a statement about indicator vectors or basis vectors. Every factor also determines a partition of the observational units into disjoint subsets, labelled or unlabelled, so the equivalence of factorial model-formula terms could equally well be be interpreted as a statement about induced partitions.

It is instructive to examine the output from the standard linear Gaussian model for log(time), in which the mean response lies in the subspace

$$E(Y) \in \mathcal{X} = \textit{species}+\textit{bark}+\textit{brand}+\textit{team},$$

the variances are constant and the covariances are zero. Note that the Latin-square column factor *saw.id* does not occur in this model. By assumption, two observations on the same saw are independent, and they are identically distributed if the two logs are of the same species and bark status.

The standard model is contrasted with one in which *saw.id* occurs as a block factor in the variance

$$E(Y) \in \mathcal{X}; \qquad \mathrm{cov}(Y_{ir}, Y_{js}) = \sigma_0^2 \delta_{ij} \delta_{rs} + \sigma_1^2 \delta_{rs}.$$

Once again, duplicates of the same brand have the same one-dimensional marginal distribution, all observations have the same variance $\sigma_0^2 + \sigma_1^2$, observations on different saws are independent, but observations on the same saw are positively correlated.

The least-squares estimates for both models can be computed from

```
fit0 <- regress(log(time)~species+bark+brand+team)
fit1 <- regress(log(time)~species+bark+brand+team, ~saw_id)
```

Ordinarily, the regression parameter estimates for these two models should be similar but not identical. Because of the balanced design, they are identical, but the standard errors are different, some a little smaller, others appreciably larger. Despite the fact that the mean square for brand replicates is not significantly larger than the mean square for residuals, the argument for a zero between-replicate variance is not compelling. Accordingly, the second version is preferred. On the other hand, additivity for species and bark effects is plausible on the log scale. Both models assume additivity for species and bark effects, which can be tested in the usual way.

If team effects were not a primary focus, they could reasonably be regarded as independent and identically distributed, in which case, the fitted model is obtained by using `team` as a block factor rather than a treatment factor

```
regress(log(time)~species+bark+brand, ~saw_id+team)
```

Because of orthogonality, the fitted values and standard errors for species and bark contrasts are exactly the same, whether team effects are fixed constants contributing to the mean or iid random variables contributing to the covariances.

2.5 REML and likelihood ratios

All of the models described above assume that the effects of species and debarking are additive on the log scale. How do we compute a likelihood ratio statistic for testing additivity in a situation where the model contains more than one variance component? For various reasons, this is a technically complicated question and there is at least one technically incorrect answer. But there is one answer that is both mathematically natural and technically correct, which is the one given by Welham and Thompson (1998): see chapter 17 for a detailed analysis. The answer that is recommended in the `lmer()` literature, which is to abandon REML and use ordinary maximum likelihood, may be technically defensible, but it is not the most natural for this setting.

The Welham-Thompson likelihood-ratio statistic on two degrees of freedom for testing the null hypothesis of additivity can be computed as follows:

```
K <- model.matrix(~species+bark+team+brand)
fit0 <- regress(log(time)~species+bark+team+brand, ~saw_id, kernel=K)
fit1 <- regress(log(time)~species*bark+team+brand, ~saw_id, kernel=K)
2*(fit1$llik - fit0$llik)    #  1.591
```

The kernel is a subspace of the observation space, which determines the likelihood criterion that is used for estimation purposes. For a valid likelihood-ratio statistic, it is essential that the kernel subspaces be the same for both fits. The kernel shown above is the REML default for the first fit, but it is not the default for the second. If we choose to follow the advice in the `lmer()` literature, we must adjust the argument to `kernel=0` in both `regress()` expressions, giving a likelihood-ratio statistic of 2.17 in place of 1.59. Although the difference is numerically not negligible, the asymptotic null distribution is χ^2_2, for which the 95th percentile is 6.0, so neither statistic indicates a departure from additivity.

If `team` is removed from the mean model but included as a block factor in the variance, the two likelihood-ratio statistics are 1.59 and 1.81 respectively. In that case, the kernel subspace for the Welham-Thompson statistic is `species+bark+brand`, which is the mean-value subspace under the null model.

2.6 Summary of conclusions

2.7 An open-ended counterfactual

The first sample unit was log number one paired with saw F; it was spruce with no bark, and was assigned to team I. The cutting time was 6.4 minutes. What would the cutting time have been if the same log had been paired with saw E and assigned to team II?

Before jumping to an answer, it is best to ask whether the question as phrased admits an answer. After all, the premiss of the question is that this particular log be cut by saw E when in fact it was cut by saw F. Bear in mind that the constraint that each log can be cut only once into $8'$ lengths is more a

matter of everyday practice in the lumber industry than a matter of physics or mathematics or even joinery. Although it would take some time, effort, expense and a water-tolerant glue, a cut log could be reassembled by butt-end gluing. With subsequent cuts offset by four inches or thereabouts, interference from previous cutting activity could be kept to negligible levels.

If an answer is deemed possible, is it most naturally given as a real number or as a probability distribution? If the answer is a probability distribution, it is presumably a conditional distribution given the data. In that case, the answer must come from a stochastic model that admits a joint distribution of cutting times for multiple cuts by different saws on the same log. The fact that we have not observed multiple cuts of this sort makes the problem all the more challenging, but it does not necessarily make it impossible.

2.8 Exercises

2.1 Suppose that intact logs are numbered 1:36, and that `species` is the species factor. Write code in R that picks uniformly at random a subset of six logs of each species for debarking, and stores the information as a Boolean treatment factor. Explain where the number $924 = 3 \times 4 \times 7 \times 11$ comes from.

2.2 The R commands

```
anova(lm(log(y)~row+team+saw_id));
anova(lm(log(y)~species*bark+team+brand+saw_id))
```

are designed to decompose the total sum of squares additively into components associated with certain subspaces, which are mutually orthogonal for this design. Explain how to compute the row sum of squares on five degrees of freedom directly from the six row averages

$$1.943,\ 1.737,\ 2.243,\ 2.129,\ 1.910,\ 2.341.$$

Arrange these six numbers in a 3×2 table, and explain the computation of the sums of squares for `species`, `bark`, and `species:bark` from this table of numbers.

2.3 Use the averages for the six saws A–F

$$2.122,\ 2.060,\ 1.975,\ 2.070,\ 2.156,\ 1.920$$

to compute the `brand` sum of squares on two degrees of freedom, the saw replicate sum of squares on three degrees of freedom, and the F-ratio (ratio of mean squares). Why is this two-part decomposition structurally different from the three-part decomposition in the preceding exercise?

2.4 Use the method described by Welham and Thompson (1998) to compute the REML likelihood-ratio statistic for comparing the two linear models

$$\mathcal{X}_0 = \mathit{species}+\mathit{bark}+\mathit{team}, \quad \mathcal{X}_1 = \mathit{species}+\mathit{bark}+\mathit{team}+\mathit{brand}$$

in the setting where *saw.id* occurs as a variance component. You may use R code as follows:

```
fit0 <- regress(log(y)~species+bark+team, ~saw_id)
K <- model.matrix(~species+bark+team)
fit1 <- regress(log(y)~species+bark+team+brand, ~saw_id, kernel=K)
c(fit1$llik, fit0$llik, fit1$llik-fit0$llik)
```

2.5 In the simple linear model setting with $\mu \in \mathcal{X}$ and $\Sigma \propto I_n$, show that the maximum value of the log likelihood is $\text{const} - n \log \|QY\|$, where $Q = I - P$ is the orthogonal projection with kernel $\mathcal{X}$, and the constant is independent of $\mathcal{X}$.

2.6 In the simple linear model setting, the F-ratio for testing the hypothesis $\mu \in \mathcal{X}_0$ versus $\mu \in \mathcal{X}_1$ is the ratio of mean squares

$$F = \frac{\|Q_0 Y\|^2 - \|Q_1 Y\|^2}{\|Q_1 Y\|^2} \frac{n - p_1}{p_1 - p_0},$$

where $\mathcal{X}_0 \subset \mathcal{X}_1$, and $p_r = \dim(\mathcal{X}_r)$. Using the expression in the preceding exercise, show that the log likelihood ratio statistic is a monotone increasing function of F:

$$2\Lambda = m \log\left(1 + \frac{(p_1 - p_0)F}{n - p_1}\right).$$

where $m = \dim(\mathbb{R}^n/\mathcal{X}_0) = n - p_0$ for the Welham-Thompson statistic, and $m = n$ for the ordinary likelihood-ratio statistic.

2.7 Check that the F-ratio for brand differences is in approximate agreement with the Welham-Thompson REML statistic computed in exercise 2.15. Explain why you need $m = 6 - 1$ rather than $m = 36 - 9$ in this comparison.

2.8 Express the random-effects models from the previous section in `lmer()` syntax, and check that the parameter estimates agree with `regress()` output.

Example 3

3.1 Drosophila mating preferences

This project concerns the experimental design and the data analysis in the paper titled *Commensal bacteria play a role in mating preference of Drosophila Melanogaster,* published in 2010 by Sharon *et al.* in Proceedings of the National Academy of Sciences, vol 107, No. 46, 20052–20056. The experimental design and the goals are straightforward in principle: do female flies have a preference for male flies that have been fed on the same diet rather than flies that have been fed a different diet? Some of the finer experimental details are crucial for model formulation, analysis and interpretation, but are easy to miss in a superficial reading. Partial information on the design and analysis is given below, so you are encouraged to read the paper for yourself for additional background.

Two breeding populations of genetically identical fruit flies were raised separately for roughly forty generations on one of two diets, here denoted by C (corn-molasses-yeast) and S (starch). At certain stages, flies destined for experimentation (test matings) were removed from the breeding populations and raised for one intermediate generation on the standard CMY diet before testing. Thus the testing for generation six was done on the offspring, so generation six is really 6+1: see Fig. 1 of the paper. Tests were done for selected generations from two to 37. The table of mating counts shown on the right, is implicit in the authors' Fig. 2. It is not given explicitly in the published paper or in the supplementary online materials, but was provided by the authors on request. It contains five columns of data, generation number, followed by the mating counts for the four types, CxC,

Table 3.1 *Drosophila* mating counts

	Type of cross			
Gen	CxC	CxS	SxC	SxS
2	12	8	9	16
6	10	5	9	10
7	17	9	9	15
9	8	7	7	9
10	18	13	5	12
11	12	5	7	14
13	14	9	8	12
15	18	9	7	15
16	14	5	5	10
17	31	22	12	27
20	23	13	10	20
21	13	7	5	14
26	30	19	12	21
31	9	7	3	10
37	20	14	11	17
111	18	11	7	16
112	16	11	8	15
113	22	13	8	13

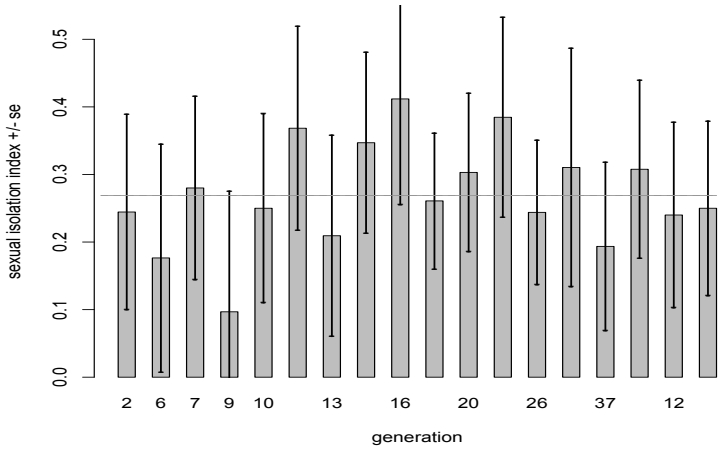

Figure 3.1: Sexual isolation index plotted against generation number.

CxS, SxC, SxS. Here SxC denotes matings of male flies whose parents were raised on diet S with females whose parents were raised on diet C. Matings of types CxC and SxS are called homogamic; the other types are heterogamic.

The experimental set-up consisted of a number of mating wells, from 20 to 70, with four flies in each well, one male and one female of each dietary type. Over a one-hour period, each mating was noted, and the totals for each type were recorded. The number of mating wells was not reported for the first three generations, but the values reported for subsequent generations were 24, 39, 20, 24, 36, 23, 70, 46, 24, 45, 23, 48, 48, 48, 48. The last three rows of data are taken from a parallel experiment run under a similar protocol, so the mating probabilities are expected to be similar, but the generation numbers should be ignored.

3.2 Initial analyses

3.2.1 Assortative mating

The main summary of the experimental data is given in the authors' Fig. 2, which is a barplot of the estimated sexual isolation index (SII) for each of 15 generations. It is similar in style to Fig. 3.1, which also includes the additional three generations. The sexual isolation index is defined as the difference $p_{\text{hom}} - p_{\text{het}} = 2p_{\text{hom}} - 1$ between the probability of a homogamic mating and the probability of a heterogamic mating. The reported values are the empirical relative frequencies observed in each generation. Random mating, or absence of assortative mating, implies $p_{\text{hom}} = 1/2$ or SII $= 0$. Under the assumption of binomial sampling, the estimate $\hat{p}_{\text{hom}}$ has variance $p_{\text{hom}}(1 - p_{\text{hom}})/n$, so the

observed isolation index has variance $4p_{\text{hom}}p_{\text{het}}/n$, or $(1-\text{SII}^2)/n$ which reduces to $1/n$ in the absence of assortative mating.

The height of each bar in Fig. 3.1 is the estimated sexual isolation index for that generation, and the whisker length is one binomial standard error, i.e., $\pm\sqrt{(1-\text{SII}^2)/n}$. The horizontal line at SII = 0.27 is the overall average estimated from the pooled data. Superficially, at least, all of these calculations seem quite standard statistically. However, there are both statistical and non-statistical reasons to have a closer look at the design and the analysis.

3.2.2 Initial questions and exercises

1. What are commensal bacteria?

2. A trained observer can distinguish male from female fruit flies. But flies raised on diet C are genetically and morphologically indistinguishable from flies raised on diet S. How did the authors determine the diet type of a mating pair?

3. Wing vibration appears to be an important part of the *Drosophila* courtship ritual. What are the implications for experimental design? Are there ways in which the design could be improved in this respect?

4. The Bernoulli model and the binomial distribution are used at various points in the authors' analyses, either explicitly or implicitly. In the context of this experiment, what assumptions are required to justify the Bernoulli model?

5. Since the segregation index starts off at zero for generation zero, one might expect the value to increase slowly over generations. Is there any evidence for this, or can it reasonably be taken as constant after 1–2 generations? Construct a suitable test statistic, explain how it addresses the question, and indicate what conclusion is warranted.

6. Are the data consistent with the assumption of independence of mating events in successive generations? Compute a relevant statistic and explain what it tells you.

7. All analyses in the paper are essentially unaffected by switching sexes, which is mathematically fair and even-handed. However, after mating, a female fruit fly is no longer receptive to courtship. By contrast, a male fly may mate a second time if a receptive female is available. Discuss briefly how sexual asymmetry might affect the design or the analysis. You may assume that each well contains four flies, each courtship/mating event lasts up to 10 minutes, the observation period lasts about 40–60 minutes, whereas the female refractory period lasts about 24 hours.

8. Pearson statistic for testing homogeneity of relative frequencies for each generation has an approximate chi-squared distribution on 3×17 degrees

of freedom under standard assumptions. How accurate is this null distributional approximation when applied to Table 3.1. Compute the permutation distribution by simulation, using random matching to keep the row and column totals fixed, and compare the histogram of simulated values with the χ^2_{51} density.

3.3 Refractory effects

3.3.1 More detailed data

Table 3.2, which was subsequently provided by the authors, contains a more detailed description of the mating events in each generation. For each mating well, either zero, one or two matings may occur during the observation period. In single-mating wells, the mating is one of four types cc, cs, sc or ss, of which two are homogamic and two heterogamic; in double-mating wells, the female refractory period constrains the set of mating combinations to four, cc.ss, cs.sc, cc.cs and sc.ss, of which one is double homogamic, one is double heterogamic, and two are mixed. The order in which the matings occur is not reported in the table and is not considered here. The combination cc.cs implies that the c-male mated with both females, while the s-male did not participate. The combination sc.ss implies that the s-male mated with both females, while the c-male did not participate or was rejected. The other combinations do not occur because of the refractory constraint. A female that has already mated does not mate a second time within about 24 hours. Each courtship ritual and mating takes approximately 10–12 minutes, so the observation period of 40–60 minutes is sufficient for one male to copulate with both females if they are receptive.

It is important to observe the fundamental difference between the two versions of the *Drosophila* data. The objects that are counted in Table 3.1 are matings, which are of four types; the objects that are counted in Table 3.2 are wells of various types, one type for each column. In the first case, each observational unit is a mating, and the response is the mating type; in the second case, each observational unit is a well, and the response is one of nine types. From a statistical standpoint, it is natural to regard the activity in one well as a multinomial event with nine activity classes that are disjoint and exhaustive, at least in the biological sense. It is also natural to regard flies as exchangeable modulo their sex and diet type, so that events in distinct wells may be taken as independent with identical distributions for all wells in the same generation. Those assumptions justify the reduction of the data to the counts in Table 3.2 as the sufficient statistic. Provided that the activity in one well is independent of that in other wells, each row is an independent multinomial random variable.

Biologically speaking, the multinomial parameters need not be constant from one generation to the next. Apart from the possibility of a monotone increasing segregation index, there are more mundane reasons for distributional heterogeneity that may be related to experimental procedure. One possibility is that the inclination to mate may depend on temperature and other environmental

Table 3.2. Number of wells having matings of each type

	Single matings				Double matings				Null	Total
Gen	cc	cs	sc	ss	cc.ss	cs.sc	cc.cs	sc.ss	–	wells
2	1	1	0	1	11	6	0	3	0	23
6	1	0	2	1	7	5	0	2	0	18
7	2	1	0	3	9	4	4	5	4	32
9	1	2	3	3	5	3	2	1	4	24
10	3	3	0	2	10	5	5	0	9	37
11	2	1	2	1	8	1	2	4	2	23
13	1	0	0	0	11	7	2	1	2	24
15	2	1	2	2	11	3	5	2	8	36
16	2	0	2	1	9	3	2	0	3	22
17	0	8	3	7	18	6	9	1	17	69
20	8	4	4	4	8	3	5	0	9	45
21	1	2	0	2	11	4	1	1	2	24
26	2	1	2	1	17	7	11	3	3	47
31	3	2	0	6	4	3	2	0	3	23
37	5	1	3	7	9	7	6	1	9	48
111	2	2	1	1	12	4	4	2	14	42
112	5	2	2	7	7	5	4	1	15	48
113	9	3	1	3	10	7	3	0	10	46

factors that vary with the season, and hence are not constant from one generation to the next. Another very real possibility is that the period set aside for observation is not quite constant from generation to generation, in which case the fraction of null-mating wells is expected to be be greater for shorter observation periods. Likewise, for purely mechanical reasons, the fraction of double-mating wells is likely to be low for shorter observation periods.

In principle, each column in Table 3.1 is derivable as a specific linear combination of the columns in Table 3.2. Each linear combination has three unit coefficients and six zeros. For example, the CxS column is the sum of columns cs, cs.sc and cc.cs, while the SxC column is the sum of sc, cs.sc and sc.ss. Both combinations include cs.sc. This linear projection structure implies that the counts in one row of Table 3.1 are correlated in a non-multinomial way, which invalidates the distributional assumptions on which the paper is based. In practice, there are a few discrepancies between the two tables, which is not uncommon in laboratory work. Unless otherwise specified, all subsequent analyses in this chapter use the data in Table 3.2.

3.3.2 Follow-up analyses

Given that the main focus is on the excess of homogamic over heterogamic matings, how should we analyze the new version of the data for evidence bearing on the issue of commensally-related assortative mating? Assuming for the moment that there is sufficient homogeneity across generations, it is natural first

to examine the aggregate counts or column totals, which are as follows:

cc	cs	sc	ss	cc.ss	cs.sc	cc.cs	sc.ss	null
50	34	27	52	177	83	67	27	114

Among all events in 163 single-mating wells, 102 are homogamic and 61 heterogamic, so the homogamic sample fraction is 0.626 and the standard error is 0.038. If mating events occurred non-preferentially according to the Bernoulli-1/2 model, we should expect about 81.5 ± 6.4 homogamic and the same number of heterogamic matings, so the observed value is a little more than 3.2 standard deviations away from the non-preferential null. Equivalently, the SII index is 0.251 with standard error $\sqrt{(1/163)}$ computed under the Bernoulli-1/2 model, and the ratio is $0.251\sqrt{(163)} = 3.2$. Three or more standard deviations is usually regarded as moderately strong evidence against the null, so even if we restrict attention to single-mating wells, the evidence for assortative mating is clearly established.

In the double-mating wells, 448 matings out of 708 are homogamic, so the homogamic fraction is 0.633. To obtain a standard error for the sample fraction, the four totals (Y_1, Y_2, Y_3, Y_4) are regarded as multinomial with index $Y_\cdot = 354$, parameter vector π, and covariance matrix $\big(\mathrm{diag}(\pi) - \pi\pi'\big)Y_\cdot$. The number of homogamic matings is the linear combination $2Y_1 + Y_3 + Y_4$, the number of heterogamic matings is $2Y_2 + Y_3 + Y_4$, and the total number of matings is $2Y_\cdot = 708$. The variance of the linear combination is a quadratic form in the multinomial covariances, whose estimate is 235.04, so the standard error of the homogamic fraction in the sample is $\sqrt{235.04}/708 = 0.022$. The observed value is six standard errors away from the null, so once again the evidence strongly supports assortative mating.

For a slightly different version of the preceding argument, the difference between the number of homogamic and heterogamic matings is $2Y_1 - 2Y_2$, which does not involve the mixed-well counts Y_3 or Y_4. Arguably, the mixed-event wells are uninformative for testing. The null hypothesis of no assortative mating implies that Y_1 has the same distribution as Y_2, so it is possible in this setting to construct an exact binomial test by conditioning on the total $Y_1 + Y_2$. However, the *estimate* of the segregation index is not independent of the mixed double-mating well counts.

In this case, the estimates obtained from the two sources 0.626 ± 0.038 and 0.633 ± 0.022 are in good agreement with one another. The standard error of the difference is the square root of $0.038^2 + 0.022^2$, which is 0.044, whereas the observed difference is only 0.007. A similar analysis on the SII scale gives an equivalent answer. The estimates may be pooled or combined in the standard manner with weights inversely proportional to variances.

3.3.3 Lexis dispersion

The Lexis dispersion statistic, which is the ratio of Pearson's chi-squared statistic to its degrees of freedom, is a natural gauge of variation in a contingency

table in which the reference value of unity is the expected value under homogeneous multinomial sampling. For the four single-mating columns in Table 3.2 the value is is 48.4/51, and for the four double-mating columns the value is 50.95/51. As we had hoped, both are satisfactorily close to unity, so there is no evidence of inter-generational inhomogeneity in mating behaviour for either the single-mating wells or the double-mating wells.

For the 18×2 matrix whose columns are the tallies for single- and double-mating wells in each generation, the Lexis dispersion statistic is $48.2/17 = 2.83$. We conclude that there is substantial heterogeneity in the fraction of single- versus double-mating wells in successive generations. This type of inhomogeneity does not invalidate the analyses proposed in the preceding section or those in subsequent sections. As mentioned earlier, it could easily be attributed to environmental variation or to incidental variation in experimental counting procedure.

3.3.4 Is under-dispersion possible?

The dispersion index for Table 3.1 is $19.14/51 = 0.37$, which shows clearly that the counts in that table are substantially under-dispersed. Over-dispersion is common in experimental and observational work, while under-dispersion is rare, so statisticians are naturally on the lookout for phenomena that give rise to under-dispersion. The main explanation for under-dispersion in this instance appears to lie in the experimental design with four flies per mating well and its interaction with the female refractory effect.

This section offers an analysis of whether the under-dispersion that is observed in Table 3.1 should be expected on the basis of its derivation from Table 3.2. The analysis is done under the following 'multinomial assumption', which seems mathematically natural for this setting.

1. Given the vector m_1 of single-mating well counts in each generation, the 18×4 table T_1 consisting of the first four columns of Table 3.2 has independent multinomial rows, and the probability vector π_1 is constant across generations.

2. Given the vector m_2 of double-mating well counts in each generation, the 18×4 table T_2 consisting of the columns 5–8 of Table 3.2 has independent multinomial rows, and the probability vector π_2 is constant across generations.

3. The tables T_1 and T_2 are conditionally independent given m_1, m_2.

Although homogeneity across generations is an important component, we refer to these collectively as 'the multinomial assumption'.

Let L be the matrix that converts double-mating well counts into mating

counts of four types:

$$L = \begin{array}{c} \\ \text{cc.ss} \\ \text{cs.sc} \\ \text{cc.cs} \\ \text{sc.ss} \end{array} \begin{array}{cccc} \text{cc} & \text{cs} & \text{sc} & \text{ss} \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{array}$$

so that $T = T_1 + T_2 L$ counts total matings of each type in each generation. Discrepancies between Table 3.1 and T have already been noted, but are not the focus of this analysis. By assumption, the rows of this table are independent. The expected mating count for generation i is the linear combination $m_{1,i}\pi_1 + m_{2,i}L'\pi_2$ of multinomial vectors. However, even if $2\pi_1 = L'\pi_2$, the distribution of T is not multinomial, so Pearson's statistic does not have its standard χ^2-reference distribution. The question to be addressed is whether it is possible under the multinomial assumption for the 18×4 table T to be under-dispersed relative to the multinomial, and to what extent.

The question can be addressed in a variety of ways, either analytically or by simulation. For a partial analytical solution, the mean vector and covariance matrix of the ith row of T are

$$\begin{aligned} \mu_i &= E(T_i) = m_{1,i}\pi_1 + m_{2,i}L'\pi_2 = (m_{1,i} + 2m_{2,i})\pi, \\ \Sigma_i &= \operatorname{cov}(T_i) = m_{1,i}V(\pi_1) + m_{2,i}L'V(\pi_2)L, \end{aligned}$$

where $V(\pi) = \operatorname{diag}(\pi) - \pi\pi'$ is the 4×4 multinomial covariance matrix. Pearson's statistic is the quadratic form

$$X^2 = \sum_{i,j} \frac{(Y_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}}$$

whose expected value is approximately

$$\sum_{i=1}^{18} \operatorname{tr}\big(\hat{\Sigma}_i \operatorname{diag}(\hat{\mu}_i^{-1})\big).$$

Using the natural moment estimates for the vectors π_1, π_2 and π, the estimated mean of X^2 is 39.39. For a more accurate approximation, we should multiply by 17/18 to account for parameter estimation. This gives $E(X^2) \simeq 37.2$, so the appropriate null reference level for the Lexis dispersion index is $39.39 \times 17/(18 \times 51) = 0.73$. The conclusion from this analysis is that under-dispersion is not only possible but also expected in this situation.

A more accurate estimate of the null distribution of Pearson's statistic can be obtained by simulation along the lines of section 3.4. First generate two conditionally independent hypergeometric tables having the same marginal totals as T_1 and T_2, combine them into a single table as in $T_1 + T_2 L$, and then compute the Pearson statistic. The conclusion from 5000 simulations is that the null mean is 37.2 and the variance is approximately 76.8. Relative to this distribution, the observed value $X^2(T) = 24.1$ falls just below the 5% point;

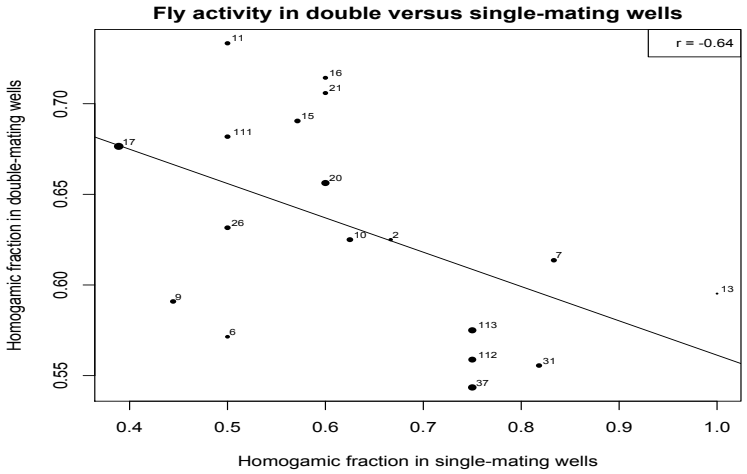

Figure 3.2: Scatterplot of homogamic mating fractions for 18 generations

the observed value for Table 3.1 is 19.1, which is below the 1% point. The conclusion is that under-dispersion is expected, though not quite to the extent observed in T or in Table 3.1.

3.3.5 Independence

One fundamental assumption in all of the foregoing analyses is that activities occurring in distinct wells must be independent. Ordinarily, the assumption of independence seems so obvious experimentally that it cannot be called into question. After all, distinct wells contain distinct flies whose activities cannot possibly be coordinated. But science rightly demands that assumptions be checked where possible, and the design of this experiment with the data in Table 3.2 provide a rare opportunity to check the independence assumption—at least in part.

Flies in single-mating wells are necessarily distinct from flies in double-mating wells, so in the absence of inter-well communication, we must expect all activity in single-mating wells to be independent of all activity in double-mating wells. This is part of the third component of the multinomial assumption in the preceding section. In particular, we must expect the homogamic fraction in single-mating wells to be statistically independent of the homogamic fraction in double-mating wells. Any failure of independence in this form must have far-reaching consequences for *Drosophila* experimentation. To paraphrase Lord Denning's notorious judgement in 1980, the possibility of coordinated mating activities in distinct wells is such an appalling vista that every sensible *Drosophila* experimentalist would say 'It cannot be right...'.

Each generation furnishes a pair of homogamic fractions, one for single-mating wells and one for double-mating wells. Figure 3.2 is a scatterplot of the

18 pairs, one pair for each generation. Contrary to expectation, it shows not only that homogamic fractions in the same generation are correlated, but also that the correlation is negative ($r = -0.64$). The null distribution of sample correlations is symmetric about zero with a standard deviation of about 0.23, so the observed correlation is far removed from the bulk of the null distribution. Hypergeometric simulation by random matching points to a left tail p-value of approximately one in 850, which is equivalent to three standard deviations from expectation on the standard normal scale. In other words, the hypothesis of independence of events in disjoint wells is firmly rejected by these data.

The correlations indicated by Fig. 3.3 are synchronous in time; there is no suggestion that the homogamic fractions in one generation are correlated with homogamic fractions in previous or subsequent generations. Inter-well communication is one potential explanation for synchronous correlations. *Drosophila* courtship rituals are not silent, so sound leakage may be possible. Pheromonal leakage may be more likely. However, in order to achieve the observed *negative* correlation, the communication must be anti-symmetric or conspiratorial, so it is unlikely that leakage alone could suffice. On balance, therefore, it seems safe to rule out inter-well communication as a likely explanation. However, no satisfactory alternative has yet been suggested.

The phenomenon analyzed in section 3.3.4 accounts for an under-dispersion factor of 0.73. Negative correlation has an additional and potentially greater effect on the variance of linear combinations ($1 - r = 0.36$), and this appears to be the main reason for the extreme under-dispersion seen in Table 3.1. For example, the marginal table of homogamic versus heterogamic counts derived from Table 3.1 has a dispersion factor of $4.29/17 = 0.25$, which matches nicely with the product $0.73(1 - r) = 0.26$.

The sample correlation reported above is weighted harmonically with weights w_t satisfying $w_t^{-1} = m_{1,t}^{-1} + m_{2,t}^{-1}$ for generation t, and the points in Fig. 3.2 are enlarged in areal proportion. Either $m_{1,t} = 0$ or $m_{2,t} = 0$ implies $w_t = 0$, which is the main reason for choosing harmonic weights. Some weighting is needed to accommodate the different sample sizes, and this harmonic weighting may not be optimal, but the correlation value is not especially sensitive to the choice of weights.

3.4 Technical points

3.4.1 Hypergeometric simulation by random matching

A two-way contingency table is a rectangular array Y whose components Y_{ij} are non-negative integers. Usually, Y_{ij} is the number of observational units for which one attribute or factor is equal to i and a second attribute is equal to j. Thus, the table is indexed by attribute values. Let $m_i = Y_{i\bullet}$ be the ith row total, and let $s_j = Y_{\bullet j}$ be the jth column total so that $m_\bullet = s_\bullet = Y_{\bullet\bullet}$ is the overall total. A random table is said to have the hypergeometric distribution if

the joint distribution is

$$\mathrm{pr}(Y = y) = \frac{\prod_i y_{i\bullet}! \prod_j y_{\bullet j}!}{y_{\bullet\bullet}! \prod_{ij} y_{ij}!}.$$

The row and column totals are arbitrary fixed positive integers, so the probability mass function is inversely proportional to $\prod_{ij} y_{ij}!$.

If Y has the hypergeometric distribution, so also does the transposed array. If Y is a random matrix whose rows Y_i are independent multinomial vectors, $Y_i \sim M(m_i, \pi)$, which are homogeneous in the sense that they have the same multinomial probability vector, then the conditional distribution given the column totals is hypergeometric.

One way to simulate a hypergeometric random table having given row and column totals is by random matching of the components of two n-component vectors. Suppose that `row` has m_r components equal to r, and `col` has s_j components equal to j, with $\sum m_r = \sum s_j = n$. Random matching permutes the components of `row` uniformly at random, does the same for `col`, and then tabulates or counts the ordered pairs (r, j) thus generated. Distributionally speaking, it is necessary only to permute one of the vectors as follows:

```
RHG <- function(rowsum, colsum){
   # rowsum and colsum are integer vectors having the same sum
   row <- rep(1:length(rowsum), rowsum)
   col <- rep(1:length(colsum), colsum)[order(runif(sum(colsum)))]
   table(row, col)
   }
```

To simulate the null distribution of Pearson's statistic or any other statistic such as the deviance, we compute the statistic for each table thus generated, and report the histogram. The analysis near the end of section 3.3.4 calls for two independent hypergeometric tables T_1, T_2, followed by Pearson's statistic computed on the linear combination $T_1 + T_2 L$. The analysis in section 3.3.5 also calls for the same pair of independent hypergeometric tables followed by a symmetric correlation statistic $R(\cdot, \cdot)$ computed as a function of the pair $(T_1, T_2 L)$.

3.4.2 Pearson's statistic

Pearson's statistic is a quadratic form in residuals, $X^2 = (Y - \mu)' \Sigma^{-1} (Y - \mu)$, which is a scalar measure of variability in the response relative to a given reference distribution whose mean vector and covariance matrix are μ and Σ. In most cases, the mean vector is estimated from the data, and Σ is a function of μ.

For counted data in the form of a contingency table, the reference distribution is usually Poisson or binomial with independent components, or multinomial with independent rows. In all of these cases, the algebraic form is the same,

$$X^2 = \sum_i \frac{(Y_i - \hat{\mu}_i)^2}{\hat{\mu}_i},$$

where $\hat{\mu}_i$ is the fitted mean value. In the binomial case, the sum extends over both response classes—failure and success—so that the net contribution from a binomial pair $(Y_{i,0}, Y_{i,1}) \sim B(m_i, \pi_i)$ for which $\hat{\mu}_{i,0} = m_i\hat{\pi}_i$ and $\hat{\mu}_{i,1} = m_i(1-\hat{\pi}_i)$ is

$$\frac{(Y_{i,0} - \hat{\mu}_{i,0})^2}{\hat{\mu}_{i,0}} + \frac{(Y_{i,1} - \hat{\mu}_{i,1})^2}{\hat{\mu}_{i,1}} = \frac{(Y_{i,0} - m_i\hat{\pi}_i)^2}{m_i\hat{\pi}_i(1-\hat{\pi}_i)}.$$

The Poisson form of Pearson's statistic differs from the binomial form only in the variance function, $\Sigma = \mathrm{diag}(\mu)$ for the Poisson covariance; and $\Sigma = \mathrm{diag}(m_i\pi(1-\pi))$ for the binomial. But the Poisson form covers both provided that we sum over both successes and failures.

The sampling distribution of the statistic depends on the distribution of Y and on the degrees of freedom used up in the estimation of μ. Exact moments are available in a few special cases, all of them null in a suitable sense. For a single multinomial $Y \sim M(m, \pi)$ with k classes and given probability vector, we have

$$\begin{aligned} E(X^2) &= k - 1, \\ \mathrm{var}(X^2) &= 2(k-1)\frac{m-1}{m} + \frac{1}{m}\sum \pi_j^{-1} - k^2/m. \end{aligned}$$

The third cumulant is given in McCullagh and Nelder (1989, p. 169). The asymptotic distribution for large m is χ^2_{k-1}.

For an $r \times c$ contingency table that is distributed according to the hypergeometric distribution with strictly positive row and column totals, the Haldane-Dawson formulae give the exact mean and variance. The mean does not depend on the row or column totals, but only on the overall total:

$$E(X^2) = (r-1)(c-1)m/(m-1).$$

The variance of X^2 depends on the sum of reciprocals of the row and column totals: see McCullagh and Nelder (1989, p. 244).

Despite warnings given freely by over-cautious computer software, the $\chi^2_{(r-1)(c-1)}$ approximation is quite accurate even for a large sparse table such as the 12×12 birth-death table where the average cell count is only 2.4. Even for Bortkewitsch's horsekick fatality data for 14 Prussian army corps over 20 years (Andrews and Herzberg, 1985, 17–18), where the mean is only 0.7 fatalities per corps per year, the χ^2_{247} approximation is reasonably good in the upper tail. The left tail is not so good. In that instance, the Haldane-Dawson values for the mean and variance are 248.3 and 419.8, so the variance-to-mean ratio is only 1.69 as opposed to 2.0 for the χ^2 approximation. The moment-matching approximation $0.85\chi^2_{294}$ is quite accurate in both tails.

Pearson's statistic has a role to play in the analysis of counted data, mainly as a metric for relative dispersion. Over the past 70 years, various authors have pointed out its inferential limitations, and have sought to modify and strengthen it in various ways (Yates, 1949; Cochran, 1954; and Armitage, 1955). Its deficiencies for significance testing are entirely unrelated to the adequacy of the χ^2 or other distributional approximation. The discussion in this section

focuses on its use as a dispersion index; it is not intended as an endorsement of its widespread use in applications as a test for independence or lack of association.

3.5 Further Drosophila project

The paper *Sex-specific responses to sexual familiarity, and the role of olfaction in Drosophila* by Tan, Løvlie, Greenway, Goodwin, Pizzari and Wigby, which was published in *Proceedings of the Royal Society, Series B* (2013), discusses a number of experiments that were designed to investigate the mating preferences of fruit flies. The main focus is on the courtship behaviour of males, and specifically whether males preferentially court novel females over familiar females. A directly familiar female is the previous mate, and a phenotypically familiar female is a sister of the previous mate. According to the abstract

> ...we show that male and female Drosophila melanogaster respond to the direct and phenotypic sexual familiarity of potential mates in fundamentally different ways. We exposed a single focal male or female to two potential partners. In the first experiment, one potential partner was novel (not previously encountered) and one was directly familiar (their previous mate); in the second experiment, one potential partner was novel (unrelated, and from a different environment from the previous mate) and one was phenotypically familiar (from the same family and rearing environment as the previous mate). We found that males preferentially courted novel females over directly or phenotypically familiar females. By contrast, females displayed a weak preference for directly and phenotypically familiar males over novel males.

As it turns out, the statistical analysis in the original paper is seriously deficient in a number of ways. In a 2014 correction note, the authors remark

> ... the statistical models we used for analysing male courtship behaviour did not take into account temporal correlations in courtship events within males. Consequently, the variance in courtship events was higher than predicted by the model, and the excess dispersion could potentially result in errors in conclusions. This highlights the general potential for high-frequency sampling of behaviours to give rise to high temporal correlations of event counts within a dataset, and the importance of correcting dispersion factors when analysing this type of data.

In other words, the courtship activity for one male was recorded on multiple occasions over a short period, and the sequence of records was analyzed as if the activities on successive occasions were independent events measured on unrelated flies.

There is nothing intrinsically wrong with high-frequency sampling provided that the statistical analysis accommodates the inevitable serial correlation in a

satisfactory way. If the activity of the focal male were recorded at 24 frames per second, we may mark each frame in which courtship is directed at the novel female by the label 'N' and those in which it is directed at the familiar female by the label 'F'. While it may be reasonable to treat a single marked frame as a Bernoulli random variable, it is obviously unreasonable to treat the *sequence* of frames as a Bernoulli sequence with independent components. For the same reason, it is unreasonable to treat the number of 'N' frames as a Poisson or binomial variable. This statement may be obvious at a sampling rate of 24 frames per second, but it applies equally at a sampling rate of one per minute or one per hour. Doubling the frame rate doubles the computational burden, but has a negligible effect on information pertaining to sexual preferences.

One possibility for analysis is to reduce the frame sequence to the fraction of time spent in each activity, and to regard these temporal fractions as a compositional response in the sense of Aitchison (19??).

The data for three of these experiments are available in the files

```
eyedat <- read.table("CoolEyeColorArchive.dat",  header=TRUE)
paintdat <- read.table("CoolPaintArchive.dat",  header=TRUE)
decapdat <- read.table("PhenoMaleDecapArchive.dat",  header=TRUE)
```

Additional information is available in the file `Coolidge.R`. Other data files are available online.

3.6 References

Much of the analysis in section 3.3 is based on an unpublished report by Dan Yekuteli, which was provided by the author.

3.7 Exercises

3.1 Use the normal approximation to the binomial to compute the probability that the horizontal line in Fig. 3.1 intersects all 18 whiskers at ± 1 standard deviations. Devise a better approximation by simulation that takes account of the fact that the SII index has been computed from the same data.

3.2 Is the total number of matings in Table 3.1 related to the number of mating wells? Is the pattern of variation different for the experiments reported in the last three rows? Explain how you address such questions.

3.3 For the experiment giving rise to the data in Table 3.2, an algebraically natural assumption is that the allowable double matings occur as a Poisson process at a rate proportional to the product of the single-mating rates. It is also natural—physically if not mathematically—to allow separate factors for single and double wells, and a reduced rate for wells in which one male does double duty. Formulate this statement as a Poisson log-linear model or four-class multinomial model, and check whether the data are in compliance with the product

assumption. For this exercise, the multinomial assumption in section 3.3.4 may be used. (The computation for this question may involve the entire table, but parameter estimates and other conclusions must be a function of the column totals only. Why so?)

3.4 Hypergeometric simulation in section 3.3.5 implies a symmetric null distribution with standard deviation 0.23 for the weighted sample correlation of homogamic pairs. One suggested alternative to random matching is to generate the null distribution by randomly permuting the vector $(\pi_{2,i}, m_{2,i})$ of double-mating homogamic fractions, keeping the sample-size attached to each fraction. Check that random permutation of generations also gives a symmetric null distribution with standard deviation at least 10% larger than the hypergeometric null. Which of these null distributions is the relevant one to use as a reference in this setting? Explain your reasoning.

3.5 The file `...birth-death.R` contains the data compiled by Phillips and Feldman (1973) on the month of birth and the month of death of 348 'famous Americans'. Investigate whether the month of death is or is not independent of the month of birth. The data are given as a 12×12 table of event counts. (This is not a generic contingency table because the row labels and the column labels are not only the same, but also cyclically ordered. Both aspects of the structure are relevant to the question posed, and both should be exploited in your analysis.)

3.6 The advice sometimes given for the validity of the χ^2 approximation to the null distribution of Pearson's statistic is that the minimum expected value should exceed a suitable threshold, usually in the range 3–5. However, the mean count for the birth-death table is 2.42, so the expected count in every cell falls below the threshold. Compute the null distribution of Pearson's statistic by hypergeometric simulation. Plot the density histogram of simulated values, and superimpose on it the χ^2_{121} density function. (This is intended as a computational exercise only. It is *not* a suggestion for data analysis aimed to address the question posed by Phillips and Feldman.)

3.7 Check the calculations reported in the last paragraph of section 3.4.2 for Bortkewitsch's horsekick data. Compute the row and column totals, and simulate the null distribution of X^2 by random matching. Superimpose the χ^2_{247} density on a histogram of the simulated values. Find two numbers a, b such that the first two moments of $a\chi^2_b$ coincide with the Haldane-Dawson moments. Superimpose this scaled chi-squared density on your histogram. (The intent of this exercise is solely to provide insight into distributional approximation. It should *not* be read as an endorsement for data analysis.)

Example 4

4.1 Plant growth: data description

The file `PlantGrowth.dat` contains the heights in mm. of 70 *Arabidopsis* plants measured every four days from day 29 to day 69 following the planting of seeds. The ultimate heights range from 19mm to 40mm, and most heights are recorded to the nearest millimetre.

The cumulative number of plants brearded was zero up to and including day 29, eight by day 33, 40 by day 37, and all 70 by day 41. Thus, sprouted plants were first recorded on day 33, and all plants had appeared by day 41. Or, to put it more accurately perhaps, no additional plants emerged after day 1. By day 65 or earlier, the growth was complete; for each plant, the height recorded on day 69 was the same as the height on day 65.

Plant age is most naturally measured from its birth at brearding rather than the date on which seed was planted. In this experiment, all seeds were planted on the same date, but the date of brearding varies from plant to plant. The brearding date is deemed to be the last date on which the height was recorded as zero rather than the first date on which the height was positive. In other words, eight plants were deemed to be born on day 29, 32 on day 33, and so on.

The typical growth trajectory for *Arabidopsis* begins at zero on day 0, reaching a plateau whose height varies from plant to plant. Regardless of the ultimate height, the semi-maximum height is attained in about 13 days, which is fairly constant from plant to plant. By inspection of the graphs in Fig. 4.1, it appears that the standard *Arabidopsis* growth trajectory is roughly $h(t) \propto t^2/(\tau^2 + t^2)$. This is sometimes referred to as 'inverse quadratic' because the inverse height is a linear function of inverse squared time, $1/h(t) = \beta_0 + \tau^2/t^2$. The parameterization is such that $h(\tau) = \frac{1}{2}h(\infty)$, so τ is the semi-max age, which is approximately 13 days.

The growth trajectories are plotted against calendar time in the top panel of Fig. 4.1, and against plant age in the lower panel. The graphs give the impression that the number of plants is no more than 30, but there are in fact 69 distinct growth trajectories for 70 plants. The illusion is caused in part by heights being rounded to 1mm, so that, at any given time, there are usually fewer than 20 distinct heights.

Two strains of plant are included in the study, the first 40 called 'cis' and

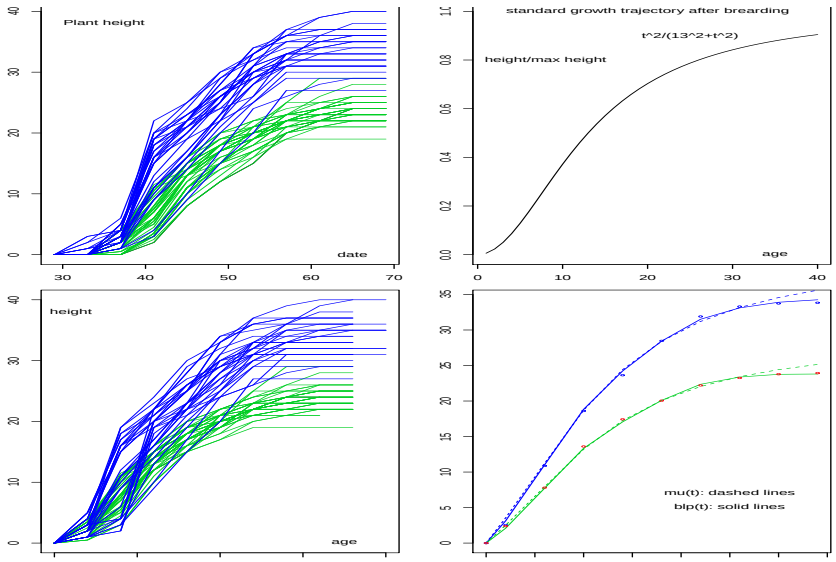

Figure 4.1: Heights in mm of 70 *Arabidopsis* plants of two strains, plotted against calendar time in panel 1, and against age in panel 3 (lower left). Lower right panel shows the fitted mean functions (dashed lines) together with the best linear predictor (solid lines) of plant height for each strain.

the remaining 30 labelled '108'. One goal of this project is to compare the two strains and to assess the significance of the observed differences. The time series plot for all plants in Fig. 4.1 reveals that both types have similar growth trajectories, but that the ultimate height of the '108'strain is about 40% greater than the 'cis' strain. The age-specific ratio of sample mean heights '108'/'cis' for plants aged 4–32 days is

Age in days	4	8	12	16	20	24	28	32
108/cis ratio	1.06	1.39	1.37	1.36	1.42	1.44	1.43	1.42

which is remarkably constant from day 8 onwards.

4.2 Growth curve models

The growth curve for plant i is modelled as a random function $\eta_i(t)$ whose value at age zero is, in principle at least, exactly zero, and whose temporal trajectory is continuous. In the analyses that follow, $s(i)$ is the strain of plant i, the mean trajectory is $\beta_{s(i)}h(t)$ with $h(0) = 0$ and $h(\infty) = 1$, so that the plateau levels β_0, β_1 depend on the strain, and the ratio of means is constant over time. The observation $Y(t) = \eta(t) + \epsilon(t)$ is contaminated by measurement error, which is assumed to have mean zero with constant variance σ_0^2, and to be independent

for all times $t > 0$ and all plants. The error distribution for values reported as zero need not be the same as the error distribution for positive measured values.

Brownian motion (BM) starting from zero at time $t = 0$ is a continuous random function with covariance function $\mathrm{cov}(B(t), B(t')) = \min(t, t')$. We are thus led initially to consider the additive Gaussian model with moments

$$\begin{aligned} E(Y_i(t)) &= \beta_{s(i)} h(t), \\ \mathrm{cov}(Y_i(t), Y_j(t')) &= \sigma_0^2 \delta_{ij} \delta_{t,t'} + \sigma_1^2 K(t, t') + \sigma_2^2 \delta_{ij} K(t, t') \end{aligned} \tag{4.3}$$

where $K(t, t') = \min(t, t')$ for the Brownian-motion model.

This formulation is incompatible with baseline information available at planting, which consists solely of two covariates, `date` and `strain`. Two deviations are noted. First, the units at planting are seeds, not plants. Second, it is most unlikely that every seed germinates, so the set of germinated plants is a random subset of planted seeds, which is not available at planting. This implies that `plant_id` is not a baseline factor nor is the germination date a baseline covariate. Formulation (4.3) with plants as observational units and t representing plant age, is compatible with baseline taken to be the plant-specific date of brearding. Nonetheless, as the discussion in the preceding section shows, the determination of that date is slightly inaccurate.

One objection sometimes raised to Brownian motion as a model for a growth curve is that it is not suficiently smooth—in fact nowhere differentiable. If a compelling argument could be made that physical growth is a differentiable function, one would have to reconsider the Brownian-motion model, perhaps replacing it with a smoother random function or a family of random functions having varying degrees of smoothness. But in the absence of a compelling demonstration of differentiability, the lack of differentiability of BM is not a strong argument against its use for growth curves. The Brownian motion component of the model can be replaced by any continuous random function deemed suitable, such as fractional Brownian motion (FBM), and the data can be permitted to discriminate among these. Despite the perception that physical growth curves are smooth in time, trajectories smoother than BM are firmly rejected by the data in favour of rougher trajectories. See variation (iii) below.

Leaving aside the measurement error component, the growth-curve model (4.3) has two additional variance components, one Brownian motion with volatility coefficient σ_1 that is common to all plants regardless of strain, and another with coefficient σ_2 that is plant-specific and independent for each plant. In other words, for $t > 0$ the measured value on plant i is a sum of one non-random term and three independent random processes

$$Y_i(t) = \beta_{s(i)} h(t) + \epsilon_{it} + \sigma_1 \eta_0(t) + \sigma_2 \eta_i(t), \tag{4.4}$$

where $\eta_0, \eta_1, \ldots, \eta_{70}$ are independent and identically distributed Brownian trajectories starting from zero at time zero. In this model, the variances

$$\begin{aligned} \mathrm{var}(Y_i(t)) &= \sigma_0^2 + (\sigma_1^2 + \sigma_2^2) t \\ \mathrm{var}(Y_i(t) - Y_j(t)) &= 2\sigma_0^2 + 2\sigma_2^2 t \end{aligned}$$

are both increasing linear functions of plant age.

In (4.3), the average height at age t of a very large number of plants of strain s is $\beta_s h(t) + \sigma_1 \eta_0(t)$. This infinite average is not a deterministic function of t; it is a Gaussian process with mean $\beta_s h(t)$ and covariance $\sigma_1^2 K(t, t')$. That is to say, $\sigma_1 \eta_0(t)$ is the deviation of $\beta_s h(t)$ from the mean trajectory averaged over infinitely many plants of strain s. From the fitted model, the estimated, or predicted, plant height trajectory $E(Y_{i^*}(t) \mid \text{data})$ for a new plant i^* is shown for both strains in the fourth panel of Fig. 1. Each fitted trajectory is the sum of the fitted mean $\hat{\beta}_s h(t)$ plus the conditional expected value of $\sigma_1 \eta_0(t)$ given the data. The latter term $E(\eta_0(t) \mid \text{data})$ is linear in the data; as a function of t it is a C^∞-spline with knots at observation times, i.e., continuous at all points and differentiable at all non-observation times.

Only the 628 response values at strictly positive plant ages are included in the likelihood computations, the heights at $t \leq 0$ being exactly zero by construction. For the mean model, $\hat{\tau} = 12.842$ days is used throughout. The three variance-components estimated by maximum residual likelihood are

parameter	estimate	$S.E.$
σ_0^2	1.040	0.151
σ_1^2	0.066	0.042
σ_2^2	0.432	0.052

with asymptotic standard errors as indicated. Asymptotic standard errors of variance components are worth reporting, but are often less reliable as indicators of significance than standard errors of regression coefficients. The first coefficient implies that the standard deviation of the measurement error is around 1mm, which is about right for laboratory measurements of plant height. The small value of σ_1^2 implies that $h(t)$ is a close approximation to the mean trajectory averaged over plants, and the relatively large standard error suggests that this term may be unnecessary. Nevertheless, the reduced model with only two variance components is demonstrably inferior: the increase in residual log likelihood is 13.78, i.e., the likelihood ratio chi-squared statistic is 27.56 on one degree of freedom. In this instance, the comparison of $\hat{\sigma}_1^2$ with its asymptotic standard error gives a misleading impression of the significance of that term.

The regression parameters governing the mean, τ included if necessary, are estimated by weighted least squares. For the 70 plants in this study, the plateau estimates in mm for the two strains are as follows:

strain	coefficient	$S.E.$
cis	28.35	1.76
108	40.13	1.92
108 − cis	11.86	0.99

Although this is a two-sample comparative design, the variance of the 108/cis-contrast estimate is substantially less than the sum of the two variances.

The Box-Tidwell method has been used here for the calculation of standard errors to make allowance for the estimation of τ. (The unadjusted standard

errors are 1.54, 1.58 and 0.90 respectively.) The parametric bootstrap is a viable alternative. This analysis makes it plain that the difference between the ultimate heights of the two strains is not zero.

4.3 Technical points

Non-linear mean model with variance components
The inverse quadratic model for the mean growth curve

$$\mu_{it} = E(Y_{it}) = \beta_{s(i)} t^2/(\tau^2 + t^2)$$

has three parameters to be estimated, the two asymptote heights β_0, β_1 and the temporal scale parameter τ. Two options for the estimation of parameters are available as follows.

The most straightforward option is to use ordinary maximum likelihood (not REML) for the estimation of all parameters jointly. Since the model for fixed τ is linear in β, this can be done by computing the profile likelihood for a range of τ values, say $12 \leq \tau \leq 14$ in suitably small steps, and using the `kernel=0` option in `regress()` as follows.

```
h <- age^2/(tau^2 + age^2)
fit0 <- regress(y~h:strain-1, ~BM+BMP, kernel=0)
fit0$llik
```

Although all ages used in the computation are strictly positive, the model formula is such that the mean height at age zero is exactly zero. We find that the log likelihood is maximized at $\hat{\tau} \simeq 12.782$. A plot of the profile log likelihood values against τ can be used to generate an approximate confidence interval if needed: the 95% limits are approximately $(11.7, 14.2)$ days.

A follow-up step is needed in order for the standard errors of the β-coefficients to be computed correctly from the Fisher information. To compensate for the estimation of τ, the derivative of the mean vector with respect to τ at $\hat{\tau}$ must be included as an additional covariate, as decribed by Box and Tidwell (197?)

```
deriv <- -2*tau * fit$fitted * h / age^2
fit0a <- regress(y~deriv+h:strain-1, ~BM+BMP, kernel=0)
```

It is a property of maximum likelihood estimators for exponential-family models that the residual vector $y - \hat{\mu}$ is orthogonal to the tangent space of the mean model (with respect to the natural inner product $\hat{\Sigma}^{-1}$). Consequently, the coefficient of `deriv` at $\hat{\tau}$ is exactly zero by construction, and all other coefficients β, σ^2 are unaffected. The ordinary maximum-likelihood estimates of the variance components are (1.0467, 0.0497, 0.4283), the plateau coefficients are $(28.293, 40.042)$ mm, and the standard error of the difference is 0.975. In this instance, the unadjusted standard error is 0.941, so the effect of the adjustment is not great.

The second method is closer in spirit to REML, where the variance components are estimated from the residual likelihood, i.e., the marginal likelihood

based on the residuals. The mean-value model has a three-dimensional tangent space at τ,

$$\mathcal{X}_\tau = \text{span}\{\partial\mu/\partial\beta_0, \partial\mu/\partial\beta_1, \partial\mu/\partial\tau\} = \text{span}\{hS_0, hS_1, deriv\}$$

where S_r is the indicator vector for strain r. The aim is to find $\hat{\tau}$ such that $\mathcal{X}_\tau$, is orthogonal to the residual vector. The only difference between this procedure and maximum likelihood is that the variance components, which determine the inner product matrix, are estimated by maximizing the residual likelihood rather than the full likelihood. If we fix τ, h and `deriv` as before, the command

```
fit0b <- regress(y~deriv+h:strain-1, ~BM+BMP)
```

uses the default kernel $\mathcal{K} = \mathcal{X}_\tau$ in the estimation of the variance components using REML, and $\tau = 12.842$ is such that the coefficient of *deriv* is zero. The estimated variance components are $(1.0400, 0.0659, 0.4319)$, the plateau coefficients are $(28.353, 40.138)$, and the standard error of the difference is 0.988.

When an iterative function such as `regress()` is used repeatedly in a loop as above, the overall efficiency can be substantially improved by supplying the previously-computed vector of variance components

```
fit0 <- regress(y~..., ~BM+BMP, start=fit0$sigma)
```

Estimated variance components may be negative provided that the combined matrix is positive definite. The argument `pos=c(0,1,1)` can be used to force positivity on selected coefficients.

Fitted and predicted values

The mean functions for the two strains are $\beta_0 h(t)$ and $\beta_1 h(t)$, and the fitted curves with β_s replaced by $\hat{\beta}_s$ are shown as dashed lines in the lower right panel of Fig. 1. The fitted mean is not to be confused with the predicted growth curve for an extra-sample plant i^* of strain s, which is deemed to have a response

$$Y_{i^*}(t) = \beta_s h(t) + \sigma_1\eta_0(t) + \sigma_2\eta_{i^*}(t) + \epsilon_{i^*t}.$$

Although this new plant is not one of those observed in the experiment, its growth trajectory is not independent of the sample responses because the covariances

$$\rho_t(i, t') = \text{cov}(Y_{i^*}(t), Y_i(t')) = \sigma_1^2 \,\text{cov}(\eta_0(t), \eta_0(t')) = \sigma_1^2 K(t, t')$$

are not all zero. The conditional distribution given the data is Gaussian with conditional mean

$$E(Y_{i^*}(t) \mid \text{data}) = \beta_s h(t) + \rho_t' W(y - \mu) \tag{4.5}$$

where ρ_t is the n-component vector of covariances, μ is the n-component vector of means, and $W = \Sigma^{-1}$ is the inverse covariance matrix of the response values for the sampled plants. The fitted conditional distribution, or the fitted

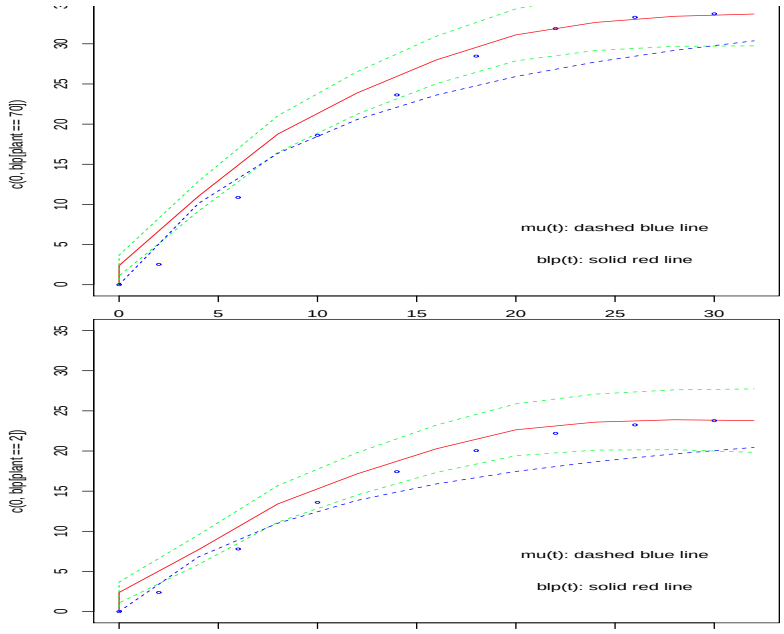

Figure 4.2: Fitted mean growth curves (dashed lines) and best linear predictors (solid lines) of plant height for two strains, using an inverse linear model for the mean trajectory and Brownian motion for the deviations. Sample average heights at each age are indicated by dots.

predictive distribution, has a mean $\hat{\beta}_s h(t) + \hat{\rho}_t' \hat{W}(y - \hat{\mu})$, called the best linear predictor (BLUP). This is shown as a pair of solid curves in the lower right panel in Fig. 1, one curve for each strain.

If we had taken the plant age to be two days rather than four at the time of the first positive measurement, and $h(t) = t/(\tau + t)$ to be linear at the origin rather than quadratic, the graphs of fitted means and best linear predictors would look rather different: see Fig. 2. Even with the reduced two-day offset for the origin, the inverse linear function is less satisfactory as a description of the growth curve than the inverse quadratic, so the variance coefficient σ_1^2 needs to be increased by a factor of roughly 7.3 to compensate for the larger deviations. Using log likelihood for model comparison, the inverse linear model is decisively rejected. Despite the less satisfactory fit, the best linear predictor for the inverse linear model (shown as the pair of solid lines in Fig. 2) is not appreciably different from the best linear predictor for the inverse quadratic model in Fig. 1. The maximum difference is approximately one millimetre (or 3%) at age 40 days. The difference between fitted means is much larger.

In certain areas of application such as animal breeding, the chief goal is to make predictions about the meat or milk production of the future progeny of a specific individual bull. This bull is not an extra-sample individual, but one of those experimental animals whose previous progeny have been observed and measured. Such predictions are seldom called for in plant studies. Nevertheless, from a probabilistic viewpoint, the procedure is no different. If i^* is one of the sampled plants and t is an arbitrary time point, the covariance of $Y_{i^*}(t)$ and $Y_i(t')$ is

$$\rho_{i^*t} = \sigma_1^2 K(t, t') + \sigma_2^2 \delta_{i,i^*} K(t, t'),$$

which involves two of the three variance components. The conditional expected value (4.5) then yields an interpolated curve for each plant.

4.4 Modelling strategies

1. Choice of temporal origin. The distinction between calendar time and plant age is fundamental. The decision to measure plant age relative to the time of brearding is crucial, and has a greater effect on conclusions than any subsequent choice.

2. Selection of a characteristic mean curve. The mean curve must pass through the origin at age zero, so a logistic function $e^t/(1 + e^t)$ cannot be used. The graphs in Fig. 4.1 suggest an inverse quadratic curve, which may or may not be appropriate for other plants.

3. Use of a non-stationary covariance model. Plant growth curves are intrinsically non-stationary because they are tied to the origin at age zero. Animal growth curves using weight in place of height are not similarly constrained.

4. Brownian motion. It seems reasonable that every growth curve should be continuous. It seems reasonable also to model the response as a sum of the actual height plus measurement error, thereby making a distinction between plant height and the measurements made at a finite set of selected times. The particular choice (BM) is not crucial, and can be improved by FBM. It is also possible to mix these by using FBM for the plant-specific deviation, and BM for the common deviation, or vice-versa.

5. Positivity. Plant heights are necessarily positive at all positive ages, whereas any Gaussian model puts positive probability on negative heights. This is one of those compromises, some major, some minor, that are frequently needed in applied work.

6. Response transformation, usually $y \mapsto \log(y)$, is an option that must always be considered. The log transformation might be reasonable for animal growth curves, but it was rejected here because of the role of zero height in determining the age origin.

7. Limiting behaviour. Plants do not grow indefinitely or live for ever, so the capacity of the growth model for prediction is limited to the life span of a typical plant.

8. Other issues. The emphasis on growth curves overlooks the possibility that the two strains may differ in other ways. In fact, the average brearding time for strain '108' is two days less than the time for strain 'cis', with a standard deviation of 0.43 days. No single summary tells the whole story.

4.5 Miscellaneous R functions

The following is a list of various R functions used in the construction of covariance matrices, and in the fitting of variance-components models.

```
BM <- outer(age, age, "pmin")      (BM covariance matrix)
FBM <- outer(age^p, age^p, "+") - abs(outer(age, age, "-"))^p
Plant <- outer(plant, plant, "==")      (plant block factor)
BMP <- BM * Plant      (component-wise matrix multiplication)
FBMP <- FBM * Plant      (iid FBM for each plant)
mht0 <- tapply(height[strain==0], age[strain==0], mean)
mht1 <- tapply(height[strain==1], age[strain==1], mean)
tapply(brearded, strain, mean)
L <- t(chol(FBM))   (Choleski factorization)
fit <- regress(y~h:strain-1, ~BM+FBMP, kernel=0)
```

Computer files

```
PlantGrowth.dat  PlantGrowth.R
```

4.6 Exercises

4.1 In the inverse quadratic model, the height of plant i at age t is Gaussian with mean $\beta_{s(i)}h(t)$ whose limit as $t \to \infty$ is $\beta_{s(i)}$. What is the variance of the ultimate height of plant i?

4.2 For the inverse linear model in which brearding is deemed to have occurred two days prior to the first positive measurement, estimate τ together with the plateau coefficients. Obtain the standard error for the estimated limiting difference of mean heights for the two strains.

4.3 The Brownian motion component of the model can be replaced with fractional Brownian motion with parameter $0 < \nu < 1$, whose covariance function is

$$\operatorname{cov}(Y(s), Y(t)) = s^{2\nu} + t^{2\nu} - |s-t|^{2\nu},$$

where $s, t \geq 0$. The index ν is called the Hurst coefficient, and $\nu = 1/2$ is ordinary Brownian motion. Show that the fit of the plant growth model can be appreciably improved by taking $\nu \simeq 1/4$.

4.4 Bearing in mind that the heights are measured to the nearest millimetre, comment briefly on the magnitude of the estimated variance components for the FBM model.

4.5 In the fractional Brownian model with $\nu < 1/2$, the temporal increments for non-overlapping intervals are negatively correlated. Suggest a plausible mechanism that could lead to negative correlation.

4.6 For 1000 equally spaced t-values in $(0, 10]$ compute the FBM covariance matrix K and its Choleski factorization $K = L'L$. (If $t = 0$ is included, K is rank deficient, and the factorization may fail.) Thence compute $Y = L'Z$, where the components of Z are independent and identically distributed standard Gaussian, and plot the FBM sample path, Y_t against t. Repeat this exercise for various values of ν in $(0, 1)$ and comment on the nature of FBM as a function of the Hurst coefficient.

4.7 Several plants reach their plateau well before the end of the observation period. How is the analysis affected if repeated values are removed from the end of each series?

4.8 Explain the Box-Tidwell method.

4.9 Investigate the relation between brearding date and ultimate plant height. Is it the case that early-sprouting plants tend to be taller than late-sprouting plants?

Example 5

5.1 Evolution of lice on captive pigeons

5.1.1 Backround

Understanding the mechanisms responsible for the origin of new species is a fundamental topic in evolutionary biology that has been the focus of numerous experiments and much speculation dating back at least to Darwin, who argued that differential natural selection in a range of environments leads to reproductive isolation and thence, eventually, to the formation of new species. Chapter 3 is concerned with speciation induced by differential diets over approximately 40 generations of *Drosophila*. This chapter considers another experiment on the same theme, but with a different system and different environmental pressures.

The paper *Rapid experimental evolution of reproductive isolation from a single natural population* published by Villa *et al.* (PNAS 2019) is concerned with reproductive isolation developing in response to body-size evolution in isolated lineages of pigeon lice. Each lineage evolved over 60 generations on a different host pigeon. Half of the hosts were normal-sized captive feral pigeons, the other half were giant runts.

To establish their claim, the authors must show evidence of two phenomena: first that louse size evolves rapidly in giant runt hosts relative to that in captive feral hosts, and second that differential louse size induces sexual isolation. The evidence for both of these phenomena is essentially statistical. The mechanism by which size differences lead to reproductive isolation is important from an evolutionary standpoint, but this chapter deals only with size evolution, i.e., whether systematic louse-size changes are detectable in a 60-generation span. Our concern is not so much with the evolutionary implications of the authors' findings, but with the experimental design, the data analysis, and the inferences that follow. The goal is solely to examine the data for evidence of systematic body-size changes in response to host size.

5.1.2 Experimental design

The following synopsis of experimental procedure is taken directly from Villa *et al.* (2019). Before the start of the experiment, resident lice on all experimental pigeons were eradicated by housing the birds in low-humidity conditions for

at least ten weeks. According to the authors, this procedure kills both lice and eggs, while avoiding residues from insecticides. To begin the experiment, 800 lice taken from wild-caught feral pigeons were transferred to 32 lice-free experimental pigeons, 25 lice per host. Half of the experimental hosts were captive feral pigeons; the other half were giant runts, a domesticated breed that is threefold larger than feral pigeons. Pigeons were housed in eight aviaries, each aviary containing four birds of the same breed. Every six months, a random sample of lice from each bird was removed, photographed, and returned to the host. The sex, body length, metathorax width, and head width of each louse was recorded.

One aspect of this design is different from that in chapter 3. After measurements were made, the lice were returned to their host. This was done in order to minimize the effect of measurement on the host-parasite system. Otherwise, the act of measurement would reduce the resident population, and introduce instability in the lineage, which is not desirable. In the design in chapter 3, the flies removed for experimental purposes were reared separately for one generation on a standard diet, so it was not possible to return them to the main breeding line. However, the breeding lines were more tightly controlled, so plans could be made in advance to accommodate the numbers needed in any particular generation.

As always in situations of this sort, the phrase 'random sample of lice from each bird' must be treated with caution, particularly with regard to size measurements. Larger lice are more visible than smaller specimens, so it would be naive to expect the random sample to behave like a simple random sample of the resident lice on a given bird. Nonetheless, size-biased sampling need not be a serious concern for this experiment provided that it affects all birds equally.

5.1.3 Deconstruction of the experimental design

Since each measurement is made on one louse, it is evident that each observational unit is either one louse or one louse on one occasion, while the response Y_u is a point in the state space, which is $\{M, F\} \times \mathbb{R}^3$. Since a louse generation is approximately 24–25 days, and measurement occasions are six months apart, we can be sure that no louse was measured on more than one occasion. While there is no practical distinction between louse and louse-occasion as the observational unit, as a matter of principle the ordered pair is the correct choice.

The lice are arranged in 32 lineages, one lineage to each bird. Thus *lineage* and *bird* are equivalent as block factors, and *aviary* is a coarser partition or block factor with eight levels. With respect to birds, *host* or *breed* is a binary classification factor.

The baseline is the time at which de-lousing was complete, and the experiment was ready to commence with new lice lineages on captive birds. The paper mentions randomization only incidentally in the 'Materials and Methods' section, and the reference there is a little ambiguous, but two crucial choices appear to have been made at baseline. First, the 800 initial lice were partitioned into 32 lineages with 25 founders for each lineage. Second, each lineage was associated with a particular bird. Regardless of the biological and mechanical

constraints in the laboratory, it seems reasonable and mathematically natural to regard each of these steps as the outcome of an independent uniform randomization scheme. Since the objective is to study selective pressure, host size is the principal treatment. If the randomization was done in two steps as indicated, treatment is assigned to lineages in step two, in which case each lineage serves as one experimental unit.

By definition, a covariate is a pre-baseline variable, and it appears that there is only one. Measurement occasion or *time* is a function on the observational units, which is a quantitative factor. However, as indicated in the preceding paragraph, lineage could be regarded as a pre-baseline block factor, and it should certainly be used as the experimental unit to assess variability of the treatment effect estimate.

In addition to *time* and *lineage*, pre-baseline vital measurements including louse *sex* are available on the 800 founder lice. All pre-baseline variables are available for use as covariates as if the values were fixed and non-random, and initial response values are no different in that respect from any other pre-baseline measurements. Randomization ensures that the distribution of initial values is the same for all treatment groups, so the initial response values are uninformative for treatment effects. Generally speaking, when the response is a time series or temporal process, it is more convenient and mathematically more natural to treat initial response measurements as an integral part of the response process. A crucial point is that the probability model for the response at $t = 0$ must be consistent with the randomization: see sections 5.2.3, 5.2.5 and 11.4.5. The joint distribution implies a conditional distribution, which is available if needed for purposes of estimation or prediction.

The paper does not discuss how birds were assigned to aviaries, but it seems reasonable to regard that too as the outcome of a balanced randomization applied to birds, subject to restrictions mentioned earlier. We presume here that birds were quarantined in their aviaries during de-lousing, in which case *aviary* is a pre-baseline block factor. Since all birds in one aviary are of the same breed, a strong argument can be made that *aviary* is the experimental unit, not *lineage* as stated earlier. Both seem to be relevant. Whether they are pre-baseline or immediately post-baseline, *time*, *lineage*, *aviary*, and *treatment* are available for purposes of analysis and model construction.

Apart from the founders, louse sex is a post-baseline variable, and thus one of four components in the response. Genetic theory leads one to expect the sex ratio should steady at 50:50, and post-baseline counts in Table 5.3 confirm this. But the same table also shows that the baseline F:M ratio is 464:336, which is significantly in excess of 50:50.

Each lineage was associated with a particular pigeon at baseline, which means that *lineage* and *pigeon* are equivalent as block factors. A subsequent remark in the paper shows that this statement is not quite correct. When a bird died during the experiment, all lice from the dead bird were transferred to a new parasite-free bird of the same type. Thus, one lineage could span two or more birds. Unfortunately the data file does not indicate when deaths might have occurred, so we have no way to check the effect on lineages of host transfers.

5.2 Data analysis

5.2.1 Role of tables and graphs

However it is measured, the response of evolutionary interest is louse size. To keep matters as simple as reasonably possible, we focus on the single response, body length. Since we plan to use additive decompositions, the log transform is more or less automatic, so Y_u is the log body length for louse u. However, the range of variation in all size measurements is only a few percent of the average, so the log transformation has little effect on conclusions.

The purpose of a table or graph is to advance the narrative thread by drawing attention to the most important patterns or features in the data such as the nature and direction of various effects. It is natural enough to emphasize the effects of scientific interest—but not at the cost of misleading the reader. Every table or graph invites the question 'What is the point of this table?' or 'What feature does this graph illustrate?'. If the answer is not clearly apparent, the narrative is not advanced, and the reader is likely to be confused. Generally speaking, the data analyst examines numerous tables and graphs. Only the most useful of these are retained for presentation.

Table 5.1. Average log body length (in μm) of lice on two pigeon hosts

		Time in months								
Sex	Host	0	6	12	18	24	30	36	42	48
F	Feral	7.883	7.883	7.883	7.874	7.866	7.886	7.880	7.872	7.864
F	G.R.	7.885	7.894	7.882	7.882	7.882	7.895	7.894	7.899	7.886
M	Feral	7.720	7.716	7.705	7.700	7.702	7.712	7.709	7.713	7.700
M	G.R.	7.720	7.718	7.717	7.716	7.713	7.723	7.726	7.731	7.720
		Differences ×100: Giant runt − Feral								
F	G−F	0.2	0.1	−0.1	0.8	1.7	0.9	1.4	2.6	2.2
M	G−F	0.0	0.2	1.2	1.7	1.1	1.1	1.7	1.8	2.0

The first four rows of Table 5.1 show the average log body length of all lice measured on each occasion. Most impressive is the stability of body length for both louse sexes over 60 generations. If anything, there is a slight decrease in length for lice on both hosts, with a slightly greater decrease for captive feral pigeons.

The numbers in Table 5.1 are accurate to three decimal places or four decimal digits, but the first three digits are essentially constant at 7.88 for females and 7.72 for females, so we say that there are only 1–2 significant decimal digits. Usually, one is not enough to gauge accurately the statistical variation in the process. However, we have chosen to leave the table in its present form to emphasize how tiny are the size differences between lice on the two hosts.

The sex difference $7.88 - 7.72 = 0.16$ means that female lice are about 16% longer than males. The last two rows show that the mean difference for hosts tends to increase over time, reaching around 2% for both sexes after 48 months.

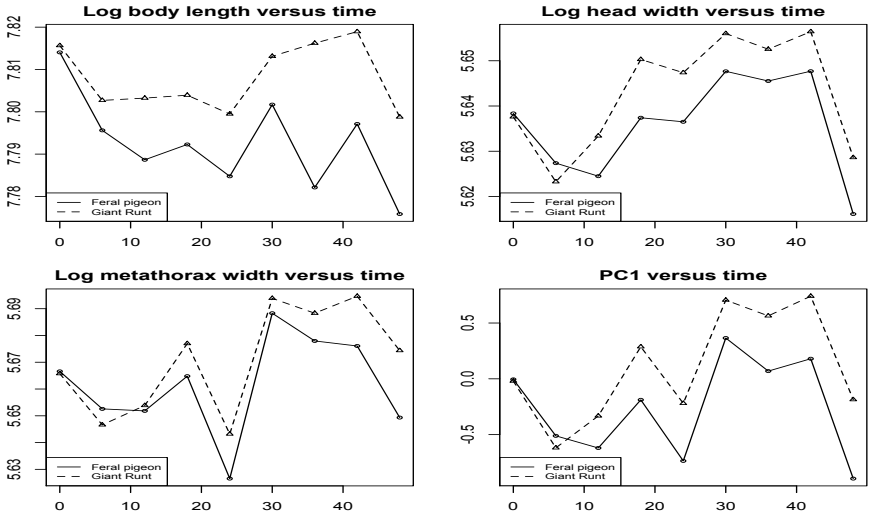

Figure 5.1: Average body sizes of lice for two hosts over time

It is remarkable that such a small size difference could have a detectable effect on sexual coupling.

The first panel of Figure 5.1 shows a plot of the same data with sexes combined. Automatic centering and re-scaling of the y-axis has the effect of exaggerating the variation and the magnitude of the divergence between the two groups. In other words, that which is emphasized by the table of averages is eliminated by the plot.

The remaining panels show similar plots for the head width, the metathorax width, and the first principal component, which is a roughly equally-weighted positive linear combination of the three standardized size variables. For all size variables, the temporal trajectory for louse size on giant runts is surprisingly similar to that for feral pigeons, and lice on giant runts are larger on average than those on feral pigeons. Apart from the uniform decrease in all size measurements in the initial and final intervals, no clear temporal trend is visible.

Ideally, it would be good to show error bars for every point. But size measurements for different lice on one pigeon are not independent, so honest error assessment is not straightforward. On balance, it is better to show no error bars than to show the naive default based on independence, which is misleading in this setting: contrast Fig. 5.2 with the table at the end of section 5.5.

5.2.2 Trends in mean squares

Table 5.1 and Fig. 5.1 illustrate temporal trends in average body size. To get a comparable impression of trends in variance, it is helpful to compute mean-squares associated with *louse sex*, *host size*, *aviary*, *lineage* and residuals at each of the nine time points.

Table 5.2. Trends in mean squares and variance components $\times 10^5$

	Time t in months								
MS	0	6	12	18	24	30	36	42	48
Host	12	340	190	996	1095	690	1024	3448	1506
Aviary	290	204	408	575	493	333	324	528	350
Lineage	83	85	85	94	87	80	100	79	152
Residual	111	52	60	68	56	52	56	57	64
	Variance-component estimates ($\times 10^5$)								
$\hat{\sigma}^2_{\text{aviary}}$	2.1	3.2	10.7	10.4	9.5	6.5	6.7	10.6	6.6
$\hat{\sigma}^2_{\text{lineage}}$	−1.1	2.8	3.3	3.8	2.8	2.2	5.5	0.3	12.1
$\hat{\sigma}^2_{\text{resid}}$	111.3	53.1	59.4	67.3	56.9	52.5	56.1	58.4	63.3

The dominant mean square is that for *louse sex* which starts off at 5.25 at baseline, drops to half that value at six months and decreases slowly to 1.64 at 48 months. For the other factors, the mean squares are shown in the top half of Table 5.2, together with the REML variance components for *aviary*, *lineage* and residual in the second half. For this fit, *host* and *sex* were eliminated as fixed effects, so the mean-squared residual does not coincide exactly with $\hat{\sigma}^2_0$.

Some of the following points are accommodated in subsequent analyses, but others are merely noted.

1. The residual variability at baseline is twice that on all subsequent occasions. One plausible explanation is that founder lice collected from wild pigeons are more variable in size than those resident on captive pigeons.

2. The lineage mean square is remarkably constant from baseline onwards. Relative to the residual mean square, it is below expectation at baseline, but not significantly so. After baseline, it is uniformly larger than the residual mean square, but not by a large factor.

3. The host mean square at baseline seems artificially low. There is strong evidence in the data, for example in the sex ratios, that the randomization scheme was more complicated that that depicted in the preceding section, so this may be a consequence of an effort to balance the randomization.

4. The between-aviary mean square at baseline is a little larger than expected from uniform random assignment: the F-ratio is 2.6, which is is at the upper 1.6% point of the reference null distribution.

5. Variance-component estimates on few degrees of freedom, such as those for aviary and lineage, have notoriously high variances.

The main issue to be addressed at this point is the size of the aviary mean square at baseline, and whether the mean square provides sufficient probative evidence to cast doubt on the randomization or to declare it inadequate or biased. The question is not whether the initial lice were labelled 1–800 and lots

drawn to determine which lice would be assigned to which birds, but whether the laboratory procedures actually employed are a reasonable facsimile of objective randomization. The only evidence before the court is shown in Table 5.1.

One traditional view is that the aviary mean square is selected for attention as the largest of three or four, so the p-value, or measure of extremity, is closer to 5%. That calculation tells us something, but it does not answer directly the question of interest to the court: 'Given the data, what is the probability that the allocation to aviaries was biased?' From another viewpoint in which sparseness prevails at odds level ρ, the odds against aviary bias given the mean-square ratio $F = 2.6$ on $6,367$ degrees of freedom are approximately $\rho\zeta_6(2.6)$, where $\zeta_6(2.6) = 3.81$. This calculation uses a modification for F-ratios of the sparsity argument in McCullagh and Polson (2018). The strength of the evidence is such that the initial presumption of innocence with probability $1/(1+\rho)$ is changed to $1/(1+\rho\zeta_6(2.6))$. For $\rho = 0.1$, which is not a strong prior presumption for this setting, the probability of a no aviary bias is changed by the evidence from 0.91 to $1/1.38 = 0.72$. So we take note and proceed with caution, giving the randomization a provisional pass. This point is revisited in section 5.4.2.

5.2.3 Initial values and factorial subspaces

If host size has an effect on louse size, it is an evolutionary development, so the effect is not immediate. Thus, *treatment* and *time* are the principal covariates whose effects are to be studied. In addition, the body length for *C. columbae* male lice is approximately 85% of that for females, so louse sex must also be taken into consideration. The effects of *lineage* and *aviary* are assumed to be additive random variables with independent and identically distributed components for each pigeon and each aviary respectively. Since their effects are additive zero-mean random variables, *lineage* and *aviary* do not contribute to the mean-value subspace.

Setting the two block factors aside temporarily, the factors *treatment* or *host size*, *time* and *sex* are to be taken into account. If we proceed to use factorial models in the naive manner, we may begin with all three main effects and check which interactions are needed. Or we may follow the authors' practice in their Tables S2–S5, which is to report the coefficients in the full three-factor interaction model. Both approaches are technically incorrect. Fitting either of the suggested models is a pointless exercise that serves only to confuse the narrative thread for this experiment.

The problem with the naive application of factorial models to this design lies in the role of time, and the fact that $t = 0$ corresponds to the experimental baseline. If Y_{ut} denotes the log body length of louse u at time t, the additive main-effects model for the conditional mean given sex and host has the form

$$E(Y_{ut} \mid s, h) = \beta_0 + \beta_1 t + \beta_2 s(u) + \beta_3 h(u),$$

in which $h(u)$ is a code for the host size, and $s(u)$ is the louse sex. At baseline,

the additive model implies

$$E(Y_{u0} \mid s, h) = \beta_0 + \beta_2 s(u) + \beta_3 h(u),$$

with three coefficients to be estimated. The presumption of randomization, which is that lice are assigned to hosts independently of their size, implies $\beta_3 = 0$. Thus, randomization contradicts both the additive model and any other factorial model that contains it as a subspace.

Whether or not randomization was explicitly employed in this experiment, it is reasonable to imagine or suppose that the initial assignment of lice was effectively randomized. Randomization has implications. The use of a model that contradicts those implications is a source for confusion; the use of a model that conforms with randomization is strongly advised.

The phenomenon described here—of time in relation to treatment and initial values—is not new. A simple example is given in Exercise 3.11 of McCullagh and Nelder (1987).

Only the most cynical reader would seriously consider the possibility that the researchers had deliberately assigned the lice differentially to hosts or to aviaries in an inappropriate manner. However, there might well be sound biological arguments for balancing the design in certain ways or for favouring females in the establishment of lineages. Deviations of this sort are normal practice, but they should be reported. Nonetheless, unintentional biased assignment can occur, so it is routine in many areas of application to check whether the baseline values are in conformity with randomization. That can be done here. While there is no indication of bias in Fig. 5.1, randomization implies that the mean squares for *aviary*, *lineage* and residual have the same expected value at baseline. However, the aviary-to-residual mean-square ratio in Table 5.2 is 2.61, which falls near the upper 98.5 percentile of the null distribution. This is not proof positive of bias, but it is a little troubling and calls for an explanation.

5.2.4 A simple variance-components model

The following linear models address directly the question that is of principal interest to an evolutionary biologist. Without straying from linearity in time, the null and alternative may be formulated as linear subspaces.

$$H_0: \quad E(Y_{ut}) = \beta_0 + \beta_1 t + \beta_2 s(u); \tag{5.1}$$

$$H_A: \quad E(Y_{ut}) = \beta_0 + \beta_{h(u)} t + \beta_2 s(u). \tag{5.2}$$

The model formulae `time+sex` and `host:time+sex` generate basis vectors for the two subspaces whose dimensions are three and four respectively. The alternative model has two linear trends in time, one for captive feral hosts $h(u) = 0$, and one for giant runts $h(u) = 1$.

For covariances, we start out following the authors' suggestion with three variance components

$$\operatorname{cov}(Y_u, Y_{u'}) = \sigma_0^2 \delta_{u,u'} + \sigma_1^2 \delta_{l,l'} + \sigma_2^2 \delta_{a,a'}, \tag{5.3}$$

where l, l' and a, a' are the lineages and aviaries respectively. This is a linear combination of three identity matrices, one on the lice, one on the lineages or pigeons with 32 blocks, and one on the aviaries with eight blocks. It is usually justified either by appeal to exchangeability based on recorded similarities of observational units, or, if that argument fails to convince, by appeal to randomization. Although neither argument carries weight in this instance, computation is cheap so we proceed.

For the log body length, the REML variance components in (5.3) paired with (5.2) are

lice	$\hat{\sigma}_0^2$	78.19×10^{-5},
lineages	$\hat{\sigma}_1^2$	1.84×10^{-5},
aviaries	$\hat{\sigma}_2^2$	1.40×10^{-5}.

Both the lineage and aviary variance components are small relative to the between-lice variance. Despite that, there is no compelling reason to declare them null simply because they are small. The fitted slope coefficients ($\times 10^4$) for the two pigeon breeds are

Parameter	Estimate	s.e.
Feral:time	−2.23	0.53
Giant:time	1.37	0.38
Difference	3.60	0.63

This analysis appears to provide reasonably strong evidence that lice transferred to captive feral pigeons decrease in size over time, and moderately strong evidence that lice transferred to giant runts increase in size over time. However, the analysis is based on linearity in time, which seems implausible given Fig. 5.1, and a covariance structure (5.3) that is both inadequate for the data and in conflict with randomization.

5.2.5 Conformity with randomization

Randomization implies that the body-size measurements at $t = 0$ are exchangeable with respect to some group of permutations, here assumed to be large enough that the responses for every pair of lice have the same joint distribution regardless of whether they are assigned to the same pigeon, to different pigeons in the same aviary or to different pigeons in different aviaries. Unfortunately, randomization implies $\sigma_1 = \sigma_2 = 0$ in (5.3).

Ever since the pioneering work of Edwards and Cavalli-Sforza (1963, 1964), Brownian motion has been the standard probabilistic model for the neutral evolution of a quantitative trait (Felsenstein, 2004, chapter 23). The conflict with randomization can be fixed only by introducing non-stationary temporal processes for the lineage and aviary effects, and the most natural way to incorporate Brownian motion is as follows:

$$\operatorname{cov}(Y_u, Y_{u'}) = \sigma_0^2 \delta_{u,u'} + \sigma_1^2 K(t,t')\, \delta_{l,l'} + \sigma_2^2 K(t,t') \delta_{a,a'} + \sigma_3^2 K(t,t'). \qquad (5.4)$$

The Brownian covariance function $K(t,t') = \min(t,t')$ is positive semi-definite, and $K(0,0) = 0$ ensures conformity with randomization. Environmental selective pressure exerts a genetic drift, and the mean model (5.2) contains one drift parameter for each host, so the differential drift is the treatment effect.

The rationale for (5.4) is as follows. The louse population as a whole evolves as a Brownian motion with volatility σ_3; each aviary evolves independently as a Brownian motion with volatility σ_2; and each lineage evolves independently as a Brownian motion with volatility σ_1. For the duration of this experiment, each louse belongs to the system, a lineage and an aviary, and the value for the louse is the sum of these three processes plus white noise. All three processes are neutral or drift-free. Drifts associated with host size occur in (5.2).

The REML log likelihood achieved by this Brownian modification exceeds that for (5.3) by approximately 57.9 units, and all four fitted coefficients are positive. Although these models are not nested, the difference is huge enough to leave no doubt that (5.3) is totally inadequate for these data.

The effect of these temporal correlations on the fitted regression coefficients is small but not negligible; their effect on standard errors is an eight-fold increase. The fitted slope coefficients ($\times 10^4$) for the two pigeon breeds are

Parameter	Estimate	s.e.
Feral:time	−4.22	4.1
Giant:time	0.33	4.1
Difference	4.55	3.3

The conclusion from this analysis is the essence of simplicity: the data are entirely consistent with neutral evolution of louse size on both hosts.

Apart from the Brownian contribution, Table 5.2 shows that the baseline variance is substantially larger than the residual variance on subsequent occasions. This observation suggests that (5.4) is not adequate on its own, and must be supplemented by an additional diagonal matrix for baseline observations. This differential baseline variance leads to a further 64.7-unit increase in the REML criterion. However its effect on conclusions is almost negligible; for comparison, the fitted coefficients ($\times 10^4$) are as follows:

Parameter	Estimate	s.e.
Feral:time	−4.39	4.3
Giant:time	0.30	4.1
Difference	4.69	3.6

The conclusion regarding neutrality of evolution is unaffected. The apparent evidence for a differential trend in the analysis at the end of section 5.2.4 is a consequence of a demonstrably inadequate variance assumption.

Brownian motion in (5.4) does a reasonable job of describing the temporal dependence, but the fit can be improved by using a low-index fractional Brownian motion. However, this and other modifications discussed in section 5.4 and exercises 5.24–5.25 have only a small effect on drift estimates.

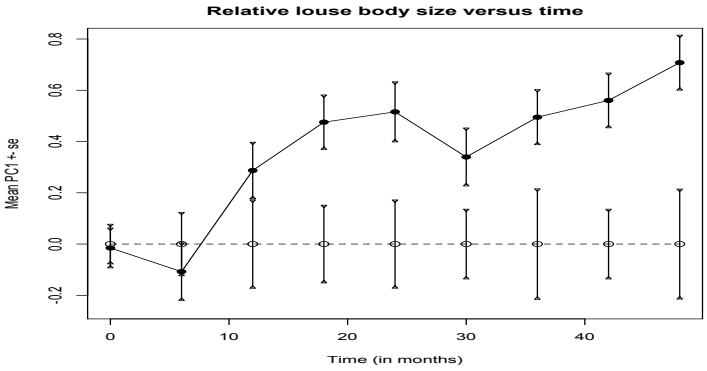

Figure 5.2: PC1 mean difference 'giant runt - feral' versus time

5.3 Critique of published claims

Villa *et al.* base their conclusions on the first principal component as a combined measure of overall louse size. Since the first principal component is essentially the standardized sum or average of the three size variables, this much is fine. The sample averages for each host are plotted in the fourth panel of Fig. 5.1, which shows that the divergence between the two mean trajectories is not appreciably greater than the temporal variability of any single trajectory. This is a disappointing conclusion for a four-year experiment, and not appealing as a headline story.

However, Villa *et al.* choose to emphasize the divergence over the variability by plotting the PC1 mean difference (giant runts minus controls) as a function of time in their Fig. 1C. A version of their plot is shown in Fig. 5.2, and is to be contrasted with the fourth panel of Fig. 5.1.

The plot symbol on the horizontal line in Fig. 1C or Fig. 5.2 is explicitly associated with controls. Error bars attached to zero are not mentioned in captions or in text. The visual impression of remarkable temporal stability of louse size on feral pigeons contrasts starkly with the rapid increase for lineages on giant runts. The plot title and the scale on the y-axis confirm those impressions, which are in line with the authors' conclusion *Lineages of lice transferred to different sized pigeons rapidly evolved differences in size.* In my opinion, Fig. 1C or Fig. 5.2 gives a grossly misleading impression of stability for feral pigeons contrasted with a substantial trend for giant runts. In fact, Table 5.1 shows that louse body-size changes are no more than 2% over the entire period.

Taking correlations into account, the error bars for the non-zero line in Fig. 1C or Fig. 5.2 are too small by a factor increasing from about 1.0 to 7.0, and roughly proportional to time.

Tables S2–S5 in the Appendix to their paper report regression coefficients and their standard errors for the full factorial model with (5.3) as the covariance structure. These tables are cited in the *Results and Discussion* section to

support the chief claim: *Over the course of 4 y, lice on giant runts increased in size, relative to lice on feral pigeon controls (Fig. 1C and SI Appendix, Tables S2–S5).* It is unclear which coefficients are meant to justify this claim, but the coefficient of *host:time* in the PC1 analysis is reported with a t-ratio of 3.15. Overlooked in this computational blizzard is the fact that both the fitted mean and the fitted covariance contradict the randomization. In addition, the covariance assumption is non-standard for an evolutionary process, and is demonstrably inadequate for the task.

The formal analysis of the first principal component by linear Gaussian models follows the lines of section 5.2.5. Although the scale of the PC1-response is very different from that of the body length, the need for the Brownian-motion component is abundantly clear, as is the additional baseline variance. When these covariances are accommodated, the slope estimates and their standard errors are

Parameter	Estimate	s.e.
Feral:time	−0.0126	0.033
Giant:time	0.0016	0.033
Difference	0.0142	0.013

Nothing in this PC1 analysis points to a departure from neutral evolution of lice on either host. In conclusion, the evolutionary divergence described by Villa *et al.* may well exist on some time scale, but the evidence for it is not to be found in their data.

5.4 Further remarks

5.4.1 Role of louse sex

The variables *host* and *lineage* are treatment factors generated immediately post-baseline by randomization, and having a known distribution. For the 800 lineage founders, louse *sex* is a a pre-baseline variable; for the remaining lice, *sex* is a random variable not generated by randomization, and not recorded immediately post-baseline. One can speculate on the joint distribution, but in principle, the sex ratio for giant runts might not be the same as the sex ratio for controls. Thus, (5.1) and (5.2) are models for the conditional mean while (5.3) and (5.4) are models for the conditional covariance—given *host* and *lineage* plus the entire *sex*-configuration for all sampled lice.

Regardless of covariance assumptions, the interpretation in (5.2) of β_h as 'the effect of treatment' must be considered in the light of the fact that any additive effect possibly attributable to an effect of treatment on *sex* has been eliminated. Although not intermediate in the temporal sense, *sex* is not dissimilar mathematically to an intermediate response. It is possible that treatment could have an effect on the intermediate response, in which case the coefficients β_h in the conditional mean describe only one part of the treatment effect.

In the context of this experiment, no effect of treatment on *sex* is anticipated. Any effect that might be present is most likely to be a sampling artifact of little

Table 5.3. Louse counts by host, sex and time

		Time in months								
Host	Sex	0	6	12	18	24	30	36	42	48
Feral	F	231	67	39	57	38	56	23	55	19
	M	169	73	44	50	37	55	31	49	22
Giant runt	F	233	95	104	105	102	104	105	105	96
	M	167	102	95	92	98	97	91	95	104

or no evolutionary interest. Nonetheless, it is not difficult to examine the sex distribution at baseline and post-baseline for both treatment groups. Table 5.3 shows the louse counts by time, host and sex.

The post-baseline total count is quite constant at 200 for giant runts, but is much more variable for captive feral pigeons. The first is presumably a design target. We are left to wonder why the the control group does not have a similar target. Nevertheless, this is not a serious criticism. In both treatment groups, females account for 58% of lice at baseline, but close to 50% thereafter. As anticipated, there is little evidence of a difference in sex ratio between groups. If anything, the difference between the ratios is below expectation at nearly every time point.

The Poisson log-linear model *time:(host+sex)* is equivalent to the statement that *host* and *sex* are independent at each time point, or equivalently, that the sex ratio is the same for both pigeon breeds, but not necessarily 50:50. The residual deviance of 2.8 on nine degrees of freedom falls at the lower third percentile (0.03) of the null distribution, which shows that sample log odds ratios are uniformly closer to constant than the Poisson model predicts. Certainly, there is no suggestion of a treatment effect on sex ratios. Apart from the imbalance at baseline, the subsequent ratios are close to 50:50, so we can regard the sex indicator post-baseline as a Bernoulli process independent of treatment.

5.4.2 Persistence of initial patterns

One unintended consequence of the Brownian covariance model (5.4) is that baseline values are independent of all subsequent values. This is a strong assumption. It is not implied by randomization, and it is not necessarily a property that we could confidently expect to be supported by detailed examination of the data. Without contradicting the randomization, it is possible to introduce temporal correlations between baseline and non-baseline values by a simple modification such as replacing the last term in (5.4) with the shifted Brownian covariance $\min(t-\tau, t'-\tau)$ for some $\tau \leq 0$. For reasons that are explained in chapter 17, the REML criterion is independent of τ, so this particular modification has no effect on fitted values, on prediction or inference for contrasts. In fact, this covariance term could be replaced with the stationary version $-|t-t'|/2$.

The analysis of variance for baseline values already casts doubt on the fair-

ness of the randomization with respect to aviaries, so it is natural to check for correlations between initial and subsequent values associated with the same aviary. Does the pattern of louse size differences among aviaries at baseline persist in subsequent generations? The question is concerned with persistence of aviary patterns, so fairness of the randomization is not presumed.

One way to introduce persistent initial patterns is to replace the aviary term in (5.4) with independent shifted Brownian motions, one per aviary. The covariance contribution is then

$$\sigma_2^2 \delta_{a,a'} \min(t - \tau, t' - \tau),$$

with a single temporal shift $\tau \leq 0$ to be estimated from the data using the REML criterion. One boundary point $\tau = 0$ coincides with (5.4), and the other limit $\tau \to -\infty$ implies a constant aviary effect as in (5.3). For $\tau < 0$, this modification implies positive correlations within aviaries at baseline, which is a size pattern that contradicts our understanding of randomization. The interpretation is that, by accident or by design, some aviaries start out with larger lice than others, and the initial pattern leaves an imprint on the subsequent evolution.

For the PC1 variable, the profile REML log likelihood values for τ at zero, $\hat{\tau} = -2.6$ and $-\infty$ are 0.0, 11.5 and -18.5, showing that the constant aviary effect is decisively rejected by the data. It appears from this analysis that the initial aviary pattern for PC1 is non-zero and that it persists in the subsequent evolution. The particular temporal offset may be pure coincidence, but $\hat{\tau} = -2.6$ months is a very close approximation to the de-lousing quarantine period during which the pigeons had to be housed somewhere.

5.4.3 Observational units

Consider the statement near the beginning of section 5.1.3: 'since each measurement is made on one louse, it is evident that each observational unit is one louse...'. The premiss—that each measurement is made on one louse—is indisputable. Nevertheless, a conclusion that is obvious literally, is not necessarily true mathematically in the sense of the definition.

According to the definition, the observational units are the objects, or points in the domain, on which the response is defined as a stochastic process. Thus, each observational unit exists at baseline, not necessarily as a physical object, but as a non-random mathematical entity. For the models in section 5.2, with louse-time pairs as observational units, there is no birth or death, and no evolving finite population—only a fixed, arbitrarily large, set of lice in each lineage. In this mathematical framework, the lice are in 1–1 correspondence with the natural numbers, they live indefinitely in the product space, and their vital statistics are random variables recorderd in the state space. To each louse there corresponds a stochastic process, so the value for each louse evolves over time, but the population itself is fixed and arbitrarily large in every lineage.

It would be wrong to say that the Gaussian model is incorrect or that its flaws are fatal, but its shortcomings for this application are clear enough. If

the application calls for a finite randomly-evolving lineage, a more complicated mathematical structure is required. The remarkable thing is not that this Gaussian model is exquisitely tailored to this evolutionary process, but that a generic model that is missing the defining aspects of life, namely birth and death, should have anything useful to contribute at all.

Certainly, the lice do not exist in the physical sense at baseline. But lineages are established at baseline, and it is the lineages that evolve. They evolve randomly in two senses—in their composition as a finite set of lice, and in their values or features. If both aspects are important for a given application, a more complicated model is needed in which the observational units are lineage-time pairs. The state space for one measurement on one louse is $\mathcal{S} = \{M, F\} \times \mathbb{R}^3$; the state space for one observational unit is the set of finite subsets of $\mathcal{S}$. One finite subset of $\mathcal{S}$ is a complete description of the population size and the vital statistics of the residents at time t. The transitions from one finite subset to another are limited by birth, death and continuity in time.

A general process of the type described in the preceding paragraph is a complicated mathematical structure, and we make no effort to develop a general theory here. But there are simple versions that are essentially equivalent to imposing a pure birth-death process independently as a cohort restriction on the domain of a Gaussian process. The distribution of the values thus generated coincides with the Gaussian model in section 5.2, and none of the subsequent analyses are affected. For that setting, birth and death are immaterial.

The possibility that individual louse values or body-sizes might be related to the sample size or lineage size from which they come has not been considered up to this point, in part because such a dependence is not possible under the models in section 5.2. The notion that a sample can be extended indefinitely from a sub-sample such that the sub-sample values remain unchanged, is usually understood in applied work as an obvious fact. The possibility that the obvious fact might fail is "such an appalling vista that every sensible person would say 'It cannot be right..."'. But, just as it turned out a decade after Lord Denning's notorious judgement in 1980, from which this quote is taken, the appalling vista is not sufficient probative evidence to establish its alternative as fact. Failure also strikes at the heart of the most cherished notion in probability and applied statistics, which is the 'obvious fact' of distributional consistency for sub-samples as formulated by Kolmogorov (1934). If lice are the observational units for this process, consistency implies that the distribution for individuals is unrelated to the size of the sample from which they are taken. Fortunately, variability of sample sizes provides a weak check to test that implication.

Each of the 32×9 lineage-time pairs provides one sample, of which 15 are empty. The louse counts range from zero to 44, they are highly variable, and they tend to decrease over time. One lineage appears to go extinct at 30 months. The safest and the simplest way to test for a dependence on sample size is to include sample size as an additional 'covariate' in (5.2), retaining (5.4) for covariances. For both log body length and PC1, the fitted coefficient is negative and approximately one half of the standard error. This analysis offers no evidence of a sample-size dependence, which provides a little reassurance that the

earlier analysis with louse as the observational unit is reasonably sound.

If some birds preened more vigorously or more thoroughly than others, and larger or older lice were preferentially removed by preening, the more assiduous preeners would then host fewer and smaller lice. Differential preening could lead to a dependence of mean louse size on lineage size or on sample size, in which case the test in the preceding paragraph is a reasonable check.

5.5 Follow-up

5.5.1 New design information

Given the severity of the discrepancy between the conclusions presented above and those published by Villa et al. (2019), it seemed only appropriate to send a copy of sections 5.1–5.4 to the authors for comment. I contacted the lead author in early December 2020. Scott Villa, responded immediately, and later at the beginning of February 2021 offering further details about the experimental design, and challenging the conclusions on several points.

The randomization was carried out according to an elaborate protocol, which involved dislodging the CO_2-anesthetized lice over a custom-made 10×14 glass grid, generating a random grid number as the starting point for collection of specimens, and placing lice sequentially and cyclically in vials labelled 1–32 until each vial contained 25+ lice. It was designed to avoid unintentional biases, and it appeared to be adequate for the task.

The following summary of key design points that had previously been partially or totally misunderstood is taken from Villa's reply.

1. At time zero, 1600+ lice were collected from wild feral pigeons. No size measurements were made on the sub-sample of 800 founder lice that were transferred to captive birds. A second sample of 800 lice was photographed, measured, and frozen for subsequent genetic analysis.

2. The 800 founder lice were assigned to hosts at random, 25 per bird. Each founding population consisted of 13–14 females and 11–12 males with a deliberate female bias to ensure that a lineage would be established on every host.

3. The 800 lice measured at time zero did not contribute to the breeding population; their assignment to lineages was randomized, but purely virtual. The virtual sample had the same sex-ratio as the founders.

4. After baseline, the lice that were measured at 6-month intervals were frozen thereafter to use for genomic analyses of the populations over time. Throughout the experiment (months 6–48), the adult and immature lice that were removed but not photographed were immediately placed back on birds, thus ensuring stability of the lineages over time.

In light of the revised information, certain statements in the 'Materials and Methods' section of the published paper seem ambiguous or oddly phrased, for example,

> *We transferred 800 lice from wild caught feral pigeons to 16 giant runt pigeons and 16 feral pigeon controls (25 lice per bird). At this time (Time 0), we also randomly sampled 800 lice from the source population on wild caught feral pigeons and measured their body size.*

This remark suggests, correctly as it turns out, that the measured lice and the founder lice might be disjoint subsets. But that thought was dispelled by an earlier remark *Once photographed, the live lice were returned to their respective host*, which now turns out to be incorrect.

To learn about a natural host-parasite system, the scientist must manipulate the system to some extent. But as the degree of interference increases, the more is learned about the interference and the less about the natural system. The strong approving remark in the second paragraph of section 5.1.2 about the necessity of returning all lice to their host seems entirely correct as a matter of principle, if only to reduce interference and to minimize the possibility of lineage extinction. Regrettably, it seems now that photographed lice were not returned, perhaps because photography is damaging or destructive. Whether that degree of interference is acceptable or excessive is a matter of biological judgement best left to subject-matter experts, not a matter on which statistical expertise carries weight. As always, the over-riding concern is that the experiment be reported as it was conducted.

5.5.2 Modifications to analyses

At this point we accept the new design information, and ask what effect it has on the appropriateness of the analyses already performed, and what modifications are required.

Consider first the information that the association of time-zero measurements with lineages is virtual. This fact implies that the information content is unchanged if time-zero values are permuted in any manner that preserves sexes, while non-baseline values stay put. A baseline permutation that preserves sexes is one in which males are permuted with other males, females with other females, and non-baseline individuals are fixed. This set of permutations is a sub-group of size $464! \times 336!$ in the larger group of size $3105!$.

Any credible analysis that accommodates the virtual randomization must be invariant with respect to this group of permutations; similar remarks apply to numerical conclusions regarding temporal trends, variance components or other effects. The authors' block-factor assumption (5.3) applies to baseline and non-baseline values, so it contradicts baseline exchangeability, virtual or otherwise. The numerical values reported in their supplementary tables S1–S5 are also not invariant.

Non-virtual baseline exchangeability as discussed in section 5.2.5 implies that the marginal distribution of the initial 800 measurements is invariant with re-

spect to sex-preserving permutation. Virtual exchangeability is a much stronger condition because it implies also that the joint distribution of all 3105 measurements is invariant. Neither condition implies independence of initial and subsequent values, but virtual exchangeability implies that the dependence must be of a trivial type, which is ignorable in practice. The Brownian model (5.4) implies $\mathrm{cov}(Y_u, Y_{u'}) = 0$ for any pair $u \neq u'$ such that $t(u) = 0$ or $t(u') = 0$. Together with (5.2), it also satisfies the virtual exchangeability condition. By contrast, the standard random-effects model (5.3) having independent and identically distributed lineage effects that are constant in time, does not satisfy even the weaker exchangeability condition. It is also incompatible with the discussion in section 5.4.2.

The Brownian-motion model is in line with the standard genetic theory for trait evolution, and is compatible with virtual randomization as described above. Thus the conclusions as stated at the end of section 5.3 are confirmed. Average size differences between the two hosts shown in Table 5.1 are less than 2% and are compatible with neutral evolution in both hosts. The sex-adjusted PC1 mean differences GR − F at each non-zero time point are very similar to the unadjusted differences displayed in Fig. 5.2, but the correctly-computed standard errors tell a very different story.

Table 5.4. PC1 mean differences Giant Runt − Feral by time

	Time in months								
	0	6	12	18	24	30	36	42	48
Diff	0.000	−0.114	0.111	0.464	0.496	0.316	0.251	0.494	0.750
s.e.	0.00	0.24	0.33	0.40	0.46	0.51	0.56	0.60	0.65
Ratio	0.00	−0.48	0.34	1.16	1.09	0.62	0.45	0.82	1.16

Both the differences and the standard errors in this table are computed from a fitted Gaussian model, in which the temporal trend, previously modelled as a zero-mean random effect with covariance $\sigma_3^2(t \wedge t')$ in (5.4), is replaced with a non-random term in the mean. The moments are

$$E(Y_u) = \beta_0 + \beta_1 s(u) + \gamma_h(t), \tag{5.5}$$

$$\mathrm{cov}(Y_u, Y_{u'}) = \sigma_0^2 \delta_{u,u'} + \sigma_1^2 \delta_{l,l'}(t \wedge t') + \sigma_2^2 \delta_{a,a'}(t \wedge t') + \sigma_3^2 \delta_{uu'} I_{t=0}. \tag{5.6}$$

The mean subspace includes an additive constant for sex, and a host-dependent temporal trend $\gamma_h(t)$. The factorial model formula

```
sex + as.factor(time):host
```

generates a subspace of dimension $1 + 9 \times 2 = 19$, but the randomization constraint implies $\gamma_0(0) = \gamma_1(0)$, which reduces the dimension by one. The fitted differences $\hat{\gamma}_1(t) - \hat{\gamma}_0(t)$ are shown in the table, together with standard errors as estimated by REML and weighted least squares. They are automatically sex-adjusted, so they are not exactly the same as the sample differences shown in Fig. 5.2.

If the randomization constraint is ignored, the fitted difference is non-zero for $t = 0$. All estimates and standard errors throughout the table are altered, but only slightly.

At no time does the observed difference reach much above one standard error, so claims for rapid divergence are not supported by this analysis or by any modifications that include non-trivial temporal dependence: see Exercise 5.24. The same applies to the overall estimate of linear temporal trend, which is 0.0142 per month with standard error 0.013. Comparable analyses for body length and other size measurements point to similar conclusions.

It is possible to satisfy the randomization constraint by restricting the block factor terms in (5.3) to post-baseline times only. But Brownian motion implies a non-trivial temporal correlation, and is a much better fit than the restricted block factor. The implication is that the evidence for non-trivial temporal correlation is very strong (see Exercises 5.24 and 5.25). Every modified analysis that takes account of such correlations leads to very similar conclusions. This analysis does not imply that divergent evolution does not exist on some time scale. But it is safe to say that no evidence for it exists in these data.

5.5.3 Further remarks

According to the reply by Scott Villa, the sex ratio of lice at baseline was intentionally biased towards females, with 13–14 females and 11–12 males as founders for each lineage. Following the initial seeding, male and female lice were sampled in approximately equal numbers, so information on the evolution of the sex ratio over time is not available. In light of this information, much of the speculation in section 5.4.1 is not relevant.

Villa also takes issue with a remark in section 5.2.1 that the overall change in body size is surprisingly small, which suggests that changes of this magnitude ($< 2\%$) cannot be biologically significant. His counter-claim is that *body size changes on this scale are biologically relevant for this species, as the effect on mating behavior shows (Villa et al. 2019, Figs 2–5).*

The coefficient of variation of body length for female lice within aviaries is very stable at 2.4%–2.6% from six months onwards; the value for males is equally stable at 2.2%–2.4%. These numbers represent natural variability of body length within freely breeding populations, which is approximately 2.4%. The mean differences between hosts are shown in Table 5.1; they are almost uniformly less than 2%.

What are the implications for mating? The root mean square size discrepancy between a random pair from the same aviary is approximately $\sqrt{2 \times 2.4^2}$, or 3.4%, so the distribution of $F-M$-size differences is approximately $N(411, 83^2)$. A 2% increase in mean size for females implies that the distribution of size differences for mixed hosts is $N(411 + 50, 83^2)$. If size discrepancy is the chief determinant of sexual compatibility, and incompatibility is rare in each population, a mean difference of 0.6 standard deviations is not sufficient to make the incompatible fraction large in the mixed population.

The two movies provided by Villa et al. (2019) illustrate size discrepancies of 1.8 and -2.6 standard deviations, so their relevance at the 0.6σ-scale is not immediately apparent. In the absence of a detailed morphological explanation, it is difficult to accept the authors' claim that body size changes on this scale ($\sim 0.6\sigma$) are biologically important for any species.

5.6 Exercises

5.1 According to the standard definition in section **??**, two observational units u, u' belong to the same experimental unit if the treatment assignment probabilities given the baseline configuration satisfy $P(T_u = T_{u'}) = 1$. Section 5.1.3 makes the argument that each louse is one observational unit, and that each lineage is one experimental unit. But the author subsequently pivots to *aviary* as the experimental unit, hedging his bets by stating that 'both seem to be relevant'. Discuss the arguments pro and con of *louse-lineage* versus *louse-aviary* versus *lineage-aviary* as the observational-experimental units. In connection with the models in section 5.2, what are the substantive implications of one choice versus another?

5.2 According to Villa *et al.*,

> *Pigeons combat feather lice by removing them with their beaks during regular bouts of preening. Columbicola columbae, a parasite of feral pigeons, avoids preening by hiding in spaces between adjacent feather barbs; preening selects for C. columbae small enough to fit between the barbs. Preening also exerts selection on traits critical for locomotion on the host.*

In light of this information, comment on the remark in section 5.1.3 *... size-biased sampling need not be a serious concern for this experiment provided that it affects all birds equally.*

5.3 Download the data, compute the averages at each time point for the two pigeon breeds, and reconstruct the plots in Fig. 5.1 and Fig. 5.2.

5.4 The coefficient of variation is the standard-deviation-to-mean ratio, which is often reported as a percentage. For *body length* or other size variables, the coefficient of variation is essentially the same as the standard deviation of the log-transformed variable. Compute the coefficient of variation of *body length* separately for male and female lice on each occasion, and report this as a table of percentages. What patterns do you see in this table for males versus females or baseline versus non-baseline?

5.5 Use `anova(...)` to re-compute the mean squares in Table 5.2. Use Bartlett's statistic to test the hypothesis that the residual mean squares have the same expected value at all time points. What assumptions are needed to justify the null distribution?

5.6 For the model (5.3), what is the expected value of the within-lineage mean square at time t? For the Brownian-motion model (5.4), show that the variance of Y_u increases linearly with time. What is the expected value of the within-lineage mean square?

5.7 Use `lmer(...)` to fit the variance-components model (5.3) to the log body length with (5.2) as the mean-value subspace. Report the two slopes, the slope difference, and the three standard errors.

5.8 Explain why (5.3) is in conflict with randomization.

5.9 Compute the four covariance matrices $V_0, \ldots, V_3$ that occur in (5.4). Let Q be the ordinary least-squares projection with kernel (5.2). Compute the four quadratic forms $Y'Q'V_rQY$ and their expected values as a linear function of the four variance components. Hence or otherwise, obtain initial estimates.

5.10 Use `regress(...)` to compute the REML estimate of the variance components in (5.4). Hence obtain the estimated slopes, their difference, and the standard errors for all three.

5.11 For $n = 100$ points $t_1, \ldots, t_n$ equally spaced in the interval $(0, 48)$, compute the matrix

$$\Sigma_{ij} = \delta_{ij} + \theta(t_i \wedge t_j)$$

for small values of θ, say $0 \le \theta \le 0.02$. Find the maximum-likelihood estimate of β in the linear model $Y \sim N_n(\alpha + \beta t,\ \Sigma)$ with Σ known, and plot the variance of $\hat{\beta}$ as a function of θ. Comment on the effect of the Brownian-motion component.

5.12 Regress the 32×9 lineage-time averages (for PC1) against sample size using sample size as weights. You should find a statistically significant positive coefficient a little larger than 0.01. Explain why the conclusions from this exercise are so different from those at the end of section 5.4.2.

5.13 In Table S2 of their Appendix, Villa *et al.* fit the eight-dimensional factorial model *host:sex:time* to the first principal component values on 3096 lice. Show that this is equivalent to fitting four separate linear regressions $E(Y_u) = \alpha + \beta t_u$, with one intercept and one slope for each of the disjoint subgroups, Fer.F, Fer.M, Gr.F, Gr.M. Feral and female are the reference levels, so $sex_u = 1$ is the indicator vector for males. Deduce that the *host:time* coefficient is equal to the slope difference $\beta_{Gr.F} - \beta_{Fer.F}$ restricted to female lice. The fitted value is 0.009. What is the fitted slope difference for male lice?

5.14 The *sex* coefficient in Table S2 is -2.437. Which combination of the four α-values in the previous exercise does this correspond to?

5.15 The *host* coefficient in Table S2 is 0.449 with standard error 0.159. What does this imply about the average or expected baseline values for the four subgroups?

5.16 For the model with persistent aviary patterns described at the end of section 5.4.2, compute and plot the REML profile log likelihood for τ in the range $0.5 \le \tau \le 24$. Use PC1 as the response, and (5.2) for the mean-value subspace. The covariance should be a linear combination of five matrices, one each for the identity matrix and the identity restricted to baseline, two Brownian-motion product matrices as in (5.4), and one τ-shifted B-M product matrix. Ten to twelve points equally spaced on the log scale should suffice for plotting.

5.17 Use the profile log likelihood plot in the previous exercise to obtain a nominal 95% confidence interval for τ.

5.18 Distributional invariance. Consider a simplified version of the louse model in which there are 16 feral and 16 giant runt pigeons, no sex differences between lice, and no correlations among measurements. Two lice are associated with each bird at baseline, and two at each subsequent time $t = 1, \ldots, 7$ for a total of 512 observations. Each louse u is associated with a host type $h(u)$, feral or giant runt, and the joint distribution is Gaussian with moments

$$E(Y_u) = \beta_0 + \beta_{h(u)} t_u; \qquad \mathrm{cov}(Y_u, Y_{u'}) = \sigma^2 \delta_{u,u'}.$$

A baseline permutation is a 1–1 mapping $u \mapsto \tau(u)$ such that $t(u) > 0$ implies $\tau(u) = u$. Distributional invariance means that the permuted vector Y^τ with components $Y^\tau_u = Y_{\tau(u)}$ has the same distribution as Y. Show that the joint distribution is invariant with respect to baseline permutations. Note that $h(\tau(u))$ is not necessarily equal to $h(u)$.

5.19 Procedural invariance. Consider a sample of 512 observations generated according to the model in the previous exercise. The estimation procedure is invariant if $\hat\beta(Y) = \hat\beta(Y^\tau)$ and $\hat\sigma(Y) = \hat\sigma(Y^\tau)$ for every baseline permutation. Is it necessarily the case that distributional invariance implies procedural invariance? Explain why least-squares and maximum-likelihood are invariant procedures.

5.20 Consider the following statement taken from section 5.5. *Any credible analysis that accommodates the virtual randomization must be invariant with respect to the same group, and similar remarks apply to numerical conclusions regarding temporal trends, variance components or other effects.* Invariance in this setting means that each distribution in the model is exchangeable, or invariant with respect to sex-preserving baseline permutations. This is a demanding standard, and it is possible that subsequent statements in that same section may not live up to it. Show that the model-formula `Host:as.factor(Time)`, which is related to Table 5.4, corresponds to a set of vectors, some of which are not group-invariant. Investigate the implications, particularly for time zero.

5.21 According to the text in section 5.5, *Virtual randomization requires the time-zero average for feral hosts to be the same as that for giant runts, but the temporal trends are otherwise unconstrained.* It appears that the model matrix spanning this subspace is not constructible using factorial model formulae. Explain how to construct the desired matrix including a constant additive sex

effect. What is its rank? Fit the model as described in the text following Table 5.4. Include independent Brownian motions for aviaries and lineages, plus an additional baseline error term with independent and identically distributed components.

5.22 Use the fitted model from the previous exercise to compute the linear trend coefficient

$$\frac{\sum t(\hat{\gamma}_1(t) - \hat{\gamma}_0(t))}{\sum t^2}$$

and its standard error. You should find both numbers in the range 0.013–0.015 per month, similar to, but not exactly the same as those reported in the text.

5.23 The model in the previous two exercises has a baseline variance that is larger than the non-baseline residual variance. What is the ratio of fitted variances?

5.24 The fact that measured lice were not returned to their hosts is an interference in the system that may reduce or eliminate temporal correlations that would otherwise be expected. One mathematically viable covariance model that is in line with virtual randomization, replaces each occurrence of $t \wedge t'$ in (5.6) with the rank-one Boolean product matrix $(t > 0)(t' > 0)$, so that the only non-zero temporal correlations are those associated with lineage and aviary as strictly post-baseline block factors. Fit this modified block-factor model to the PC1 response with (5.5) for the mean subspace. Which model fits better? Is the log likelihood difference small or large? An informal comparison suffices at this point.

5.25 Construct two versions of Table 5.4, one based on the modified block-factor model, and one based on the combined variance model that includes both. Comment on any major discrepancy or difference in conclusions based on the various models.

5.26 What was the matter that Lord Denning refused to accept in his 1980 appeals-court judgement when he referred so melodramatically to the 'appalling vista that every sensible person would reject'? And why was this phenomenon so abhorrent to him?

Example 6

6.1 A meteorological temperature series

The UK Meteorological Office, maintains the longest continuous instrumental temperature record in the world. According to the MET office website,

> These daily and monthly temperatures are representative of a roughly triangular area of the United Kingdom enclosed by Lancashire, London and Bristol. The monthly series, which begins in 1659, is the longest available instrumental record of temperature in the world. The daily mean-temperature series begins in 1772.

Here we examine the Central England daily temperature series, from January 1, 1772 to Dec 31, 2019. The series length is 90 580 days over 248 years.

The data in tenths of a degree Celsius can be downloaded from the address

`https://www.metoffice.gov.uk/hadobs/hadcet/cetdl1772on.dat`

The values for each year are arranged in a 31×12 array, one column for each month and one row for each day in standard Gregorian format. Non-existent days are padded with the placeholder 'value' -999. For present purposes, we assume that the data have been rearranged in standard data-frame format with one row for each of $n = 90\,580$ days. Each column is one variable. Apart from `temp` and `day`, it may be convenient to include the first and second-order annual harmonics

$$c(t) = \cos(2\pi t/\tau),\ s(t) = \sin(2\pi t/\tau);\ \ c(2t) = \cos(4\pi t/\tau),\ s(2t) = \sin(4\pi t/\tau),$$

where t is time measured in days counted from Jan 1, 1772, and $\tau = 365.2425$ is the mean number of days in one Gregorian year.

As is often the case with very extensive data, much can be learned from simple graphs and other summaries without resorting to formal stochastic models. We first examine the nature of the annual seasonal cycle.

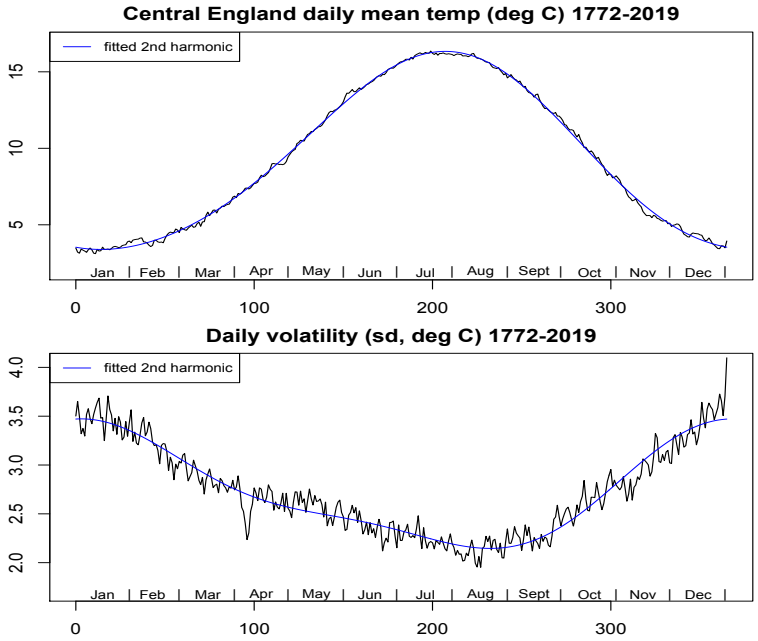

Figure 6.1: Mean temperature and volatility by day of the year.

6.2 Seasonal cycles

6.2.1 Means and variances

The average temperature for each date in the year is computed by associating with each day a calendar date, either the Gregorian calendar date or some version thereof. In the Gregorian system, each date is an integer in the range 0–365, beginning with Jan. 1 coded as zero. February 28 and March 1 are coded as 58 and 60 respectively, whether these are consecutive days or not. For present purposes, it suffices to code `day` as sequential integers $0{:}(n-1)$, where $n = 90\,580$, and to use the mathematical calendar date `date <- trunc(day%%tau)`, which is an integer in the range 0–365. Whatever version of the calendar is used, the average for each date is computed as follows:

```
dailymeantemp <- tapply(temp, date, "mean")
```

Our mathematical dates do not correspond exactly with the Gregorian calendar date, mostly because the leap day is intercalated at the end of December rather than at the end of February. Thus, each calendar date 0–364 occurs 248 times, and these dates are always consecutive days, whereas the leap date occurs only 60 times, so date 0 follows 365 in leap years and 364 in non-leap years. Similar remarks apply to date number 59 (Feb. 29) in the Gregorian system. Wherever the leap date is intercalated, a minor discontinuity may be introduced, as can

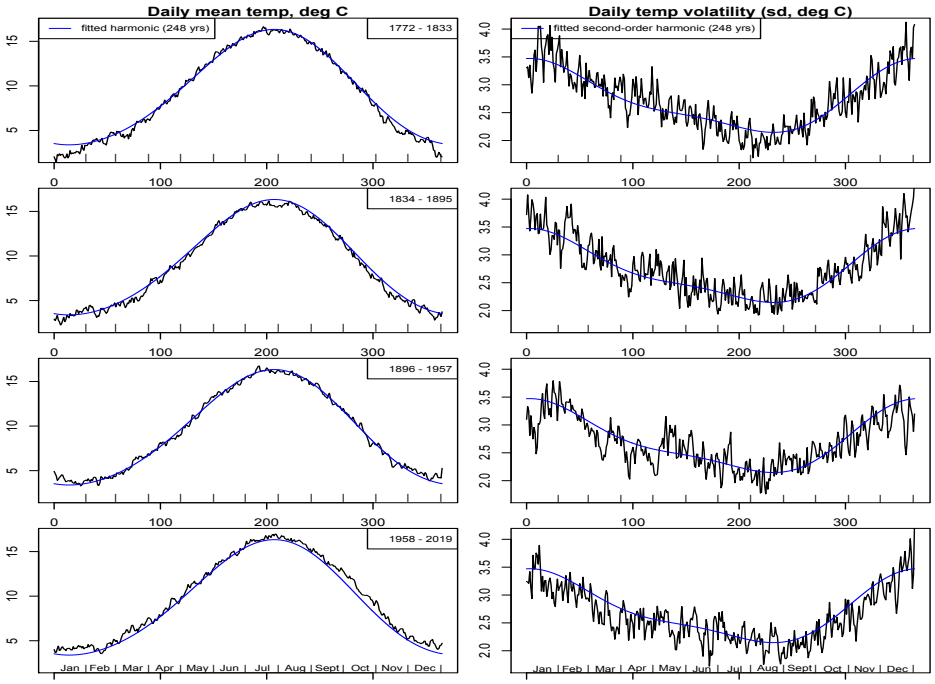

Figure 6.2: Mean temperature and volatility by day of the year in consecutive 62-year blocks

be seen in the volatility series in Figure 6.1. If the Gregorian date is used, the discontinuity at Dec. 31/Jan 1 disappears, but does not reappear at Feb. 29.

Neither the mean series nor the volatility series is adequately described by a first-order harmonic function, which is a linear combination of the three basis vectors $\mathbf{1}, c(t), s(t)$, but both are reasonably well described by second-order harmonic functions. The fitted harmonics shown in Fig. 6.1 were computed by ordinary least squares,

```
fit <- lm(dailymeantemp~c1+s1+c2+s2)
```

which is perfectly adequate for graphical purposes, but technically sub-optimal because of serial correlation. Note that all vectors at this stage, including the harmonic functions, are functions of the date, so each vector has 366 components. The leap date could be given reduced weight in the analysis, but this has not been done here. Alternatively, the leap date could be omitted, with a corresponding modification in the harmonic functions.

Apart from the discontinuity at the leap day (Dec 32), which results in a spike in volatility, there is a curious anomaly of reduced temperature volatility around April 5–9. The depression in volatility is spread over several days, and is evident also in plots using Gregorian dates. Whatever its cause—social, ecclesiastical or

meteorological—the volatility plots in Fig. 6.2 show that the phenomenon has persisted for over 200 years.

Figure 6.2 is the same as Fig. 6.1 except that the period has been split into four non-overlapping blocks of 62 years in order that long-term trends and variations in the annual cycle might be revealed. To simplify cross-block comparisons, the plotting scales are fixed for each block, and the second-order harmonic is also fixed to serve as a historical reference.

It is evident that there has been no major shift in the seasonal cycle over this period. However, winter temperatures, particularly in January, have risen by several degrees throughout this period, and that increase began even in the 19th century. The low summer and autumn temperatures in the late 19th century are well known and are often attributed to volcanic effects such as the Krakatowa eruption in 1883. However, the lowest annual mean in this series occurs in 1879, four years before the eruption, and the expected volcanic effects are not readily apparent in the annual averages for the decade that follows: see Fig. 6.4. Other than the winter increase, the annual pattern in the early 20th century is remarkably close to that in the early 19th century. The phenomenon that stands out in Fig. 6.2 is the uniformly high temperature throughout the year in the most recent period. Only on 35 dates do the daily averages for 1958–2019 fall below the historical reference curve.

6.2.2 Skewness and kurtosis

Fisher's k-statistics of order 2–4 for a sample of n observations are

$$\begin{aligned}
(n-1)\,k_{2,n}(x) &= \sum (x_i - \bar{x}_n)^2, \\
(n-1)^{\downarrow 2}\,k_{3,n}(x) &= n \sum (x_i - \bar{x}_n)^3, \\
(n-1)^{\downarrow 3}\,k_{4,n}(x) &= n(n+1) \sum (x_i - \bar{x}_n)^4 - 3(n-1)^3 k_{2,n}^2(x),
\end{aligned}$$

where $n^{\downarrow r} = n(n-1)\cdots(n-r+1)$ is the descending factorial, and $k_{r,n}$ is defined for $n \geq r$ only. For an iid sample, the expected values are the population cumulants $E(k_{r,n}) = \kappa_r$, which are zero for $r \geq 3$ in Gaussian samples. The third and fourth standardized k-statistics are $k_3/k_2^{3/2}$ and k_4/k_2^2, which are invariant with respect to affine transformation $x_i \mapsto a + bx_i$ with $b > 0$. Thus, the fact that the temperature is recorded in °C rather than °F has no effect on the standardized values. These statistics are frequently used to gauge departures from normality. Here we are looking at cumulant variations as a periodic annual time series.

The standardized values are plotted by calendar date in Fig. 6.3, so each skewness and kurtosis coefficient is computed using 248 replicate temperature values for every non-leap date, or 60 for the leap date. The average skewness is close to zero, but there is a distinct sinusoidal cycle with a summer maximum, which is in phase with the mean temperature cycle. Winter temperatures are skewed negatively, summer values positively. The kurtosis values are more widely scattered with no clear pattern, but summer values are slightly larger on

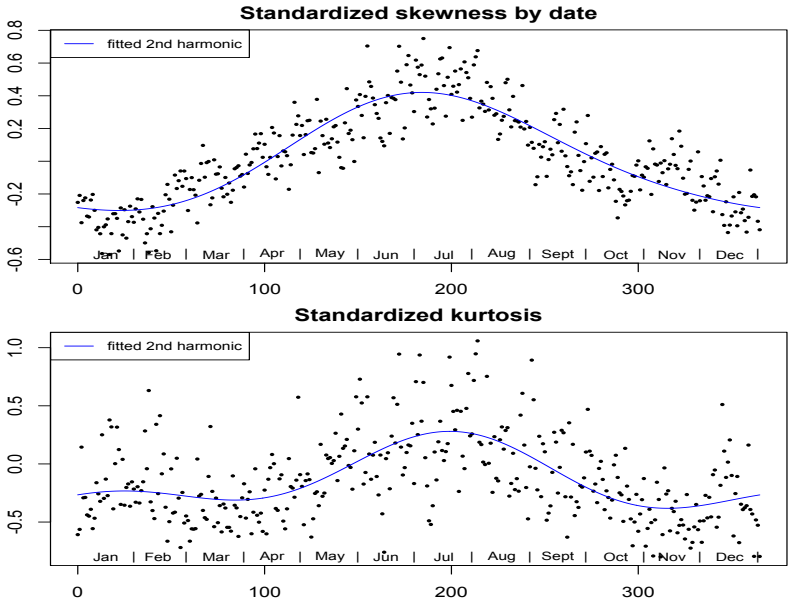

Figure 6.3: Skewness and kurtosis of temperatures by date.

average than those in other months. Two thirds of the k_4-values are negative, indicating that tails are shorter than Gaussian. The sinusoidal trend in the skewness plot is clear evidence of non-normality, but that is not an adequate reason to abandon methods of analysis based on linear decompositions.

It is worthwhile recalling the inheritance property of sample statistics $k_{r,n}$, and more general U-statistics, computed for sub-samples of various sizes. Let $[N]$ be the population and $S \subset [N]$ a sample of size $n \leq N$; let $Y[S]$ be the sample temperatures and $k_{r,n}(Y[S])$ the sample statistic. Given the population statistic $k_{r,N} \equiv k_{r,N}(Y[N])$, the average over samples of size n satisfies

$$\underset{S \subset [N]}{\text{ave}}\, k_{r,n}(Y[S]) = k_{r,N}(Y[N]).$$

Thus, given that the variance for April 7 is low relative to April 1 or April 12 in the population of 248 years, we should expect the same to hold on average for simple random samples or simple random partitions. Although a sequential block of 62 years is not a simple random sample, it may behave as such in the absence of serial correlation, in which case the depression seen for April 5–9 variances in successive 62-year blocks in Fig. 6.2 is expected and not a surprise.

This inheritance argument does not apply to the standardized skewness or standardized kurtosis, which are not U-statistics. Nevertheless, an approximate version of inheritance does hold.

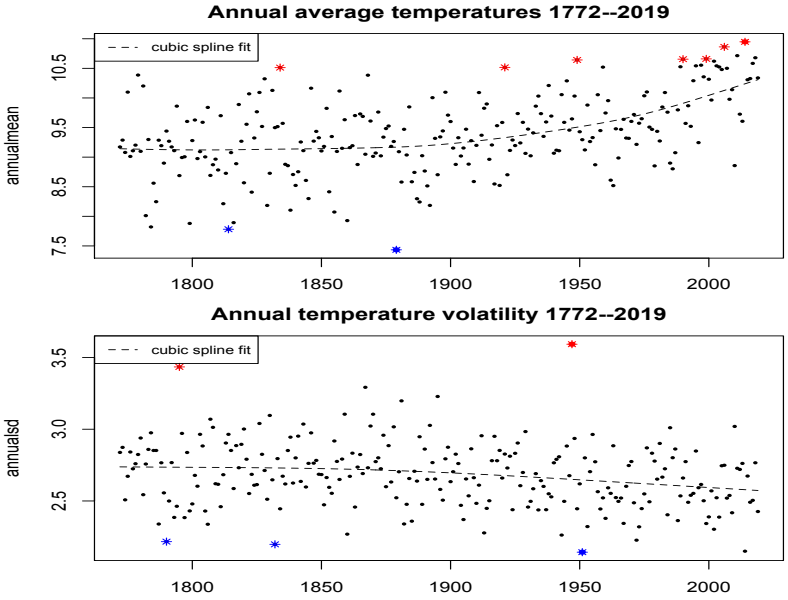

Figure 6.4: Mean temperature and volatility over 248 consecutive years, with record highs and lows indicated.

6.3 Annual statistics

6.3.1 Means and variances

The top panel of Fig. 6.4 shows the annual average temperature for each year over the 248-year period. Post-1790 record lows and record highs are indicated: year t is a record high if $Y_t = \max\{Y_1, \ldots, Y_t\}$, and a record low if $Y_t = \min\{Y_1, \ldots, Y_t\}$. The record lows occur in 1814 and 1879, the record highs in 1834, 1921, 1949, 1990, 1999, 2006 and 2014. By visual inspection, the mean trend is constant up to about 1900, increasing slowly to about 1950, and more rapidly thereafter. The maximum-likelihood cubic-spline fit has been superimposed as a summary of the mean trend. Computational details are given in the following section.

The second panel of Fig. 6.4 is similar to the first, except that it shows the within-year standard deviation measured as the deviation from the second-order harmonic fit. The harmonic term is removed so that the effect of seasonal variation is kept to a minimum. Post 1790 record lows and highs are highlighted; the lows occur in 1790, 1832 and 1951, the highs in 1795 and 1947. The trend in volatility is downwards as indicated by the cubic spline fit, but it is not significantly non-linear over this period.

Changes in meteorological technology over the centuries must have an effect on variability of measurements, but this effect seems unlikely to be large for temperature measurements. Temperatures are well calibrated relative to the

freezing and boiling points of water, so the effects of technological innovation on measurements of annual average temperatures are likely to be small, if not entirely negligible.

6.3.2 Variance of block averages

The focus in this section is on the behaviour of block averages $\bar{Y}_{t:t+h}$ for contiguous blocks of length h. To eliminate seasonal variation, we restrict attention to blocks whose length is an integer number of years. In the following table, the sample averages for 5000 blocks of length h years or $365.25h$ days were obtained, and the sample variance of these block averages was computed. Blocks were sampled uniformly at random, not necessarily starting on Jan 1. Standard theory for finite samples tells us that, in the absence of correlation, the sample variance of the averages should be proportional to $(1-f)/h$, where $f = h/248$ is the sampling fraction and $1-f$ is the finite-population correction factor. Accordingly, the second line reports the corrected variance of the block averages, with $C_h = 1/(1-f)$.

	Block length in years					
h	4	8	16	32	64	128
$C_h \operatorname{var}(\bar{Y}_h)$	0.213	0.158	0.133	0.087	0.058	0.036
$h \times C_h \operatorname{var}(\bar{Y}_h)$	0.853	1.261	2.130	2.778	3.719	4.624
$h^{1/2} \times C_h \operatorname{var}(\bar{Y}_h)$	0.427	0.446	0.533	0.491	0.465	0.409

The standard theory for uncorrelated values also applies asymptotically to block averages from a stationary processes provided that the correlations decay at a sufficiently fast rate. For a short-range dependent process the product $hC_h \operatorname{var}(\bar{Y}_h)$ shown in the middle line should be approximately constant in h, at least for large h. However, this product is clearly increasing as a function of the block size. The third line suggests that $h^{1/2}C_h \operatorname{var}(\bar{Y}_h)$ is approximately constant, and hence that the variance of block averages behaves inversely as the square root of the block size rather than $O(h^{-1})$. This phenomenon is known as long-range dependence. The behaviour observed here for block averages is consistent with the assertion that the covariance function does not have a finite integral. It is incompatible with short-range dependence such as $e^{-|s|}$ or $P(s)e^{-|s|}$ for any polynomial P, or any of the finite-range Matérn models.

6.3.3 Variogram at short and long lags

The variogram of a stationary process at lag h is the expected value of the squared difference $|Y_t - Y_{t+h}|^2$, which is non-negative and symmetric in h. If the process has a covariance function $K(|t-t'|)$, the variogram is

$$\gamma_h = E\big(|Y_t - Y_{t+h}|^2\big) = 2K(0) - 2K(h) = 2\sigma^2\big(1-\rho(h)\big),$$

where $\rho(h)$ is the autocorrelation at lag h, and σ^2 is the variance. The semi-variogram is one half the variogram, and $\rho(h)$ is the autocorrelation function.

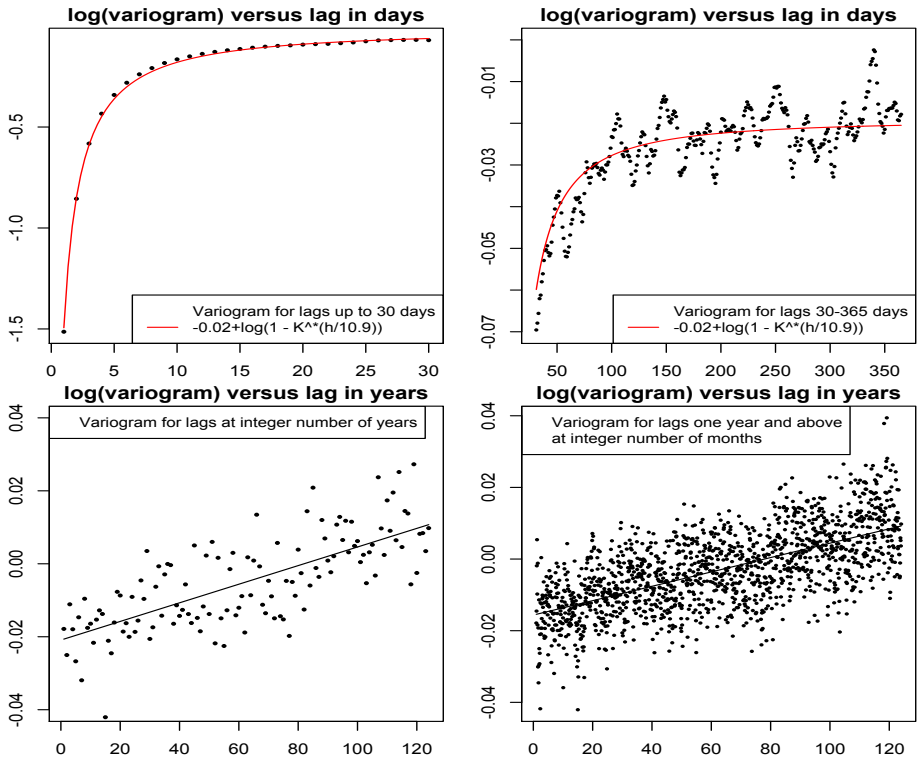

Figure 6.5: Empirical log variogram of temperatures split into short, medium and long lags. A least-squares fitted curve for short and medium lags taken together is shown in the top two panels. For the longer lags, the least-squares straight line with slope 0.20–0.25 per millennium is shown.

The empirical variogram is the average squared difference of sample values

$$\tilde{\gamma}_h = \frac{1}{n-h}\sum_{t=1}^{n-h}(Y_t - Y_{t+h})^2.$$

If the process has a non-constant mean, but is otherwise stationary, the residuals are used instead. The empirical variogram provides a decomposition of the total sum of squares by lags:

$$\frac{1}{n}\sum_{h=1}^{n}(n-h)\tilde{\gamma}_h = \sum(Y_i - \bar{Y})^2.$$

However, it is not an orthogonal decomposition.

Figure 6.5 shows the log variogram split by short, medium and long lags. All plots are based on residuals after eliminating first and second-order seasonal harmonics. The first panel shows the typical monotone increase for lags 1–30

days; the second panel is re-scaled to show a similar, but less steep, increase for lags of 30–365 days. A particular curve fitted by non-linear least-squares is superimposed on the log variogram, the same curve $-0.02+\log(1-K^*(h/10.9)$ in both panels. Details concerning the SD-1/2 covariance function K^* are given in Section ?; the behahiour for large h in excess of about 5 is $4K^*(h) \simeq 1.25h^{-3/2} - h^{-2}$. Overall, the fitted curve tracks the observed values quite well over lags $1 \leq h \leq 365$ measured in days, but it is essentially constant after about 12–18 months. The standardized variogram and the autocorrelations implied by the fitted curve for lags up to one week are as follows:

h	1	2	3	4	5	6	7
$\tilde{\gamma}(h)/(2s^2)$	0.220	0.425	0.559	0.648	0.711	0.755	0.788
$\hat{\gamma}(h)/(2\hat{\sigma}^2)$	0.224	0.441	0.572	0.658	0.718	0.761	0.794
$\hat{\rho}_h$	0.776	0.559	0.428	0.342	0.282	0.239	0.206

For lags $2 \leq h \leq 4$, the SD-1/2 autocorrelations satisfy $\hat{\rho}_h < \hat{\rho}_1^h$, the inequality being reversed for $h > 4$.

The third and fourth panels show the behaviour for very long lags in the range 1–120 years. The third panel is restricted to lags that are an integer multiple of one year, so that the sequence of values is not affected by the elimination of seasonal cycles. This graph indicates that the log variogram increases at the rate 0.25 units per millennium

$$\log \hat{\gamma}(h) = \text{const} + 0.25h/1000$$

over the range $1 \leq h \leq 120$ years. The fourth panel shows lags that are integer multiples of one month over the same range. The least-squares fitted line in this case is a little flatter with slope 0.20 units per millennium. Neither scatterplot suggests a substantial deviation from linearity over the range $1 \leq h \leq 120$ years.

It is striking that the SD-1/2 variogram curve $\gamma(h) = \sigma_0^2 - \sigma_1^2 K^*(h/\lambda)$, which fits the empirical variogram reasonably well for lags up to and well beyond one year, has a finite limit $\gamma(\infty) = \sigma_0^2$, and thus fails completely to capture the non-constant behaviour of the variogram at very long lags. Although the long-range trend is difficult to deny, the implied annual increase is almost imperceptible and is comparable to the width of one plotting symbol in the second panel.

6.4 Stochastic models for the seasonal cycle

6.4.1 Structure of observational units

The observational units in a time series are the time points at which measurements are made. Usually, there are no replicate measurements at the same time. In this instance, each day is one observational unit, the observational units are completely ordered and are associated with the natural numbers, i.e., equally spaced points on the real line. In the absence of further structure, we have at our disposal only one fundamental covariate, which is time measured in days

beginning at an arbitrary point, which is taken to be zero for Jan 1, 1772. There is, however, one crucial piece of additional information, which is the length of the Gregorian year, $\tau = 365.2425$ days. From this we arrive at the first-order harmonics,

$$c(t) = \cos(2\pi t/\tau),\ s(t) = \sin(2\pi t/\tau)$$

whose period is one calendar year. The kth-order harmonics $c(kt), s(kt)$ have a period of $1/k$ years. These are the only plausible functions that are available for use as covariates in the model for the mean temperature. One crucial property of harmonics is that the subspace spanned by each pair $c(t), s(t)$ is closed with respect to temporal translation:

$$\operatorname{span}\{\cos(t), \sin(t)\} = \operatorname{span}\{\cos(t+h), \sin(t+h)\}$$

for each displacement h, and the same holds for the pair $c(kt), s(kt)$. The space $\mathcal{H}_k$ of harmonics of degree $\leq k$ is a vector space of dimension $2k+1$, in which $\mathcal{H}_0 = \mathbf{1}$ is the space of constant functions.

The Fourier basis vectors $c(kt), s(kt)$ are exactly orthogonal in the continuous setting as functions on $(0, 2\pi)$, and they are exactly orthogonal in certain uniformly-spaced discrete-time settings. In the present discrete setting, they are not quite orthogonal because of the leap-year complication, but this effect is very small.

6.4.2 Seasonal structure

The fitted second-order harmonic shown in Fig. 6.1 is a reasonably accurate description of the seasonal pattern in mean temperature. Although the deviations from this curve are small, they are far from independent, as the following analysis demonstrates.

The additional structure on observational units comes not in the form of covariates, but in the form of relationships between pairs of observational units (t, t'). The most obvious relationship is the Euclidean metric $|t - t'|$ on the real line, but there are also at least three periodic semi-metrics that are more natural for the description of seasonal cycles:

$$\begin{aligned}
\chi(t,t') &= \frac{\tau}{\pi} \sin\Big(\frac{\pi|t-t'|}{\tau}\Big), \\
\ell(t,t') &= \min\{t-t',\ t'-t\}, \\
d(t,t') &= (t-t')\big(\tau - (t-t')\big)/\tau.
\end{aligned}$$

The first two are respectively the chordal distance and the arc length on the annual circle whose perimeter is τ. In all three expressions, t, t' are understood as points in the space $\mathbb{R} \pmod{\tau}$, with addition modulo τ, so that $0 \leq t - t' < \tau$ and $t' - t = \tau - (t - t')$ are complementary arc lengths. With this understanding, it can be verified that d is a metric. For each metric, the maximum values τ/π, $\tau/2$ and $\tau/4$ occur at diametrically opposite points $t - t' = \tau/2$.

For most statistical work, d and χ are essentially equivalent: see Exercises 6.10–13.

6.4.3 Continuous periodic processes

For each $\lambda > 0$, the function $K(t,t') = \exp(-\chi(t,t')/\lambda)$ is positive definite on $[0,\tau)$, and also stationary and positive semi-definite on the real line with period τ. The Gaussian random function $\eta \sim \mathrm{GP}(0,K)$ is periodic and continuous, and is reasonably well suited as a statistical description of the temperature deviations from the seasonal harmonic in Fig. 6.1. For fixed λ, the Gaussian model in which $\mu = E(Y)$ belongs to the space of harmonics of degree two, and $\mathrm{cov}(Y) = \sigma_0^2 I_n + \sigma_1^2 K$, is linear in the parameters. The model can be fitted using the R command

```
fit <- regress(dailymeantemp~c1+s1+c2+s2, ~K)
```

In this discrete computational setting, $\tau = 366$, or 365 if the leap day is dropped, and K is a symmetric matrix of the same order. The identity matrix, or nugget effect, is included by default, so there are two variance components and five regression coefficients to be estimated. As it happens, the nugget variance estimate is zero, or even slightly negative if not constrained. The maximized log likelihood plotted against λ has a maximum at $\hat\lambda \simeq 4.3$ days, and the fitted variance coefficients are $\hat\sigma_0^2 = 0$, $\hat\sigma_1^2 = 3.62$. The log likelihood is distinctly non-quadratic in λ, but it is approximately quadratic in λ^{-1} with a finite long-range limit as $\lambda \to \infty$.

The positivity constraint is enforced either through the optional argument `pos=c(1,1)` or, in this instance, by nugget omission `identity=FALSE`. The residual log likelihood for the covariance model $\sigma_0^2 I_n + \sigma_1^2 K$ exceeds that for the iid sub model with $\sigma_1^2 = 0$ by 167 units, leaving no doubt about strength of the residual serial correlation.

If the arc length is substituted for chordal distance, the resulting process is essentially an autoregressive process of order one, but with a periodic constraint. The dependence is local and confined to a few days, so there is little difference between the chordal and arc-length models.

The quadratic metric is closely related to the Brownian-bridge process: see Exercise 6.??.

6.5 Estimation of secular trend

6.5.1 Gaussian estimation and prediction

Suppose that $Y = (Y_0, Y_1)$ is a pair of random vectors that are jointly Gaussian with moments

$$E(Y) = \begin{pmatrix} \mu_0 \\ \mu_1 \end{pmatrix}, \qquad \mathrm{cov}(Y) = \begin{pmatrix} \Sigma_{00}\, \Sigma_{01} \\ \Sigma_{10}\, \Sigma_{11} \end{pmatrix}$$

in partitioned-matrix form. Then each marginal distributions is Gaussian $Y_0 \sim N(\mu_0, \Sigma_{00})$, and $Y_1 \sim N(\mu_1, \Sigma_{11})$. The conditional distribution of Y_1 given Y_0

is also Gaussian. The mean is linear in Y_0 and the variance is constant. If Σ_{00} is invertible, the moments are

$$E(Y_1 \mid Y_0) = \mu_1 + \Sigma_{10}\Sigma_{00}^{-1}(Y_0 - \mu_0),$$
$$\operatorname{cov}(Y_1 \mid Y_0) = \Sigma_{11} - \Sigma_{10}\Sigma_{00}^{-1}\Sigma_{01}.$$

In statistical applications of this formula for estimation and prediction, Y_0 is the observation vector, and Y_1 is an unobserved random variable—the long-range secular trend. Usually, maximum-likelihood estimates are used for all unknown parameters as needed.

6.5.2 Application to trend estimation

The series of annual averages is modelled as the sum $Y(t) = \eta(t) + \varepsilon(t)$ of two independent Gaussian processes in which the components of ε are independent and identically distributed with mean zero. The secular trend is a smooth random function whose covariance function exhibits long-range dependence. Intuitively, a long-range secular trend conjures up an image of a smooth function in time, so the choice for $K \propto \operatorname{cov}(\eta)$ must force a specific degree of smoothness on the function η. Typically, the mean of η is constant or linear $\mu(t) = \beta_0 + \beta_1 t$, but more general expressions are also possible if the circumstances require it. The statistical goal is to estimate the secular trend by computing the conditional expectation $E\big(\eta(\cdot) \mid \text{data}\big)$, which is called the Bayes estimator. When maximum-likelihood estimates of the fitted parameters are inserted, this is known as the empirical Bayes estimator. In other types of application, the conditional expected value is sometimes called the *best linear predictor* or the Kriging estimate.

It is worth emphasizing that continuity requires $\eta(\cdot)$ to be defined at all points in $\mathbb{R}$, even though the process Y is observed or recorded only at a finite set of points. The conditional expected value $E(\eta(s) \mid \text{data})$ is linear in the observations; its behaviour as a function of s is a linear combination of covariances $K(s, t_i) = \operatorname{cov}(\eta(s), Y(t_i))$ for observation times $t_1, \ldots, t_n$. In practice, the degree of smoothness of K on the diagonal is crucial.

6.5.3 Matérn models

For the present illustration, we choose the Matérn covariance function with index ν

$$K(t, t') = x^{\nu}\mathcal{K}_{\nu}(x),$$

where $\mathcal{K}_{\nu}$ is the Bessel function of order $\nu > 0$, and $x = |t - t'|/\lambda$ is the standardized temporal difference. The exponential covariance function corresponds to $\nu = 1/2$, and in this case $\lambda \log 2$ is the range at which the serial correlation in η is reduced by half. The index range $\nu \geq 1/2$ guarantees continuity of $\eta(\cdot)$ as a random function, $\nu \geq 1$ guarantees continuity of first derivatives, and $\nu \geq 3/2$ guarantees continuity of second derivatives. For computational illustration, we set $\nu = 3/2$ and $\lambda = 1000$ (in units of years). The large value of λ means not

only that serial correlation persists well beyond the observation period but also that the behaviour of η is governed by the behaviour of K near the diagonal.

```
nu <- 3/2;  lambda <- 1000;   x <- abs(outer(yr, yr, "-"))/lambda;
K <- x^nu * besselK(x, nu);   diag(K) <- 2^(nu-1)*gamma(nu)
fit <- regress(annualmean~yr, ~K)
blp <- fit$fitted + fit$sigma[2]*K %*% fit$W %*% (annualmean-fit$fitted)
lines(yr, blp)
```

In the formula for the conditional expectation, `fit$fitted` is the fitted mean vector $\hat{\beta}_0 + \hat{\beta}_1 t$ with 248 components, `fit$W` is the fitted inverse covariance matrix for the observations, `fit$sigma` is the vector of fitted variance components, and `fit$sigma[2]*K` is the matrix of covariances $\mathrm{cov}(\eta(t), Y(t'))$ for t, t' among the observation points. If we wish to make predictions beyond the range of observation times, say to 2020 or 2021, it is necessary to extend the vector of fitted means and the matrix of covariances in the obvious way.

6.5.4 Statistical tests and likelihood ratios

The fitted variance components are $\hat{\sigma}^2 = (0.316, 151.6)$, and the log likelihood ratio statistic for testing the linear sub-model $\sigma_\eta^2 = 0$ against this alternative is $2(16.14 - 8.46) = 15.36$ on one degree of freedom. Note that the null hypothesis sub-model is not stationary, but the mean trend is constrained to be linear in time. Using the standard asymptotic approximation for the distribution of the likelihood-ratio statistic, the tail probability is less than 10^{-4}, so the evidence for non-linearity and/or long-range correlation is fairly strong.

If we wish to test the hypothesis of no trend versus a continuous non-linear trend, we could proceed computationally as follows:

```
nu <- 1/2;  lambda <- 1000;   x <- abs(outer(yr, yr, "-"))/lambda;
K <- x^nu * besselK(x, nu);   diag(K) <- 2^(nu-1)*gamma(nu)
fit0 <- regress(annualmean~1)
fit1 <- regress(annualmean~1, ~K)
2*(fit1$llik - fit0$llik)   #   LLR=71.68
```

To be clear, the null hypothesis of no trend is interpreted here as iid Gaussian observations, and it is this hypothesis that is decisively rejected by the likelihood ratio statistic of 71.68. However, this interpretation of the null hypothesis is arguably unfair because short-term correlation between consecutive annual averages seems inevitable, and no trend is not the same as no correlation.

The difficulty here is that it is unclear statistically what is implied by the null-hypothesis phrase 'no long-term trend'. After all, the Gaussian process with constant mean and covariance $\sigma_0^2 I_n + \sigma_1^2 K$ is temporally stationary. The last snippet of code generates a test statistic that is sensitive to long-range correlation, which is arguably indistinguishable from long-term trend.

6.5.5 Rough paths versus smooth paths

To appreciate the effect of the Bessel index, it is worthwhile computing and plotting the Bayes estimate for the fitted mode shown above with $\nu = 1/2$. For $\nu = 3/2$ the conditional mean is piecewise cubic—a cubic spline which has at least two continuous derivatives at all points. For $\nu = 1/2$ the conditional mean is piecewise linear—a linear spline with a knot at each observation. The linear spline is constrained only by continuity; the cubic spline is constrained by continuity of two derivatives, so it is less flexible and much smoother in appearance. Intermediate values such as $\nu = 1$ have Bayes estimates that are intermediate in appearance.

In the majority of applications of this sort, the likelihood function is close to constant in ν, but also slowly decreasing. In other words, the data are relatively uninformative about smoothness of η, but there is a slight preference for rougher trajectories. For visual extrapolation, however, a smooth curve is a more compelling image and tells a more convincing story than a rough curve. Continuity of second derivatives seems about right, and $\nu = 3/2$ is a reasonable compromise for a graphical summary.

In principle, the range parameter λ can also be estimated by maximum likelihood, but for most practical work the value is effectively infinite, in which case the limit process may used directly. In both examples, we have used $\lambda = 1000$ for illustration, but the maximum is achieved in the long-range limit. Details of the limit process, also called the cubic spline model, are given in chapter ?. The first snippet of code shown above is satisfactory for $\nu \leq 3/2$, but it is not recommended for $\nu > 3/2$ if λ is large. The second snippet with constant mean is satisfactory for $\nu \leq 1/2$, but is not recommended for $\nu > 1/2$ if λ is large.

6.5.6 Smooth and ultra-smooth paths

The Matérn covariance function is convenient in many ways, but it is not essential to the argument. An alternative strategy for accommodating long-range trends is to use an inverse-polynomial covariance function of the form $1/(1+x^2)$, where $x = |t-t'|/\lambda$. Each realization of this process is an infinitely differentiable function, so the conditional expected value $E(\eta \mid \text{data})$ is also a C^∞-function. Unlike the Matérn process, the long-range limit of the inverse-quadratic process is not well-behaved. Consequently, it is necessary to fix a finite range or to estimate the range, and $\hat{\lambda} \simeq 99$ years is the value suggested by the sequence of annual averages. The conditional expected-value curve is not appreciably different from the cubic spline shown in Fig. 6.4.

The 'Gaussian' covariance function $\exp(-|t-t'|^2/\lambda)$ also gives rise to C^∞ trajectories. Usually this choice is not recommended for applied work because the ultra-smooth trajectories give rise to ultra-smooth, non-local, predictions whose apparent accuracy may be misleading. In addition, the long-range limit is not well-behaved as a process, so a finite range is needed for computation.

6.6 Exercises

6.1 Given the variance components, the Bayes estimate of the secular trend is a linear combination of the fitted mean vector and the fitted residual

$$\tilde{\mu} = PY + L\Sigma^{-1}QY,$$

where PY and QY are independent Gaussian vectors. Use this representation to approximate $\mathrm{cov}(\tilde{\mu})$.

6.2 The U.K. Met Office maintains a longer record of monthly average and annual average temperatures for Central England from 1659 onwards in the file

`https://www.metoffice.gov.uk/hadobs/hadcet/cetml1659on.dat`

Check the format, download the data, and plot the annual average temperature as a time series. For the annual mean data up to Dec 31 of the past year, fit the Matérn model with $\nu = 3/2$ and range $\lambda = 1000$ as described in section ?, and plot the Bayes estimate of the secular trend. Repeat the calculation for $\nu = 1$ and range $\lambda = 1000$, and superimpose the two Bayes estimates. Comment briefly on the shape of the fitted curves prior to 1772.

6.3 For the cubic and quadratic models described in the preceding exercise, compute the predicted temperature for next year, i.e., the conditional distribution of the mean temperature for next year given the series of annual averages up to December 31 of the past year. The two models should give slightly different predictive distributions.

6.4 A variety of other smoothing techniques can be employed to illustrate long-term secular trends. Pick your favourite kernel density smoother, apply it to the temperature series, and compare the fitted curve with the Bayes estimates described above.

6.5 Compute the annual average temperatures for the years 1772–2019, and duplicate the first plot in Fig. 2.4. Include the Bayes estimate of the long-term secular trend up to 2025 using the Matérn covariance function with $\nu = 3/2$ and range $\lambda = 1000$ years. Compute the pointwise standard deviation of the Bayes estimate, and include the 95% prediction interval on your plot.

6.6 The U.K. Met Office site `https://www.metoffice.gov.uk/` keeps long-term weather records—temperature, rainfall, and so on—for a range of stations in Great Britain and Northern Ireland. Monthly rainfall totals for Oxford from 1853 are available in the file

`/pub/data/weather/uk/climate/stationdata/oxforddata.txt`

Check the format, download the data, and plot the monthly average rainfall as a seasonal series. Take note of the units of measurement, and include this information on the graph.

6.7 This exercise is concerned with two versions of the Bayes estimate of the seasonal rainfall component, where it is required to compute $E(\eta_m \mid \text{data})$ for each of 12 months. As usual, χ is the chordal distance as measured on the clock whose perimeter is 12 units, and `month` is a factor having 12 levels. In computer notation, the code for fitting the two models is as follows, where $K = \text{const} - \chi$ is positive-definite of order $n \times n$ and rank 12:

```
fit0 <- regress(rain~1, ~month)
fit1 <- regress(rain~1, ~K)
```

Positive definiteness is not required for computation so the constant is immaterial, but K is positive definite if the constant exceeds $2\tau/\pi^2 = 24/\pi^2$. Compute the Bayes estimate of monthly means for each model and superimpose these points on the plot of monthly averages.

6.8 For the Oxford rainfall data up to Dec 2019, the first Bayes estimate in the preceding exercise is a flat 10% shrinkage of monthly averages towards the annual average; the second Bayes estimate is different. For example, the average rainfall for September is 55.6mm, which is slightly above the overall average of 54.7, so the first Bayes estimate is 55.5mm. The second Bayes estimate is 57.5mm. Explain this phenomenon—why the September component, which is already above the annual average, is shifted even further from the overall average.

6.9 Let $(\varepsilon_k, \varepsilon'_k)_{k\geq 0}$ be independent and identically distributed standard Gaussian variables. For real coefficients σ_k, show that the random function

$$\eta(t) = \sum_{k=0}^{\infty} \sigma_k \varepsilon_k \cos(kt) + \sigma_k \varepsilon'_k \sin(kt)$$

is stationary with covariance

$$\operatorname{cov}\bigl(\eta(t), \eta(t')\bigr) = \sum_{k=0}^{\infty} \sigma_k^2 \cos\bigl(k(t-t')\bigr)$$

provided that the series converges in a suitable sense.

6.10 Verify the following trigonometric integral for integer k:

$$\int_0^{2\pi} \sin(x/2)\cos(kx)\,dx = \frac{-4}{4k^2-1}.$$

Hence find the coefficients λ_k in the Fourier expansion of the function

$$2/\pi - \sin(x/2) = \sum_{k=0}^{\infty} \lambda_k \cos(kx)$$

for $0 \leq x < 2\pi$, and show that they are all positive.

6.11 In this exercise, χ is the chordal metric on the unit circle. From the results of the preceding exercise, show that $4/\pi - \chi(t, t')$ is positive definite on $[0, 2\pi)$. Use the Choleski decomposition to simulate and plot a random function having this covariance function as follows:

```
n <- 1000; t <- (1:n)*2*pi/n; chi <- 2*abs(sin(outer(t, t, "-")/2))
eta <- t(chol(4/pi - chi)) %*% rnorm(n)
plot(t, eta, type="l")
```

6.12 Show that the quadratic function

$$\frac{2\pi^2}{3} - x\big(2\pi - x\big)$$

on $[0, 2\pi)$ has Fourier cosine coefficients $4\pi/k^2$ for $k \geq 1$. Hence or otherwise, investigate the function

$$K(t, t') = \frac{2\pi^2}{3} - |t - t'|\big(2\pi - |t - t'|\big)$$

as a candidate covariance function for a process on $[0, 2\pi)$, and by extension to a stationary periodic process on the real line. Plot a simulation of the process on $[0, 4\pi)$, and verify continuity.

6.13 Suppose that $\eta \sim \mathrm{GP}(0, K)$, with K as defined in the preceding exercise. The tied-down process $\zeta(t) = \eta(t) - \eta(0)$ is periodic and zero at integer multiples of 2π. Find its covariance function, and investigate its connection with the classical Brownian bridge.

6.14 Simulate and plot a random function $\eta(\cdot)$ on $(0, 2\pi)$ whose covariance is $\pi/2 - \ell(t, t')$, where $\ell(\cdot)$ is the arc-length metric. This function is less well behaved than the chordal function because half of its Fourier coefficients are zero, so the simulation code must be modified to accommodate singularities.

Example 7

7.1 Frequency-domain analyses

7.1.1 Fourier transformation

A time series is, in the first instance, a function $t \mapsto Y(t)$ on time points, either $t \in \mathbb{R}$ for a continuous-time process, or $t \in \mathbb{Z}$ for a discrete-time process. Most meteorological processes exist in continuous time, but are recorded in discrete time, either as noontime values, daily totals, daily averages or daily maxima. Similar remarks apply to plant and animal growth curves, personal health as a time series, and most business series and economic series. For statistical purposes, either in modelling or analysis, it is helpful to proceed as if the process exists in continuous time but is observed discretely at a finite collection of time points. Growth curves and personal health series are typically recorded at a small collection of irregularly-spaced time points. The methods of analysis described in this section are most suitable for a long series that is observed at a large collection of equally-spaced time points.

Let $Y(t)$ be the value recorded at time $t = 1, \ldots, n$, so that the recording period is an interval of length n in suitable time units. Frequency is measured in cycles per recording interval, and the focus is on Fourier frequencies, which correspond to an integer number of cycles. The discrete Fourier transformation $\omega \mapsto \hat{Y}(\omega)$ at frequency ω is a complex number

$$\hat{Y}(\omega) = \sum_{t=1}^{n} e^{2\pi i \omega t/n} Y_t,$$

so that $\hat{Y}(0) = \hat{Y}(n) = Y_\cdot$ is the total, which is real in most applications. For integer frequencies $0 \leq \omega \leq n$, the real and imaginary parts are the linear combinations

$$\begin{aligned}
\hat{Y}(1) &= \sum_t \cos(2\pi t/n) Y_t + i \sum_t \sin(2\pi t/n) Y_t, \\
\hat{Y}(\omega) &= \sum_t \cos(2\pi \omega t/n) Y_t + i \sum_t \sin(2\pi \omega t/n) Y_t.
\end{aligned}$$

For real Y, the Fourier coefficients satisfy the aliasing identity

$$\hat{Y}(n-\omega) = \sum_{t=1}^{n} e^{2\pi i t(n-\omega)/n} Y_t = \sum_{t=1}^{n} e^{-2\pi i t\omega/n} Y_t = \overline{\hat{Y}}(\omega).$$

so that $\hat{Y}(\omega)$ and $\hat{Y}(n-\omega)$ is a complex-conjugate pair. Thus, $\hat{Y}(0) = \hat{Y}(n) = Y_\bullet$ is real, and if $n = 2m$ is even, the middle value $\hat{Y}(m)$ is also real.

7.1.2 Anova decomposition by frequency

If we set aside the zero-frequency component, and split the non-redundant components into real and imaginary parts, the Fourier transformation $\hat{Y} = FY$ is a linear transformation $\mathbb{R}^n \to \mathbb{R}^{n-1}$. The cosine and sine components of the Fourier matrix for frequency ω are the real and imaginary parts of roots of unity:

$$F^{(c)}_{\omega,t} = \cos(2\pi\omega t/n); \quad F^{(s)}_{\omega,t} = \sin(2\pi\omega t/n).$$

The identity $F\mathbf{1} = 0$ defines the kernel subspace, the rows are mutually orthogonal n-vectors, $FF' = (n/2)I_{n-1}$ is the identity of order $n-1$, and $2F'F/n = I_n - J_n/n$ is the orthogonal projection in $\mathbb{R}^n$ with kernel $\mathbf{1}$. The projection matrix $2FF'/n$ can be expressed as a sum of $\lfloor (n-1)/2 \rfloor$ rank-2 projection matrices, P_ω, one for each frequency, plus an additional rank-1 matrix for frequency $n/2$ if n is even. The net result is that the total sum of squares has an analysis-of-variance decomposition by frequencies

$$\sum_t (Y_t - \bar{Y}_\bullet)^2 = \sum_{\omega=1}^{\lfloor n/2 \rfloor} \|P_\omega Y\|^2 = \frac{1}{n} \sum_{\omega=1}^{n-1} |\hat{Y}_\omega|^2.$$

The last expression includes both conjugates $\hat{Y}_\omega, \hat{Y}_{n-\omega}$, so the sum of squares for frequency $1 \le \omega < n/2$ is

$$\|P_\omega Y\|^2 = 2|\hat{Y}_\omega|^2/n$$

on two degrees of freedom. As a function of ω, this is called the sample power spectrum, or the power spectrum.

7.2 Temperature spectrum

7.2.1 Spectral plots

The Central England daily temperature series for 248 years has a length of 90580 days. The analysis and interpretation are easier if the observation period is an integer number of years, so the partial year at the end is not included in the analysis. Harmonics associated with the annual cycle are expected to have large amplitudes, so it is helpful for plotting purposes to separate out the seasonal frequencies (integer multiples of 248) from the rest.

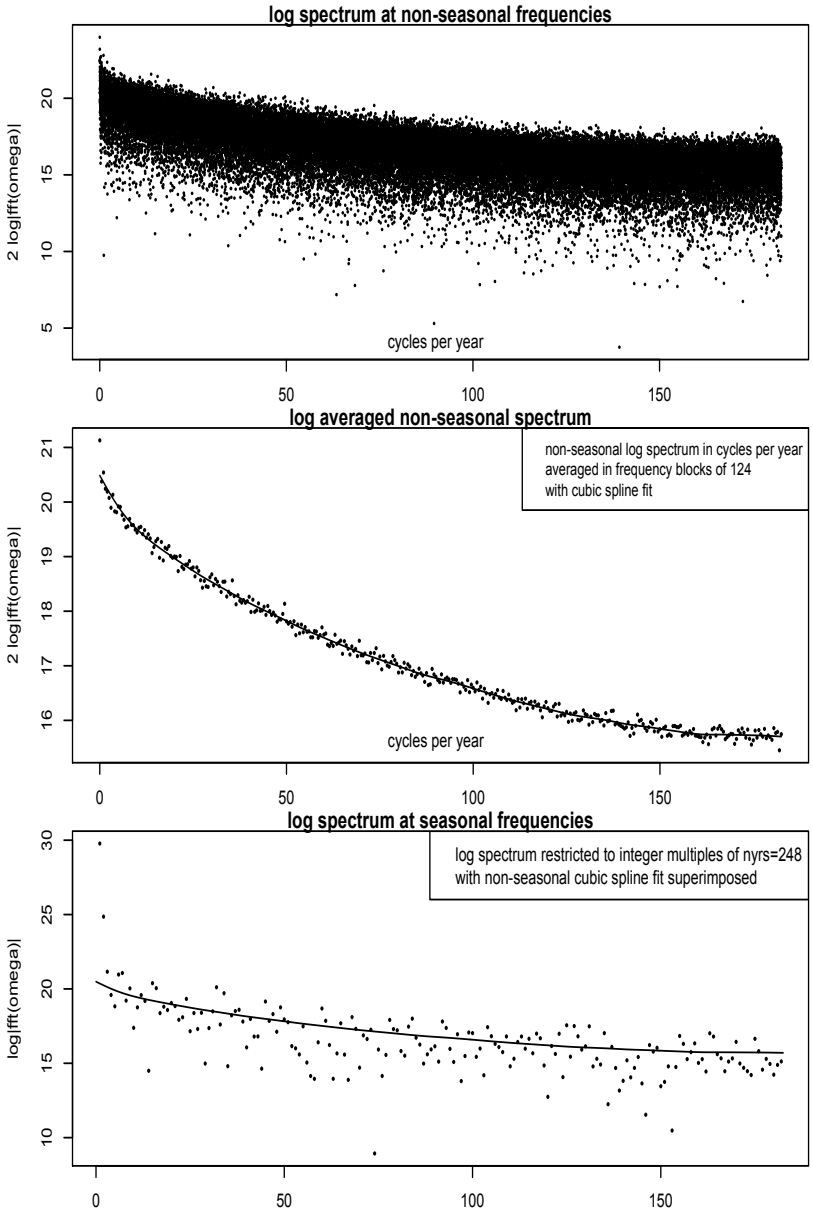

Figure 7.1: Log power spectrum for the Central England temperature series separated by seasonal and non-seasonal frequencies.

The first panel of Fig. 7.1 is a scatterplot of $\log|\hat{Y}_\omega|^2$ against frequency, which has been re-coded in units of cycles per year rather than cycles per observation period of 248 years. Seasonal frequencies have been excluded, partly because the annual and biannual coefficients are so large. The general trend is quite clear for the mean, but the high variability and density of points tends to obscure matters. Ordinarily, we should expect $\|P_\omega Y\|^2$ to be approximately exponentially distributed, in which case $\log\|P_\omega Y\|^2$ should have constant variance, $\pi^2/6 \simeq 1.28^2$, and the distribution should be skewed to the left. The plot is reasonably consistent with those expectations.

In the middle panel, the squared Fourier components $\|P_\omega Y\|^2$ have been averaged in consecutive non-overlapping frequency blocks, and the log averages are plotted against average frequency, again coded in cycles per year. In this manner, the variability is much reduced, so the trend in mean becomes clearly delineated. Note that the goal here is to estimate the spectrum, which is $E(|\hat{Y}_\omega|^2)$ as a function of ω, so all averaging takes place on that scale, not on the log scale.

Finally, the log spectrum for seasonal frequencies is shown in the third panel with the non-seasonal cubic spline superimposed for comparison. Apart from the first and second harmonics, the variation or energy at other seasonal frequencies decreases with frequency in conformity with the decrease observed in the second panel for non-seasonal frequencies. Certainly the variation at seasonal frequencies above three per year is not greater on average than the variation at neighbouring non-seasonal frequencies. The distinction between seasonal and non-seasonal seems to matter only for the first two annual harmonics.

These spectrum plots tends to emphasize the variation at higher frequencies, on the order of 20–150 cycles per year. However, it is the behaviour of the spectrum at low frequencies and the limiting behaviour as $\omega \to 0$ that is crucial for understanding long-range behaviour of temperatures. To clarify the picture, and to give greater emphasis to lower frequencies, Fig. 7.2 consists of the same points as the middle panel in Fig. 7.1, but the values are plotted against the square root of the frequency. To a first order of approximation, the log spectrum in linear in $\omega^{1/2}$ over the bulk of the frequency range.

7.2.2 A parametric spectral model

According to the theory discussed in the next section, the transformed coefficients $|\hat{Y}(\omega)|^2$ for non-seasonal frequencies are approximately independent exponential random variables. With this in mind, it is natural to fit a two-component additive spectral model

$$E|\hat{Y}_\omega|^2 = n\sigma_0^2 + n\sigma_1^2 \exp(-|2\pi\lambda\omega|^{1/2})$$

with three non-negative parameters $\sigma_0, \sigma_1, \lambda$ to be estimated. Aggregation by frequency blocks is helpful for plotting, but it is not needed for fitting this model, which, for fixed λ, is a gamma-type generalized linear model with unit dispersion. Additivity on the power-spectrum scale rather than on the log scale is natural if the temperature series is to be regarded as the sum $Y(t) = \sigma_0\varepsilon(t) + \sigma_1\eta(t)$

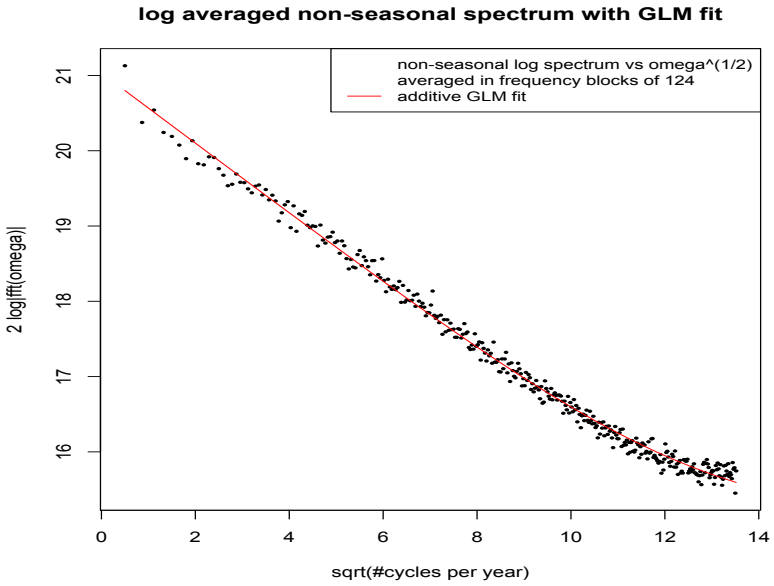

Figure 7.2: Log power spectrum plotted against $\omega^{1/2}$. The solid line is the additive GLM spectral fit $E|\hat{Y}(\omega)|^2 \propto 1 + 391\exp(-|0.219\omega|^{1/2})$.

of independent processes. Typically, ε is white noise, and η is an independent serially-correlated process whose sample paths are continuous in a suitable sense—either continuous with probability one or mean-square continuous. The spectral density proposed here decays sufficiently fast at high frequencies that η has continuous derivatives of all orders.

With ω measured in cycles per year, the fitted exponential model is

$$\hat{K}(\omega) = n\hat{\sigma}_0^2 + n\hat{\sigma}_1^2 \exp(-|2\pi\hat{\lambda}\omega|^{1/2}),$$

where $\hat{\lambda} = 0.0347$ years, or 12.67 days, $\hat{\sigma}_1/\hat{\sigma}_0 = 19.8$ for the volatility ratio, and $\hat{\sigma}_0 = 0.62$ for the nugget standard deviation in degrees Celsius. Note that he second component is formally the characteristic function of the α-stable distribution for $\alpha = 1/2$, so the associated covariance function is the density of that distribution.

The additive gamma model fits the non-seasonal power spectrum reasonably well, but it is not perfect. Small systematic deviations are apparent in Fig. 7.2 at low frequencies, and there is approximately 3% excess dispersion relative to the exponential distribution. In other words, the variance of the standardized spectral coefficients $|\hat{Y}_\omega|^2/\hat{K}_\omega$ is 1.032, while the exponential model predicts unit variance. This is a very small deviation in absolute terms, but, with 45 108 non-seasonal Fourier frequencies, a 3% deviation in variance is moderately unlikely.

In the residual plots shown in Fig. 7.3, the 3% deviation is too small to be noticed. Overall, the residual distribution seems to match the extreme-value

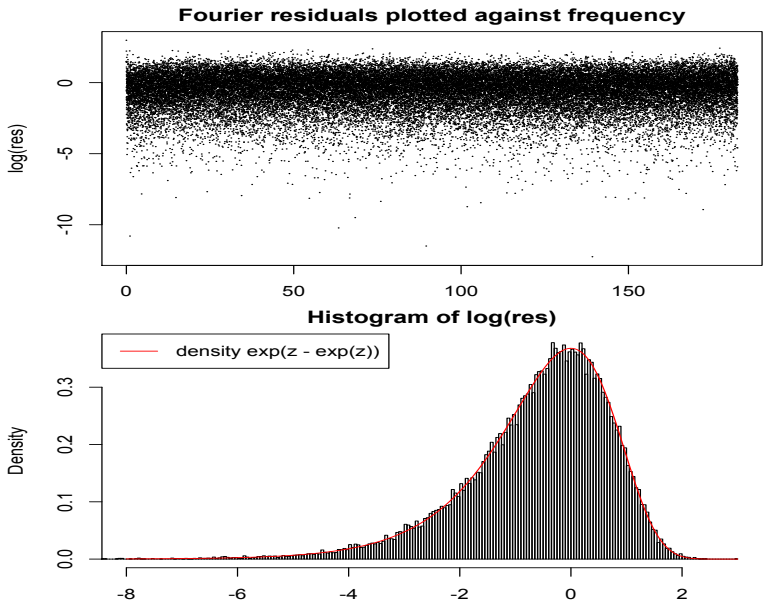

Figure 7.3: (a) Log Fourier residuals $\log\bigl(|\hat{Y}_\omega|^2/\text{fitted}_\omega\bigr)$ plotted against frequency; (b) Histogram of log residuals with theoretical extreme-value density superimposed.

distribution very closely. The mean of the log residuals is -0.584, and the variance is 1.661 versus the theoretical values $-\gamma = -0.577$ (Euler's constant), and $\sigma^2 = \pi^2/6 = 1.645$.

On the negative side, the ratios of the squared Fourier coefficients to the fitted values for the ten lowest frequencies are

$$19.67,\ 4.44,\ 9.19,\ 4.40,\ 2.15,\ 0.73,\ 4.28,\ 0.55,\ 0.36,\ 2.99,$$

and the next largest ratio is 11.3, which occurs at one of the highest frequencies. Despite the apparent success of this parametric model for the bulk of the frequency range, these low-frequency values are not consistent with the fitted model, which predicts independent standard exponential values. The expected value of the largest of n standard exponentials is approximately $\log(n) \simeq 10.7$, and the standard deviation is approximately $\pi/\sqrt{6} \simeq 1.3$. Given that it was so selected, the second-largest ratio is entirely consistent with the fitted model, but the first 4–5 Fourier coefficients are not.

The behaviour of the low-frequency Fourier coefficients is strongly tied to the behaviour of the covariance function or variogram at the longest lags. Bearing in mind the variogram phenomenon observed in the third and fourth panels of Fig. 6.5, which is compatible with a slow random walk or an autoregressive process with semi-range λ on the order of one millennium, it is natural to look for a corresponding phenomenon in the Fourier domain. The corresponding

phenomenon is an additive spectral component proportional to $1/(1+\lambda^2\omega^2)$, which is essentially a multiple of ω^{-2}. Inclusion of the inverse-square frequency as a further covariate the spectral model reduces the deviance by 52+ units, which is a substantial improvement to the fit, showing conclusively that the slow linear trend seen in the variogram plots is a real phenomenon and not a statistical artifact.

7.3 Stationary temporal processes

7.3.1 Stationarity

This section is concerned with real-valued processes that are defined pointwise on the domain, and specifically with stationary Gaussian processes on the real line. The first part of the statement means that to each point t in the domain there corresponds a value $Y(t)$, which is a real number. First-order stationarity implies that for each pair of points t, $t+h$ in the domain the values $Y(t)$ and $Y(t+h)$ have the same distribution. For a time series, the domain is either the integers or the real line, and translation implies that the domain is a group acting on itself by addition. More generally, stationarity implies that for each ordered n-configuration $\mathbf{t} = (t_1, \ldots, t_n)$ and each h-translate $\mathbf{t}+h = (t_1+h, \ldots, t_n+h)$, the values

$$Y[\mathbf{t}] = \big(Y(t_1), \ldots, Y(t_n)\big) \quad \text{and} \quad Y[\mathbf{t}+h] = \big(Y(t_1+h), \ldots, Y(t_n+h)\big)$$

have the same joint distribution in $\mathbb{R}^n$.

The focus is on Gaussian processes, which are defined by the mean function $\mu(t) = E(Y(t))$ and the covariance function $K(t,t') = \operatorname{cov}\big(Y(t), Y(t')\big)$, which is a symmetric positive semi-definite function on the domain. Stationarity implies that the mean is constant, $\mu(t) = \mu(0)$, and that $K(t,t') = K(|t-t'|)$ is a function of the temporal separation. For example, $e^{-|t-t'|}$ is the covariance function for the standard first-order autoregressive process, and $(1+|t-t'|)e^{-|t-t'|}$ is a related covariance function in the Matérn class.

The restriction to processes defined pointwise is not vacuous because there exist temporal processes that are not defined pointwise. For example, standard white noise is a zero-mean Gaussian process defined on domain *subsets* such that $\operatorname{cov}(Y(A), Y(B)) = \Lambda(A \cap B) < \infty$ is the Lebesgue measure of the intersection. The distribution is invariant with respect to translation, so the process is certainly stationary. However, the pointwise definition of stationarity is not satisfied because $Y(t)$ is not defined for such processes. If we attempt to define $Y(t)$ as a limit over subsets converging to $\{t\}$, then $Y(t) = 0$; if we regard $Y(A)$ as an integral of $Y(t)$ over A then $Y(t)$ cannot have finite variance. Neither of these implications is satisfactory.

The definition of stationarity given above is to be read conditionally as follows. If Y is defined pointwise, then Y is stationary if and only if, for every positive integer n, every n-configuration $\mathbf{t}$, and every $h \in \mathbb{R}$, the random variable $Y[\mathbf{t}+h]$ has the same distribution as $Y[\mathbf{t}]$.

For the more general definition of stationarity, the process is defined on an index set consisting of points or subsets or measures on the domain. Thus, the domain need not coincide with the index set of the process. To consider stationarity, the domain ($\mathbb{Z}$ or $\mathbb{R}$ or $\mathbb{C}$ or $\mathbb{R}^2$) is necessarily a group, and the index set is closed under domain translation. The process is first-order stationary if, for each object A in the index set, and for each point h in the domain, the value $Y(A)$ has the same distribution as the value $Y(A+h)$ taken on the h-translated object. Strict stationarity is defined in the same way for joint distributions. According to this definition, it is possible to make sense of the statement that $-|t-t'|$ is the covariance function for a Gaussian process or time series, sometimes called a generalized process because the index set does not coincide with the domain. Likewise for the functions $-|t-t'|^{1/2}$ and $-\log|t-t'|$. Moreover, these processes are strictly stationary. This definition paves the way to consider other group actions such as rigid motions or Euclidean congruences or similarity transformations, which are associated with isotropy and self-similarity.

7.3.2 Spectral density

7.3.3 Visualization of trajectories

To understand what the SD-1/2 process with spectral density $\exp(-\omega^{1/2})$ looks like, i.e., how a typical trajectory behaves as a function, it is helpful to compute, simulate and plot. The first step is to compute the covariance function by inversion of the spectral density. In general, this is a non-trivial computational exercise. Fortunately, this spectral density is a special case of the characteristic function of the α-stable class. The series expansion for the density (Feller 1971, vol II, p. 582) can be simplified for $\alpha = 1/2$. We remark only that K is strictly positive, and monotone as a function of temporal separation. It is infinitely differentiable at all points on the real line, but it is not complex-analytic in any neighbourhood of the origin. Accuracy to two or three significant decimal digits suffices for graphical representation of the covariance function, but at least eight-digit accuracy is needed to simulate trajectories.

Four covariance functions are shown in Fig. 7.4. At first glance, the differences among them appear to be slight: all four are continuous, symmetric and are equal at the origin and at ± 1. The behaviour in a neighbourhood of the origin is an important characteristic, which is shown in 5x-magnified form on the right of each panel. The Matérn functions have zero, one and two derivatives at the origin, whereas the fourth has infinitely many. The first, $1-|x|+o(x)$, is easy to see by inspection, but the others are not, even in magnified form: the approximate behaviour for $\nu = 1$ is $1 + x^2 \log|x|/2 - 0.31x^2 + o(x^2)$, so the first derivative is zero and the second does not exist. The behaviour in the tail is the characteristic that distinguishes long-range dependent processes from short-range. Once again, this is easier to see in hindsight than in foresight—especially if zero has not been included for visual reference in the graph.

All four curves in Fig. 7.4 are non-negative and have have finite integrals,

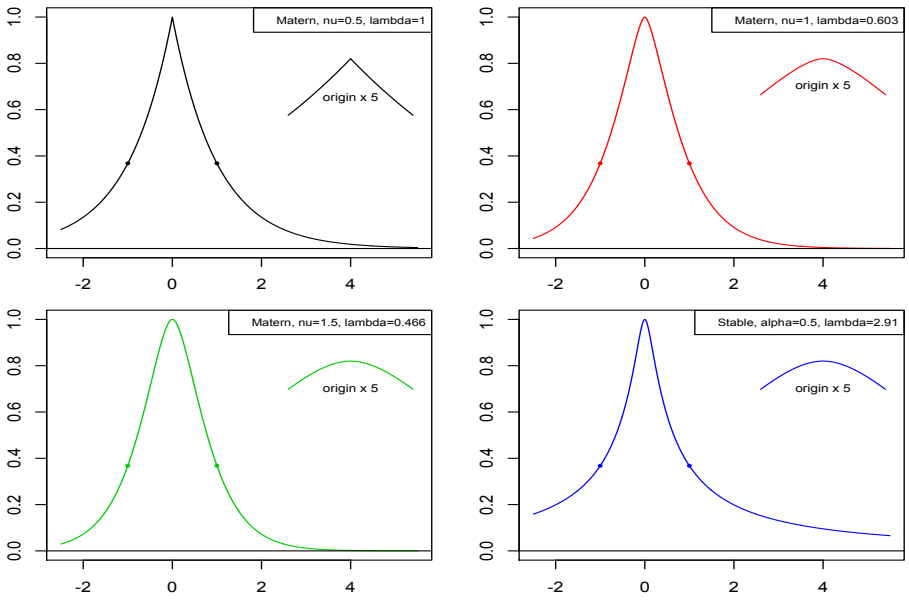

Figure 7.4: Four covariance functions standardized to have unit variance and lag-one autocorrelation e^{-1}. The fourth is a multiple of the α-stable density function for $\alpha = 1/2$.

so each is proportional to a symmetric probability distribution on the real line. The integrals are 2.0, 1.89, 1.86 and 4.58 respectively, or more generally, 2λ, $\pi\lambda$, 4λ and $\pi\lambda/2$ for the scaled versions. The first Matérn covariance is a multiple of the Laplace density; the SD-1/2 covariance is a multiple of the α-stable density for $\alpha = 1/2$.

Given computer code for the covariance function, the covariance matrix Σ for the process at 1000 points may be computed, followed by simulation of a 1000-component Gaussian variable $Y \sim N(0, \Sigma)$. These are the values of the process at the selected points in the domain. Special cases can be simulated more efficiently, but this straightforward recipe suffices for present purposes. Each of the curves in Fig. 7.5 is plotted using the values $(x, Y(x))$ at 1000 equally-spaced points in the interval $(0, 10)$.

To establish a 'normal range' of patterns, Fig. 7.5 shows the trajectories of three Matérn processes in the 'typical' index range, plus the SD-1/2 process. Each family has a variance parameter and a range parameter, both of which are strictly positive real numbers. For purposes of comparison, each covariance function is scaled to have unit variance, and the same lag-one autocorrelation $e^{-1} = 0.368$, which matches the standard order-one autoregressive process shown in the top panel. Each of the Matérn processes has a distinct character, with continuous derivatives of order zero, one and two for $\nu = 0.5$, $\nu = 1.0$ and $\nu = 1.5$ respectively. Visually speaking, the differences among these three are

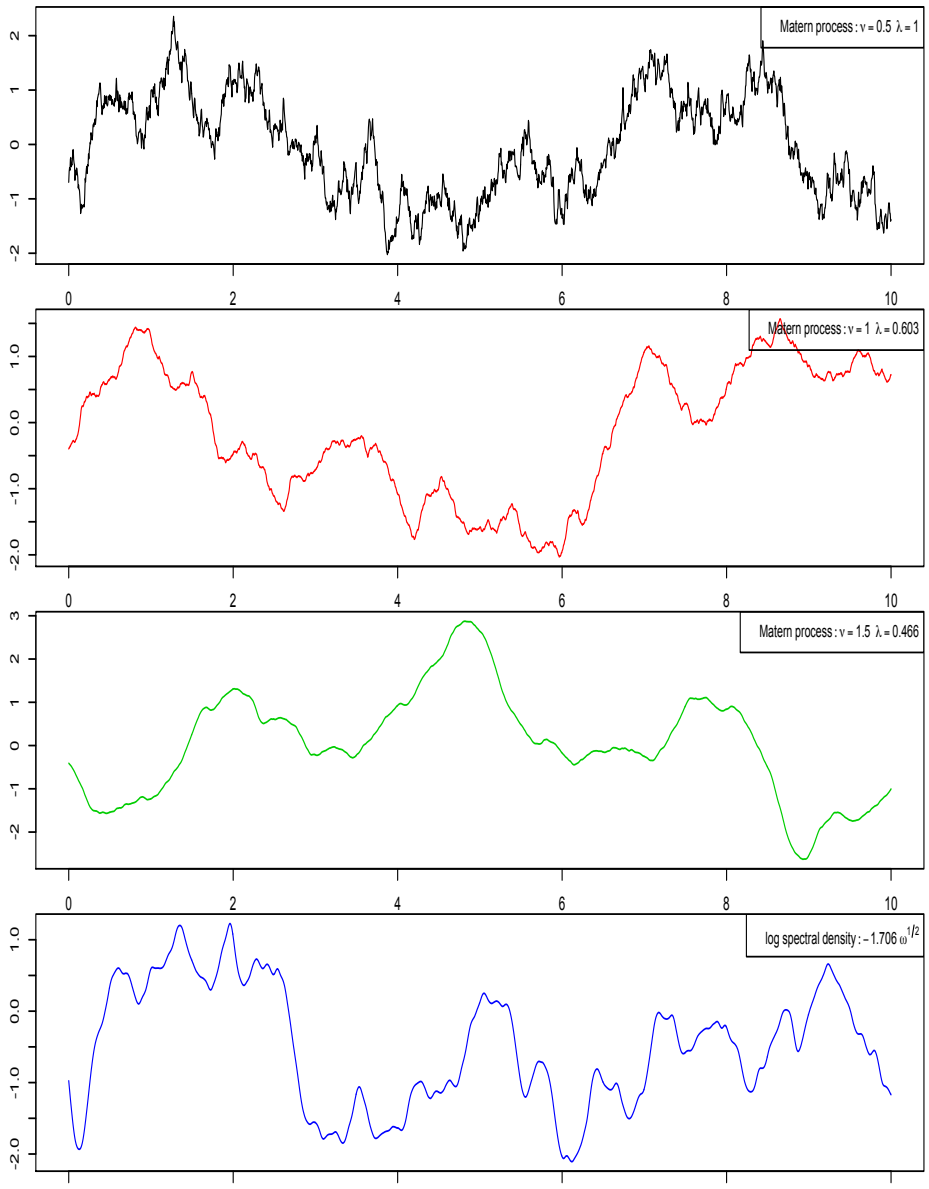

Figure 7.5: A comparison of trajectories of four stationary continuous-time processes, three in the Matérn class, and one specified by its spectral density $e^{-|\omega|^{1/2}}$. The standard Matérn covariance function is $K_\nu(x) = \|x\|^\nu \mathcal{K}_\nu(\|x\|)$; special cases include $K_{1/2}(x) = e^{-\|x\|}$ and $K_{3/2}(x) = (1 + \|x\|)e^{-\|x\|}$. The spectral density is $1/(1 + \omega^2)^{\nu+1/2}$. For visual comparison over the interval $(0, 10)$, all four processes are standardized to have unit variance and the same lag-one autocorrelation.

concerned with the degree of smoothness, which is a local property. The SD-1/2 process has its own distinct character. It has continuous derivatives of all orders so it is smoother than any Matérn process, even those with $\nu > 3/2$. However, its medium-range oscillations are distinctly more pronounced, and somewhat similar to those of the AR1 process ($\nu = 1/2$).

In addition, although it is hard to point to visual consequences in the trajectories, the long-range autocorrelation of the SD-1/2 process is algebraic of order $|x - x'|^{-3/2}$, whereas the long-range Matérn correlations decay faster than any polynomial. As a result, the lag 5–10 autocorrelations are sizable 0.073–0.031 for the SD-1/2 process, but negligible for all Matérn processes and decreasing as a function of ν. Long-range dependence appears to be universal for processes in nature, both for time series and for spatial processes. For such applications, we should bear in mind that each finite-range Matérn covariance has exponentially-decreasing tails whereas the SD-1/2 covariance has regularly-varying tails of order $|x - x'|^{-3/2}$. All four covariance functions have finite integrals, so all four processes are short-range dependent. On the other hand, each Matérn process has a well-behaved infinite-range limit, whereas the SD-1/2 process does not.

Algebraic, or inverse-polynomial, decay of autocorrelations is known as long-range dependence. One consequence is that the sample average over the interval $(0, t)$

$$\bar{Y}_{(0,t)} = t^{-1} \int_0^t Y(s)\, ds \quad \text{or} \quad \bar{Y}_{1:t} = t^{-1} \sum_{s=1}^{t} Y_s,$$

has a variance that tends to zero as $t \to \infty$ at a slower rate than $O(t^{-1})$. The rate is $O(t^{-1})$ for short-range dependent series, including every Matérn process, but only $O(t^{-1/2})$ for the SD-1/2 process. Empirically, we find that the variance of the temperature average over randomly-sampled blocks of h successive years is as follows:

	Block length in years					
h	4	8	16	32	64	128
$C_h \operatorname{var}(\bar{Y}_h)$	0.213	0.158	0.133	0.087	0.058	0.036
$h^{1/2} C_h \operatorname{var}(\bar{Y}_h)$	0.427	0.446	0.533	0.491	0.465	0.409
$h C_h \operatorname{var}(\bar{Y}_h)$	0.853	1.261	2.130	2.778	3.719	4.624

In this table $\operatorname{var}(\bar{Y}_h)$ is the sample variance of 5000 randomly-sampled block averages. The factor $C_h^{-1} = 1 - h/248$, which is the average overlap between pairs of blocks of length h, is a finite-population bias-correction factor for sample overlap. Whole-year blocks were used to eliminate the effect of seasonal cycles. It is apparent from the table that $\operatorname{var}(\bar{Y}_h) \propto h^{-1/2}$ is the dominant term for the variance of block averages, at least up to $h \simeq 128$ years.

7.3.4 Fourier transform

When the circumstances permit it, i,e., when a series is recorded over a large number of equally-spaced points, the advantages of working in the frequency

domain are considerable. For a stationary series with covariance function K, Whittle (19??) shows that the Fourier coefficients are approximately Gaussian and approximately independent for large n, with moments

$$E\big(\hat{Y}(\omega)\overline{\hat{Y}(\omega')}\big) = \begin{cases} n\hat{K}(2\pi\omega/n) + O(1) & \omega = \omega', \\ O(1) & \omega \neq \omega', \end{cases}$$

where $\hat{K}$ is the spectral density of K. Using this approximation, we may treat the frequencies as observational units, and the Fourier coefficients as independent complex-Gaussian observations. In the standard technical sense, the squared moduli $|\hat{Y}(\omega)|^2$ are sufficient for the spectral density. To accommodate the seasonal cycle in the present application, it is necessary either to restrict attention to non-seasonal frequencies or to eliminate the first few seasonal harmonics.

7.4 Exercises

7.1 Let $Y_1, \ldots, Y_n$ be independent and identically distributed standard exponential variables, and let $0 \leq Y_{(1)} \leq Y_{(2)} \leq \cdots \leq Y_{(n)}$ be the order statistics. Show that $Y_{(1)} \geq t$ if and only if $Y_i \geq t$ for $1 \leq i \leq n$, and deduce that $nY_{(1)}$ is exponentially distributed with unit mean. Hence or otherwise, show that the increments $(n-r)(Y_{(r+1)} - Y_{(r)})$ are independent and identically distributed. Find the mean and variance of the maximum $Y_{(n)}$, with asymptotic values for large n.

7.2 Use `fft()` to compute the Fourier coefficients for the temperature series on a whole number of years, identify and remove the frequencies that are seasonal, average the power-spectrum values in successive non-overlapping frequency blocks of a suitable size, and plot the log averages against the square root of the frequency in cycles per year.

7.3 For the non-seasonal frequencies, use `glm()` to fit the additive exponential model

$$E(|\hat{Y}(\omega)|^2) = \beta_0 + \beta_1 \exp(-|2\pi\lambda\omega|^{1/2})$$

for various values of λ in the range 7–14 days or 0.02–0.04 years. In this setting, the distributional family is `Gamma`, the link is `identity`, and the dispersion parameter is one. Plot the residual deviance against λ to find the maximum-likelihood estimate. Check that the fitted coefficients are non-negative. Superimpose the fitted curve on the graph of log-block-averages in the previous exercise. Plot the standardized residuals (log-ratio of observed to fitted) against frequency, and comment on any departures that are evident.

7.4 Include the inverse-square frequency as an additional covariate in the exponential model for the power spectrum. In principle, this means re-computing $\hat{\lambda}$. Compute the Wilks statistic, which is the reduction in deviance or twice the increase in log likelihood. Also compute on the Wald statistic, which is the squared ratio of the ω^{-2}-coefficient to its standard error as given by the inverse

Fisher information matrix. Recall that the dispersion parameter is one, which is not the default in `summary()`. Standard asymptotic theory for large sample sizes tells us that the difference between these two statistics is $o_p(1)$, i.e., that the difference tends to zero as $n \to \infty$, and also that the null distribution is χ^2_1 for both. In this setting the sample size is the number of non-seasonal frequencies. Comment on any discrepancy between theory and practice in this instance, and provide an explanation.

7.5 Ordinarily, Wald's likelihood ratio statistic is essentially the same as Wilks's statistic, which in one-parameter problems, is the squared ratio of the estimate to its standard error. But there are exceptional cases where a substantial discrepancy may occur, and variance-components models provide good examples. In order to understand the source of the discrepancy, simulate data with simple structure as follows:

```
set.seed(3142);  n <- 1000;  x <- 1:n
rx2 <- 1/x^2;  beta <- c(1,20);
X <- cbind(1, rx2); mu <- as.vector(X%*%beta)
y <- -log(runif(n))*mu
```

The null hypothesis is that $\mu \propto \mathbf{1}$ is constant, and the alternative is that $\mu = X\beta$ for some β with non-negative components. Test this hypothesis using Wilks's likelihood ratio statistic, and also using the Wald statistic. Recall the exponential assumption, which implies that the dispersion parameter is one.

7.6 If you used the function `glm(y~rx2, family=Gamma(link=identity)` in the preceding exercise, you may have experienced a failure to converge. Write your own Newton-Raphson function with steps on the log scale, which forces the β-components to be strictly positive. As part of this exercise, you will need to compute the Fisher information matrix, `t(X/mu^2) %*% X` for β. Report the value of I_β at the null hypothesis $\hat\beta_0$ and also at $\hat\beta$. What does this tell you about the Wald-Wilks discrepancy?

7.7 For $0 < \alpha \leq 2$, the α-stable distribution on the real line is symmetric with characteristic function $e^{-|\omega|^\alpha}$. For the sub-range $0 < \alpha < 1$, Feller (1971, eqn. 6.5) gives the series expansion for the density

$$p(t;\alpha) = \Re \frac{i}{\pi t} \sum_{k=0}^{\infty} (-1)^{k+1} \frac{\Gamma(k\alpha+1)}{k!} t^{-k\alpha} e^{-\pi i k\alpha/2}$$

which is convergent for $t > 0$. The goal of this exercise is to simplify the density for $\alpha = 1/2$ by splitting the sum into four parts according to $k \pmod 4$. Show that one of the four parts is zero, that the odd parts may be combined into a multiple of $t^{-3/2}\sin(1/(4t) + \pi/4)$, and that the remaining part is $O(t^{-2})$ as $t \to \infty$.

7.8 From the cosine integral $\int \cos(\omega t) e^{-|\omega|^\alpha}\, d\omega$, deduce that the α-stable density has a Taylor series at the origin which begins

$$\log p(t;1/2) = \text{const} - 60t^2 + O(t^4).$$

Find the general term in this expansion and deduce the radius of convergence.

Example 8

8.1 Out of Africa

This chapter is concerned with the linguistic hypothesis and the data analysis in the paper *Phonemic diversity supports a serial founder effect model of language expansion from Africa* published by Q.D. Atkinson in *Science* (15 April 2011). It is recommended that you read the paper and the supplementary material, which are available at

`http://www.sciencemag.org/content/332/6027/346.full.pdf`

The data and supplementary files can be found at `.../OOAdir/`

Like the genetic thesis for human migration and evolution, the 'out-of-Africa' thesis for linguistic diversity holds that language evolved somewhere in Africa, and diffused from there to Asia, Europe and elsewhere as populations split and migrated. Since the genetic and linguistic diversity of a population is intrinsically related to its size, a small migrating subset carries less diversity than the population from which it originated. Accordingly, a subpopulation that splits and migrates carries less diversity than the descendants of the ancestral population that remains. Although tones and sounds are continuously gained and lost in all languages, the loss is supposedly higher for small migrating founder populations than for the ancestral population. In this way, the diversity of sounds becomes progressively reduced as the distance from the origin increases.

Atkinson's paper is concerned with the hypothesis that human language developed in a single location and spread from there by migration. He aims to test that hypothesis by examining the relationship between the diversity of sounds in 504 extant languages and their geographic distance from a putative origin in Africa or elsewhere, taking account of speaker population size.

8.2 Phoneme inventory

The data on which Atkinson bases his analysis is a list of 504 languages from various parts of the world. The list of languages is not exhaustive, nor is it close to geographically uniform with respect to current population density. The diversity measure is not a measure of variability of sounds in the ordinary statistical sense, nor is it an inventory or list of sounds, but simply the number of

distinct phonemes that the language employs. There are three distinct values for vowel inventory, three for tone inventory, five for consonant inventory, and 40 for total phoneme inventory. In principle, inventories should be all be non-negative integers, but the values have been standardized or normalized in an unspecified way. The sample means are close to zero, and the sample variances for the three phoneme constituents are close to one.

Data on phoneme inventory size were taken from the World Atlas of Language Structures, (WALS). The file `OOAdir/S1.dat` contains the main part of Anderson's data, which is the list of 504 languages together with the following twelve variables

1. `Lname` Language name: text, e.g. Abkhaz, Aikan??, B?©t?©, . . ., Zuni
2. `WALS` three character code, e.g. abk, aik, bet,...
3. `Fam` Language Family: text, e.g. Arawakan, Indo-European, Sino-Tibetan,...
4. `Lat` Latitude as a decimal number, e.g. -12.67
5. `Long` Longitude as a decimal number, e.g. -60.67 (meaning $60^\circ\ 40'$W)
6. `Nvd` Normalized vowel diversity based on WALS feature No. 2
7. `Ncd` Normalized consonant diversity based on WALS feature No. 1
8. `Ntd` Normalized tone diversity based on WALS feature No. 13
9. `Tnpd` Total normalized phoneme diversity
10. `Iso` ISO codes (one or more three-character codes)
11. `Popn` Estimated speaker population: integer 1–873 014 298
12. `Dbo` Distance in km. from Atkinson's best-fit origin

Regardless of its geographical range, each language is associated with a single point on the sphere, which is not necessarily the geographic centroid of the speaker domain. International languages such as English, Spanish and French are associated with their ancestral capitals. For example, English is Indo-European and is located at latitude 52.0, longitude 0.0; the speaker population 309M is dominated by parts of the former empire. Spanish is located at 40.0N, 4.0W with a population size 322M most of whom are in Latin America; Mandarin is located at 34.0N, 110.0E, with a population of 873M. The guiding principle for inclusion is not evident. Among European languages, Albanian, Basque, Catalan, Breton, Romansch and Saami are included, but not Portuguese (178M), Italian (60M), Dutch (22M), Ukranian (37M), Belarusian, Slovak, Slovene, Serbian or Croatian.

For whatever reason, phoneme values are rounded and normalized. In addition, the number of distinct values for each variable is very limited. For example, English, French, German and Korean have exactly the same diversity profile $(1.39, 0.12, -0.77)$, which is shared by Turkish and 21 other languages; The values in the file are reported to seven or eight decimal digits. Donegal Irish shares its consonant-dominant diversity profile $(-0.48, 1.80, 0.18)$ with 13 other languages including Kwakw'ala, Lezgian and Saami.

8.3 Distances

The languages were partitioned into six continental groups, Africa, Europe, Asia, Oceania, N.Amer, and S.Amer. These coincide closely with the geographic continents, but not exactly so: Malagasy, the national language of Madagascar, belongs to Oceania, not Africa.

For his analyses, Atkinson used great circle distances between points x, x' for pairs of languages belonging to the same continental group. Otherwise, for languages in different continental groups, distances were measured for the shortest path passing through certain choke points (supplementary material, Fig. S8). For example, the shortest linguistic path from Europe to N. America consists of three great-circle arcs passing through Istanbul and the Bering Strait. The great-circle distance from `Aghem = (10.0,6.67)` in the Congo to `Malagasy = (47.0,-20.0)` in Madagascar is 5014 km, but since the latter belongs to region 4, the linguistic distance through Cairo and Phnom Penh is 18475 km.

The R-executable file `OOAdir/S1.R` contains the following commands:

```
S1 <- read.csv(file="OOAdir/S1.dat")
S4 <- read.csv(file="OOAdir/S4.dat", header=FALSE)
```

The same file also contains the coded list `lregion` of 504 linguistic groups, the geocoordinates of a small number of major `cities` including the choke points `chokes`, some code for geographic plotting, and functions for computing distances, as follows.

1. `gcdist(x1, x2)` great circle distance in km:
   `gcdist(Aghem, Malagasy) = 5013.585`
   `gcdist(Paris, Chicago) = 6651.991`
   The format used here for geocoordinates is `x=(long, lat)` in decimal degrees.
2. `chokedist(x1, x2, r1, r2)` linguistic distance:
   `chokedist(Aghem, Malagasy, 1, 4) = 18474.65`
   `chokedist(Paris, Chicago, 2, 5) = 15761.26`
3. `vdist(Dublin, 2)` a list of 504 linguistic distances from `Dublin=(-6.25, 53.33)`, regarded as a member of linguistic region 2. For great-circle distances, use `vdist(Dublin, 0)`. (The 504 language coordinates are assumed to be in `S1$Long, S1$Lat`.)

8.4 Maps and scatterplots

The file `OOAdir/S1.R` also contains standard R code for plotting world maps and subsets thereof, using `ggplot2` and `rgeos` (references ???). For illustration, Fig. 8.1 shows the location of the African and European languages used in the analysis. The African languages are heavily concentrated in equatorial Africa, roughly 10°S to 15°N. Among the eleven African countries south of 10°S, only

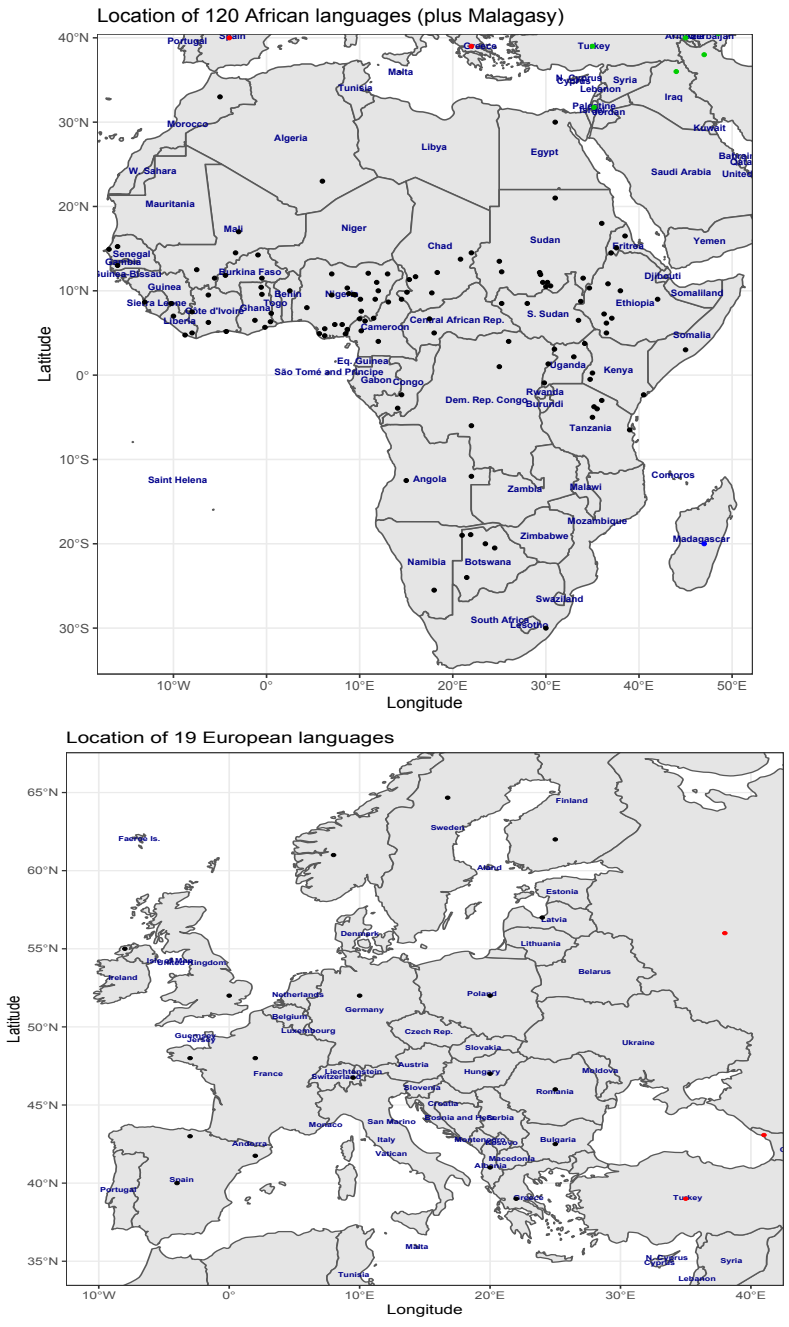

Figure 8.6: Geographic distribution of African and European languages in Atkinson's sample.

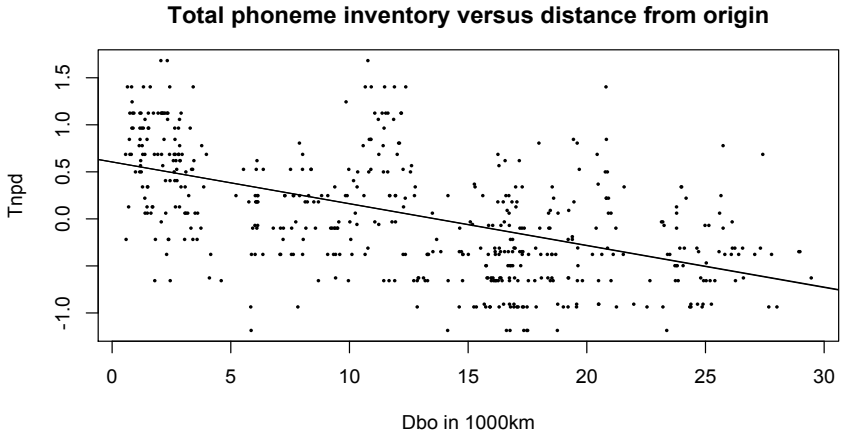

Figure 8.7: Total phoneme diversity plotted against the distance from the best-fitting origin as reported by Atkinson at 9.5°E, 1.25°S

four—Angola, Botswana, Namibia and South Africa— are represented, and Botswana has four or five. Madagascar is represented by Malagasy, whose roots are non-African. Similar anomalies are evident in the European sample.

The putative linguistic origin is a geo-coordinate point x_0 for which total phoneme diversity for language i satisfies

$$E\big(\mathrm{Tnpd}_i\big) = \beta_0 + \beta_1 \|x_i - x_0\|,$$

where, $\|x_i - x_0\|$ is the linguistic distance between the origin and the geo-coordinate of language i. The least-squares criterion is thus

$$\sum_i \big(\mathrm{Tnpd}_i - \beta_0 - \beta_1 \|x_i - x_0\|\big)^2$$

which is to be minimized with respect to the four parameters β_0, β_1 plus the two components of x_0. Atkinson reports the best-fitting origin at 1.25°S, 9.30°E in West Africa: see Fig. 8.4.

Support for Atkinson's thesis, that phoneme diversity—or more correctly phoneme inventory—decreases with distance from the origin is best illustrated by the scatter plot of total phoneme diversity against distance from the best-fitting origin. The least-squares fitted line, which is superimposed in Fig. 8.2, has a definite negative slope.

What the scatterplot Fig. 8.2 fails to show is that all of the distances up to 4.5 units are in Africa, many of those in the interval 5.5–8.5 are in Europe, most of those in the interval 5.5–13.5 are in Asia, and so on. Consequently, it is natural to plot each continental group separately, which is done in Fig. 8.3. This exercise reveals that the relation between distance and phoneme inventory is negative chiefly in Africa. The negative least-squares slope for Oceania is

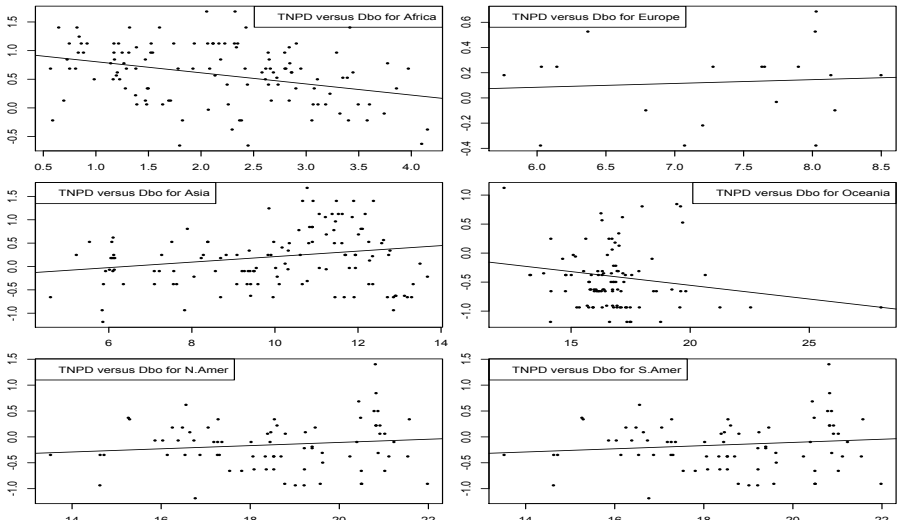

Figure 8.8: Total phoneme diversity plotted against the distance from the best-fitting origin for each of the six linguistic groups separately

almost entirely due to the single point (Malagasy), which has high leverage on account of its remoteness, and a low phoneme inventory. For each of the other linguistic groups, the relation between inventory and distance is either negligible or positive. While the aggregate scatterplot in Fig. 8.1 seems to support Atkinson's thesis, the disaggregated continent-by-continent plots paint a different picture. This is an instance of Simpson's paradox for continuous data, with positive slopes on most continental subsets, but a negative slope overall.

8.5 Point estimates and confidence regions

8.5.1 Simple version

If we buy into the out-of-Africa linguistic theory, it is natural to seek a region of plausible origins of language. In order to do this, it is necessary to know something about the statistical properties of the observations that are available. Independence of components is certainly not a reasonable assumption for this setting, but it is the easiest place to start, and it suffices to illustrate the method in its simplest form. For illustration, we assume that total phoneme inventory is related to population size and distance to the origin as follows

$$E\big(\mathrm{Tnpd}_i\big) = \beta_0 + \beta_1 \|x_i - x_0\| + \beta_2 \log(\mathrm{Popn}_i),$$

where $\|x - x_0\|$ denotes the linguistic distance. Given the range of speaker populations, linearity in log population size seems more reasonable than linearity in population size, which is in agreement with Atkinson's principal analysis. For

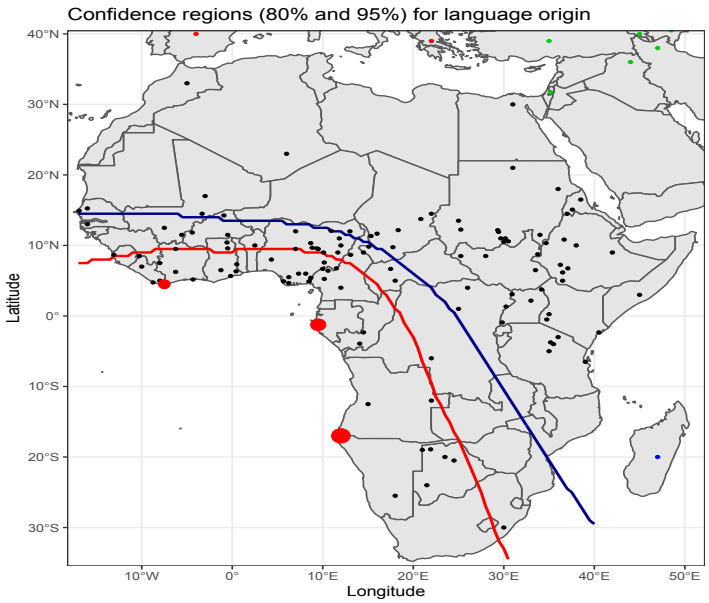

Figure 8.9: Point estimate and confidence regions (80% and 95%) for the language origin, assuming independent observations. The RSS values for the three coastal marked points are 143.4, 143.1 and 143.0 in west-to-east order.

the moment, we assume also that the components are independent with constant variance σ^2.

For arbitrary fixed origin, the model is linear in the remaining three regression parameters, so the least-squares estimates can be obtained in the standard way. Denote by $\mathrm{RSS}(x)$ the residual sum of squares on $n-3=501$ degrees of freedom for fixed x. The point $\hat{x}$ that minimizes the residual sum of squares is the non-linear least-squares estimate. For these data, the minimum over the rectangular grid that covers Africa occurs at the south west corner, near 17°W, 35°S in the south Atlantic. However, the RSS function varies little throughout the bight of Africa, and is almost constant along the coast from Liberia to Cape Town. For this exercise, we restrict the parameter space to terra firma. The minimum over continental Africa occurs on the coast near the border between Angola and Namibia, roughly at 12°E, 17°S as indicated in Fig. 8.4, with $\mathrm{RSS} = 143.02$. By the narrowest of margins, Atkinson's fitted point on the Congo coast appears to be a local minimum with $\mathrm{RSS} = 143.07$.

The standard recipe for the formation of a confidence set in non-linear least-squares problems uses a selected contour of the restricted residual sum of squares function function $\mathrm{RSS}(x)$ as the boundary. The residual mean square $s^2 = \mathrm{RSS}(\hat{x})/(n-5)$ serves as the variance estimate, which is distributed approximately as $\sigma^2\chi^2_{n-5}$ under the stated assumptions. Moreover, s^2 is distributed approximately independently of the difference $\mathrm{RSS}(x_0) - \mathrm{RSS}(\hat{x})$, which is dis-

tributed as $\sigma^2\chi^2_2$—on two degrees of freedom because the parameter space for x is locally a two-dimensional manifold. Thus, the mean-square ratio $(\mathrm{RSS}(x) - \mathrm{RSS}(\hat{x}))/(2s^2)$ is distributed according to Fisher's $F_{2,n-5}$ distribution. Accordingly, the region

$$\left\{x : \frac{\mathrm{RSS}(x) - \mathrm{RSS}(\hat{x})}{s^2} \leq 2F_{2,n-5,\alpha}\right\}$$

is a $1-\alpha$ confidence region for the linguistic origin. For $\alpha = 0.05$, the F-percentile is 3.01, so the right hand side is a little over 6.0. Atkinson uses a factor of four (BIC units) in place of six at this point, which gives only 86% coverage. Four BIC units is the 95% coverage factor for one degree of freedom, not for two.

Fig. 8.4 shows best-fitting origin at the Angolan-Namibian border, together with the 80% and 95% confidence regions computed according to the above formula. The 95% region includes most of western and southern Africa. If the unrestricted maximum had been used, confidence regions at low confidence levels would cover water only, which is a difficult case for a linguist to make. However, the unrestricted 95% confidence region matches reasonably closely the restricted 80% region.

Shortcomings

The preceding analysis overlooks the fact that one of the regularity conditions fails. The restricted maximum occurs at a 1D-boundary point on the coast, so the local 2D manifold argument fails. If the linguistic hypothesis also stated that the origin must be a coastal point, the 1D argument would naturally prevail. But that is not a part of the thesis, so it seems preferable to use the more conservative 2D allowance. An alternative option is to resort to simulation—but that is not an easy answer nor a satisfactory answer. In any event, there are more consequential effects that have so far been ignored.

8.5.2 Accommodating correlations

In statistical modelling, the difference between two means $E(Y_i) - E(Y_j)$ is associated with the difference $x_i - x_j$ between their recorded covariates: $\mu_i - \mu_j = (x_i - x_j)'\beta$. The difference between two covariances $\mathrm{cov}(Y_i, Y_{i'})$ and $\mathrm{cov}(Y_j, Y_{j'})$ is associated with the difference $R_{i,i'} - R_{j,j'}$) between their recorded relationships, usually but not necessarily in a linear manner: $\Sigma_{ii'} - \Sigma_{j,j'} = (R_{ii'} - R_{j,j'})'\tau$, where τ is a list of variance components.

For the current setting, the available covariates are the population size, continental group and the geographic location $i \mapsto x_i$. A relationship is a function on pairs of observational units, and the most obviously relevant relationships are inter-point distance $(i,j) \mapsto \|x_i - x_j\|$ and language family $(i,j) \mapsto F_{ij}$ as a Boolean matrix such that $F_{ij} = 1$ if i, j belong to the same family, and zero otherwise. These are both symmetric $n \times n$ matrices, so it is only natural that they should occur in the specification of the response covariance matrix. For

this project, we consider only the simplest additive model such that

$$\mathrm{cov}(Y_i, Y_j) = \sigma_0^2 \delta_{ij} + \sigma_1^2 F_{ij} + \sigma_2^2 e^{-\|x_i - x_j\|/\lambda},$$

depending on three variance components and one range parameter λ. As it happens, linguistic distance is a little more effective than great circle distance, and for that choice the fitted range is $\hat{\lambda} \simeq 820$km. The linguistic family effect is not negligible, but the distance effect dominates. There are also linguistic sub-families, whose effects are not taken into account in this analysis.

The likelihood-based nominal 95% confidence region is the set of all candidate source points whose log likelihood is sufficiently high compared with the maximum, i.e.,

$$\{x : 2l(\hat{x}) - 2l(x) \leq \chi^2_{2,0.95}\}.$$

Here $l(x)$ denotes the profile log likelihood maximized over all other parameters. For the model suggested here, the 95% confidence region includes all of Africa except for a portion of lower Egypt; the 99% region also includes the Levant (Israel, Syria, Turkey, Jordan, Iraq) and all of Europe except for Russia and the Caucasus.

Generally speaking, failure to accommodate correlations has the effect of making the data seem more informative than they are, so the resulting confidence intervals are unrealistically narrow. Thus, it is no surprise that Fig. 8.4 is misleading in its portrayal of the strength of information in the data.

Three points of clarification

The code used for computing the log likelihood for one candidate point `x0` belonging to linguistic region `lregn` has two parts:

```
ldist <- vdist(x0, lregn)  # vector of distances from x0
fit <- regress(Tnpd~ldist+log(Popn), ~Fam+V, data=S1)
```

Here `S1$Fam` is the linguistic family coded as a factor, and `V` is a matrix with components $V_{ij} = \exp(-\|x_i - x_j\|/\lambda)$, which do not depend on `x0`. As it stands, this code is both computationally inefficient and technically incorrect on two counts. The efficiency can be improved substantially by including the optional argument `start=fit$sigma`, which makes the previously-computed variance components available as the starting point for iteration.

The first technical issue is that the default likelihood function that is maximized by `regress()` is the REML likelihood for the observation Y in the space $\mathbb{R}^n/\mathcal{X}$ of residuals modulo the subspace $\mathcal{X}$ of mean values. Ultimately, our goal is to compute a likelihood ratio for one candidate center versus another, and the problem with the code as shown is that the mean-value subspace for one candidate point is not the same as the mean-value subspace for another candidate. The REML log likelihoods are not comparable as log likelihoods. In order for this to be done correctly, it is necessary to use the optional argument `kernel=K` to over-ride the default kernel. While `K=0` and `K=1` are valid zero

and one-dimensional options, the more natural choice is the two-dimensional intersection subspace `K <- model.matrix(~log(Popn), data=S1)`.

The second technical issue is that the log likelihood for a given `x0` should also be maximized over λ, which is a substantial computational overhead. For simplicity in the analysis described above, $\lambda = 820$km has been treated as a known constant.

Shortcomings

An essential part of the Out-of-Africa thesis is that if x_0 is the linguistic origin, the regression coefficient of phoneme inventory on the linguistic distance vector $\|x_i - x_0\|$ must be negative. However, one piece of information (negativity) has not been used at any point in the analysis, and sign constraints have not been enforced in likelihood calculations—either by Atkinson or by me. For the rough calculations in the preceding section, the candidate points considered were restricted to existing linguistic centers in each region. In some cases, the weighted least-squares regression coefficient was positive. The fraction of negative coefficients varies considerably depending on which continental region x_0 belongs to:

Region(x_0)	Africa	Europe	Asia	Oceania	N.Amer	S.Amer
Negative fraction	1.0	1.0	1.0	0.87	0.05	0.00

Imposition of negativity constraints has no effect for candidate centers in Africa, Europe an Asia, but it must decrease the likelihood for some centers elsewhere. Given that the emphasis has been on Africa as the most plausible location, failure to impose the negativity constraint has a negligible effect on conclusions.

8.6 Matters for further consideration

8.6.1 Phoneme inventory as response

Suppose that it were possible to extract from the WALS database, the actual phoneme inventory of each language rather than the phoneme count. This statement implies a finite master list of m phonemes together with a Boolean variable $Y : [m] \to \{0, 1\}$ for each language indicating the subset of the master list that occurs in the given language. The phonemes may be labelled by type, vowel, consonant or tone, but the problem is already difficult enough without this added complication.

Without altering notation, we may regard the phoneme inventory Y_i for language i either as a Boolean vector or as a subset $Y_i \subset [m]$ of the master list. Thus Y_i is the inventory for language i, the usual component-wise product $Y_i Y_j$ is the inventory common to a pair of languages, and the k-fold product $Y_{i_1} \cdots Y_{i_k}$ is the inventory common to a specific subset of k languages.

Setting aside the complication of phoneme type, it is mathematically natural to ask for an analysis that is invariant with respect to re-labelling of phonemes

in the master list. That condition implies an analysis that depends only on phoneme inventory counts

$$i \mapsto \#Y_i, \quad (i,j) \mapsto \#(Y_iY_j), \quad (i,j,k) \mapsto \#(Y_iY_jY_k)$$

and so on. The analyses presented in this chapter use only the n-vector of first-order counts. But inventory data also provide symmetric $n \times n$ matrices of second-order counts, symmetric tensors of third-order counts, and so on.

Geography and distance are essential components in the *Out-of-Africa* hypothesis. What bearing does the hypothesis have on data collected in inventory format? Each language i is associated with a geographical location x_i; each pair may be associated with a pair of points $\{x_i, x_j\}$, with a line segment (x_i, x_j) or with a weighted centroid; each triple may be associated with a set of points, the convex hull of those points, or with their centroid, and so on. How is distance to be measured for singletons, pairs, triples and so on? How are the questions to be formulated statistically? Given answers to these questions, how might the analysis proceed to estimate relevant parameters and to check whether the data are consistent with the hypothesis?

8.6.2 Vowels, consonants and tones

If it is taken at face value, the *Out-of-Africa* hypothesis applies equally to vowels, to consonants and to tones. Is the evidence from these three sources consistent? This is something that can be checked by analyzing the three variables separately. We leave it as an exercise for the reader.

8.6.3 Granularity

In our analysis, we have ignored the fact that phoneme inventory variables are discrete, with only a few distinct values. What effect does granularity have on conclusions derived from a Gaussian model?

8.7 Exercises

8.1 Use the `table()` function to extract the distinct values for vowel diversity, consonant diversity, tone diversity and total phoneme diversity. What does this tell you?

8.2 Use the function `cov(cbind(S1[...]))` to compute the sample covariance matrix of the four phoneme inventory variables. What does this tell you?

8.3 Use the function `qr(cbind(S1[...]))$rank` to deduce that total phoneme diversity is a linear combination of the three constituents. Find the coefficient vector.

8.4 The function `vdist(x0, 1)` returns a list of linguistic distances from the designated point `x0` in linguistic region 1 to each of the 504 language locations.

Show that Atkinson's distance variable `S1$Dbo` implies that his best-fitting origin lies somewhere in the box 9–10°E, 1–2°S. Find the point and locate it on a map.

8.5 Assuming the Out-of-Africa hypothesis, total phoneme inventory necessarily depends not just on distance to the origin but also on the speaker population size. By minimizing the residual sum of squares over continental Africa, find the best-fitting origin under the linearity assumption

$$E\big(\mathrm{Tnpd}_i\big) = \beta_0 + \beta_1 \|x_i - x_0\| + \beta_2 \log(\mathrm{popn}_i).$$

You should not assume that the best-fitting origin lies in or near the box 9–10°E, 1–2°S.

8.6

Example 9

9.1 Effects of atmospheric warming

9.1.1 The experiment

This project concerns an experiment conducted at two sites in Minnesota over the period 2009–11 to determine the effects of climate warming on photosynthesis in juvenile trees of 11 different species. The following excerpt is taken from the data archive:

> To test how climate warming and variation in soil moisture supply will jointly influence photosynthesis of southern boreal forest tree species we measured gas exchange rates of 11 species in an open-air warming experiment at two sites in northern Minnesota, USA. The experiment ran for three years and used juveniles of 11 temperate and boreal tree species under ambient and seasonally warmed (+3.4°C above- and below-ground) conditions. We measured in situ light-saturated net photosynthesis (Anet) and leaf diffusive conductance (gs) on numerous days across the three growing seasons. Soil and plant temperatures and soil moisture were continuously measured from sensor arrays.

Details of the experimental design, the site preparation, the species selected, the variables measured, and so on, are provided in the paper by Reich et al (2018) `https://www.nature.com/articles/s41586-018-0582-4`.

The two sites are roughly 100 miles apart, each site consisting of 12 plots arranged in three blocks of four. Each plot is a circular area, roughly three metres in diameter, which is adequate for 30–40 juvenile specimens. Plots in the same block are sufficiently far apart that the treatment applied to one plot is judged to have negligible effect on neighbouring plots. Several specimens of each species were planted in each plot. The heat treatment was applied to plots during the growing season only. On 50 days, roughly 15–18 days per year from mid-June till late September, measurements were made on trees in several plots. For administrative reasons, all measurements on one day occurred at the same site. On these days, soil water content was recorded for each plot together with several measurements (photosynthesis, conductance, vapour pressure gradient,

temperature,...) on selected trees in each plot. Each photosynthesis and vapour-pressure measurement was made on a single leaf.

9.1.2 The data

Details concerning the variables recorded are provided at the archive

`https://portal.edirepository.org/nis/metadataviewer?packageid=edi.229.2`

For present purposes, the file

`.../borealwarming.csv`

consists of a 2049×18 spreadsheet containing the data to be used for this exercise. Each row consists of several measurements on one leaf on one day. This file differs slightly from the archived data.

The questions that follow are intended as a guide for analysis in an examination setting. You should first read he *Nature* paper and then answer the questions as asked. But you are not restricted to the points mentioned below; you are free to examine the data in any way you please, so your approach might not follow the path suggested.

1. Before any trees were planted, a grid of underground electrical cables was laid down at a depth of 15 cm in each plot, with sufficiently small separation that the heating effect could be deemed uniform. The cables in the treated plots were used as heating elements. What was the purpose of the cables in the control plots?

2. Use the data to reproduce the authors' plots in Fig. 1 and Fig. 2 in a similar format. Show your code for Fig. 1. What, if anything, do these plots tell you about the effect of elevated temperature on deciduous versus coniferous trees?

3. Soil water content is expected to vary from day to day with the most recent weather and from plot to plot depending on the topography, for example, exposure, topsoil depth, drainage capacity of the sub-soil, and so on. You are asked to examine the effect of treatment on the soil water content taking appropriate account of such variations. You may assume initially that the treatment effect is constant over sites and over years. For purposes of this analysis, you may set aside missing response values and ignore all leaf-specific variables.

   Justify your selection of terms in a suitable linear Gaussian model with soil water content as response. Explain how you fitted the model, and show the parameter estimates with standard errors.

4. You should have found a small negative treatment effect in the preceding question. Is the treatment effect constant across sites as assumed in part 3? Is it is constant across years? Explain how you might address these questions, and report your answers.

5. This question is concerned with the red-oak subset of the data, which is coded as level `queru` for the factor `species`. You are asked to analyze the relation between photosynthesis (`Asat`) and other non-leaf variables including warming treatment and soil water content. Your analysis should accommodate block, plot and temporal effects as needed. Give a brief summary of the conclusions reached on the basis of the fitted model.

9.2 Further effects of atmospheric warming

9.2.1 The second experiment

9.3 The plight of the bumble bee

Example 10

10.1 Factors affecting female fulmar fitness

The northern fulmar *Fulmarus glacialis* is a seabird found mainly off the coasts of Iceland, the British Isles and parts of Norway. Fulmars have a life expectancy of 30–40 years, and some may live up to 60 years. They have a long adolescence and commence breeding at around 7–12 years of age. Adults are monogamous, they form long-term pair bonds, and they return to the same nest site year after year. Breeding begins in May; a single egg is laid and incubated by both parents for about 50 days. The chick is brooded for about two weeks and fully fledged after about three weeks.

A survey of the Eynhallow fulmar colony in the Orkney Islands was started in 1951 by Robert Carrick and George Dunnet; the data for this exercise concern the breeding record of 428 adult female birds for the period from 1958 to 1995. They were provided by Steven Orzack in 2006. A few of the birds were active breeders when first observed, but most were observed annually from their first attempts at breeding until 1995 or the bird's presumed death. No single bird was alive for the entire period of observation. Further details concerning the Eynhallow population can be found in the 2011 paper *Static and dynamic expression of life history traits in the northern fulmar Fulmarus glaciali* by Orzack, Steiner, Tuljapurkar and Thompson (Oikos 120: 369–380).

The record for each year shows whether the bird laid an egg, and if so, whether the egg hatched, and, if hatched, whether the chick fledged, which is the event of primary interest. The data are presented in the file `.../Fulmar-1.dat` as a matrix in which each row corresponds to a bird, the first column is the bird identifier and the next 38 columns give the reproductive event observed for that bird during each of the 38 years of the study. The code for reproductive events is:

0: no reproductive event observed;

2: egg laid;

3: egg hatched;

4: chick successfully fledged (event of primary interest).

From a statistician's point of view, the code-zero description is annoyingly ambiguous. If "not observed" refers to the bird, then a reproductive event could

have occurred at Eynhallow or elsewhere without being observed or recorded. If "not observed" refers to the egg, it implies that the female was observed and did not lay an egg, i.e., "bird observed but no egg laid". The information content of these interpretations is quite different. Given what we know about the fulmar life span, it is clear that some zero codes are meant to be interpreted in the second sense. However, it appears that others are meant to be interpreted in the first sense. How can we tell which is which?

The potential for ambiguity in the coding is at least as much a matter of ornithology as it is a matter of logic and semantics. It is the ornithologist's interpretation that must prevail, so the statistician is obliged to defer—if only provisionally, reluctantly and with skepticism. Since the goal is to study reproductive success, the ornithologist is interested primarily in sexually mature birds, and female sexual maturity is deemed to be attained when the bird lays her first egg. The annual record for bird 215 for the years '58–'95 occupies one row in the file `Fulmar-1.dat` as follows:

0, 4, 2, 3, 0, 4, 4, 4, 3, 4, 4, 4, 4, 0, 2, 2, 0, 4, 4, 4, 2, 3, 2, 4, 4, 3, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0

According to S. Orzack, this sequence means that the bird reached maturity in 1959, and remained reproductively active for 26 years. Following this interpretation, the first nonzero code for each bird is deemed to occur in its first year of breeding, and the last nonzero code occurs either in the last year of life or in the last year of the study. Each intermediate zero is interpreted as "bird observed but no egg laid".

To help with subsequent computations, the observations have been rearranged into a four-column spreadsheet format in the file `Fulmar-2.dat` in which each row shows the breeding record for one adult bird in one year. The records are arranged chronologically, and the birds are coded sequentially 1–428 in column 1, which is not the same as the bird code in `Fulmar-1.dat`. The second column is the year, the third is the bird's presumed reproductive age, and the fourth is the code for the reproductive event. For each bird, years before the first and after the last nonzero event are excluded. There are 3790 such records.

10.2 Suggestions for analysis

Initially, at least, we suggest that you work with the data as coded, treating the coded value as a quantitative response. Simple questions may be addressed by computing and plotting marginal means as appropriate, and we strongly suggest that you do not attempt to fit any statistical model without first examining a range of plots or tables.

1. Eliminate leading and trailing zeros from the record for each bird in `Fulmar-1.dat`, and check that your reconstruction is in agreement with `Fulmar-2.dat`.

2. Compute the annual average reproductive score, and plot this as a time series. What is special about the year that has the maximum average

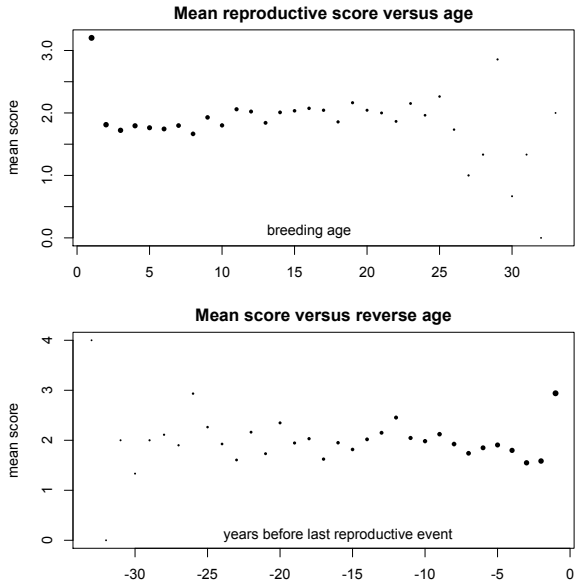

Figure 10.1: Average reproductive score versus age in the top panel, and reverse age coded backwards in time in the lower panel. Points are scaled to reflect sample size.

reproductive score? Is there an overall temporal trend? Is there any evidence of serial correlation? What formal statistical techniques can you bring to bear on these questions?

3. Use the function `tapply(...)` to compute the average reproductive score as a function of breeding age, and examine the scatterplot. This plot, the top panel of Fig. 10.1, shows that fulmars are much more successful in their first breeding year than they are in subsequent years. Suggest two explanations contributing to this unexpected anomaly.

4. Compute the average reproductive score as a function of breeding age counted backwards from last recorded breeding year, and examine the scatterplot of averages versus reverse breeding age. Comment on features that this plot has in common with averages in forward time.

5. The paper by Orzack *et al.* (2011) is concerned in part with senescence of reproduction and the biology of aging. Do age and experience of the female contribute positively or negatively to reproductive success? With this goal in mind, fit a linear Gaussian model for the scored responses (that is, 0, 2, 3, 4). Apart from reproductive age, your model should accommodate variations associated with calendar year, the bird identifier, and any anomalies that could reasonably be attributed to censoring or

sampling bias. Is the trend with age positive, negative or zero? Compute a suitable regression coefficient and its standard error.

6. One way to accommodate the anomalies seen in Fig. 10.1 is to include two additive indicator vectors, one for first breeding year and one for last breeding year. Another way is to restrict the analysis to the subset of years that are not known to be the first or last breeding year for that bird. This subset includes eight breeding birds in 1958, 55 in 1995 and 3039 others. Discuss the pros and cons of analysis by elimination of records versus analysis by inclusion of initial and terminal effects.

7. Is there any evidence to suggest that some birds are consistently more successful breeders than others? Explain.

8. It is expected that weather variations in Orkney might make some years more favourable for breeding than others. Comment on the magnitude of the year-to-year variation versus bird-to-bird variation.

9. Carry out whatever diagnostic procedures you deem appropriate to assess the appropriateness of the model you fit in part 5. Describe your results. Discuss what impact any deviations you find from your assumed model might have on your inferences.

10. Dichotomize the responses in some appropriate fashion and use logistic regression to study which factors have an effect on reproductive success. Figure 10.1 points to a bias coming from initial and final records for each bird. Discuss ways to accommodate, reduce or eliminate this source of bias. Again carry out any diagnostics you consider appropriate to assess the validity of the model.

11. Provide a one-paragraph summary of your findings that would be suitable for an ornithologist with a modest statistical background. Compare your findings with those of Orzack *et al.* (2011).

12. The information in Fig. 10.1 could have been presented using a boxplot. Why do you think the sequence of averages was chosen in preference to a sequence of boxes?

13. *The problem with the analysis in part 5 is that the observed scores are so far from normally distributed that the conclusions derived from a Gaussian model cannot be trusted.* Whether or not you agree with this sentiment, suggest a remedy, implement it, and comment on how much or how little the conclusions are altered.

10.3 Further references

Dunnett (1991) gives a concise historical account of the background, initiation and development of the study of fulmars at Eynhallow.

Chapter 11

Basic concepts

11.1 Stochastic process

11.1.1 Process

Probabilistic reasoning is the foundation of theoretical and applied statistics, and the fundamental concept that provides the basis for probabilistic reasoning is the notion of a process, and specifically a stochastic process. A process is nothing more than a function $Y : \mathcal{U} \to \mathcal{S}$ from a domain or index set $\mathcal{U}$ into another set $\mathcal{S}$ called the state space: to each point or object $u \in \mathcal{U}$, the function Y associates a point $Y(u)$ or Y_u in the state space. A stochastic process is a probabilistic description of a random function $\mathcal{U} \to \mathcal{S}$.

The domain for a Markov chain or a time series is either the integers or the natural numbers; the domain for a continuous-time temporal process is the real line $\mathbb{R}$; the domain for a spatial process may be the real plane or the complex plane, or possibly $\mathbb{R}^d$. The domain for a planar point process is not [the set of points in] the plane but the set of Borel subsets of the plane. Likewise, the domain for planar white noise is not $\mathbb{R}^2$, but the set of Borel subsets.

In a setting such as an agricultural field trial, the domain for the yield process is usually described loosely as the set of plots; this description is adequate for the field, but it is interpreted mathematically as the set of planar Borel subsets. The domain for a simple clinical trial for a COVID-19 vaccine is usually described loosely as the set of patients; this is interpreted to mean all eligible patients whether or not they were recruited and observed in the AstraZeneca trial. The domain for a study of speciation or sexual compatibility of fruit flies is the set of male-female pairs—again meaning all possible pairs having the genetic characteristics of interest. The domain for a competition experiment such as a chess or tennis tournament is the set of ordered pairs of competitors—again meaning all pairs whether or not they met in Wimbledon.

In each case, the state space is a set such as $\{0, 1\}$, $\mathbb{R}$ or $\mathbb{R}^2$, as a measurable space with Borel events. Depending on the setting, the response function may have context-specific properties, such as anti-symmetry $Y_{i,j} = 1 - Y_{j,i}$ in the case

of a pairwise competition, or additivity $Y(A \cup B) = Y(A) + Y(B)$ for disjoint plots in a field experiment.

In applied statistics, $\mathcal{U}$ is frequently called the set of observational units, and Y is called the outcome or response. A sample is a finite subset $S \subset \mathcal{U}$, and the observation is the restriction $Y[S]$ of the process to the sample.

11.1.2 Probability

A stochastic process, is nothing more than a probabilistic description of the function $Y : \mathcal{U} \to \mathcal{S}$ as a random variable or a collection of random variables $\{Y_u : u \in \mathcal{U}\}$. To each event $A \subset \mathcal{S}^{\mathcal{U}}$ the stochastic description associates a number $0 \leq P(A) \leq 1$, satisfying the rules of probability. Probability implies expectations, means, variances and so on. Given a sample $S \subset \mathcal{U}$ and an observation point $y \in \mathcal{S}^S$, the process associates a conditional probability $0 \leq P(A \mid Y[S] = y) \leq 1$. This implies conditional expectations, conditional variances and so on.

The simplest processes have independent components. In other words, to each $u \in \mathcal{U}$ there corresponds a probability distribution P_u on the state space. Independence means that for any sample $(u_1, \ldots, u_n)$ consisting of n distinct units, the joint distribution of $Y_{u_1}, \ldots, Y_{u_n}$ satisfies

$$P_{u_1,\ldots,u_n}(A_1 \times \cdots \times A_n) = P_{u_1}(A_1) \times \cdots \times P_{u_n}(A_n)$$

for arbitrary events $A_r \subset \mathcal{S}$. All generalized linear models have independent components, which are usually not identically distributed because different units may have different covariate values. By general agreement in applied work, P_u may depend *only* on covariates, so $x_u = x_{u'}$ implies $P_u = P_{u'}$.

Gaussian processes having independent and identically distributed components are the building blocks for more general processes such as those encountered in Examples 1 and 2. More general spatial and temporal processes are used throughout the examples.

11.1.3 Consistency

The dismissive phrase *nothing more than a probabilistic description of the function...*, which occurs at the beginning of the previous section, grossly underrates the difficulty of the assigned task. To understand the difficulty, consider a longitudinal design in which a given subject may be observed at an arbitrary finite collection of time points $\mathbf{t} \subset \mathbb{R}$ with $t_1 < t_2 < \cdots < t_k$. With all covariates fixed, it is necessary to specify for each $k \geq 1$ and each $\mathbf{t}$, the k-dimensional joint distribution $P_{\mathbf{t}}(\cdot)$ on $\mathcal{S}^k$. Since the event $(Y_1, Y_4) \in A \times A'$ is the same as the event $(Y_1, Y_2, Y_4) \in A \times \mathcal{S} \times A'$, these distributional specifications are subject to logical consistency conditions such as

$$\begin{aligned} P_{1,4}(A \times A') &= P_{1,2,4}(A \times \mathcal{S} \times A') \\ &= P_{1,2.5,4}(A \times \mathcal{S} \times A') = P_{1,2,3,4}(A \times \mathcal{S}^2 \times A'). \end{aligned}$$

Without consistency, alternative ways of computing the probability of a given event give different answers. Kolmogorov consistency is the mathematical glue that holds it all together, and makes statistical activities such as prediction possible.

Consistent specifications are not easy to find, and a formulation that looks plausible may well be self-contradictory. In a longitudinal setting where the response is real-valued and Gaussian, it may seem safe and natural to construct the joint distribution as a product of one-dimensional conditional distributions given past observations. This means specifying the conditional mean and the conditional variance given past observations—both times and values. If the joint distribution is to be Gaussian, the mean must be linear, and the variance constant, as a function of past values. However, the dependence on past observation times must also be specified, and this is not linear. It may be feasible to specify a continuous-time process sequentially and consistently if it is Markovian; otherwise a sequential specification is most unlikely to be consistent.

Apart from Kolmogorov consistency, other forms of consistency or inconsistency sometimes arise in statistical work. Example 5, illustrates a probability model that is incompatible with randomization.

Self-consistency is an important consideration, but not necessarily a dominant part of the story. On the one hand, a consistent specification is not necessarily well-suited to a given task. On the other hand, statistical conclusions derived from an inconsistent specification are not necessarily dangerous or disastrously wrong. It all depends on the nature of the inconsistency. Nonetheless, incompatibilities and self-contradictory specifications are strongly discouraged.

11.1.4 Statistical model

A statistical model is a non-empty set of stochastic processes $\{P_\theta : \theta \in \Theta\}$ on the same state space. It is indexed by points θ in the parameter space. Operationally speaking, to each parameter point θ there corresponds a function P_θ, and $P_\theta(A)$ is the probability of the event A in the process associated with θ. For example, $N(\mu, \sigma^2) \equiv N_{(\mu,\sigma^2)}$ denotes the normal distribution on $\mathbb{R}$, and, by extension, the process whose components are independent and identically distributed. Thus, $N_{(0,1)}(-1, 1) \simeq 0.683$ is the probability assigned to the interval $(-1, 1) \subset \mathbb{R}$, and $N_{(0,1)}((-1, 1)^{18}) \simeq 0.683^{18} \simeq 1/964$ is the probability assigned to the event in $\mathbb{R}^{18}$ that the first 18 components in an iid sequence are all less than one in absolute value.

Every distribution on a given space can be extended automatically to an iid sequence on the product space. However, this extension is not always natural or relevant. Most of the processes considered in this book do not have identically distributed components, and most do not have independent components, so the extension alluded to on the previous paragraph is not one that should be taken for granted.

11.2 Basics of Experimental design

11.2.1 Baseline

Every experiment and every observational study has a temporal component. The baseline is the temporal origin or reference point marking the beginning of the study. Mathematically speaking, the baseline is a point at which the observational units $u \in \mathcal{U}$ have been assembled, together with all of the information about them that is needed to specify the probability of arbitrary outcomes. All statistical inferences are based on probabilities, and the probability model is said to be *registered at baseline*.

Generally speaking, the units available for study are not homogeneous. The baseline information records sex, age, and, in principle anything else that is available at baseline that can reasonably be deemed to have a bearing on outcome probabilities. In practice, a certain restraint or professional judgement is needed to decide what is likely to be relevant and what is not. In a field experiment, the geometric layout of the plots is ordinarily part of the registered baseline information, and is almost always relevant in that it affects outcome probabilities. Information about crop, treatment and yield in the previous season is sometimes available and might be judged relevant if the new plots were well-aligned with the previous plots. In a clinical trial with human patients, ethnic background might be relevant as a block factor, but the number of letters in the patient's name is unlikely to be considered relevant for clinical outcomes.

For a randomized study, randomization occurs at or immediately after baseline. The randomization protocol is registered at baseline, but the randomization outcome is not. Model specification begins with randomization probabilities $p(\mathbf{t}) = \mathrm{pr}(T = \mathbf{t})$ for each treatment assignment vector $\mathbf{t} = (t_i)_{i \in S}$, also called the treatment factor. Even if one assignment list is a permutation of the other, two assignment vectors $\mathbf{t}, \mathbf{t}'$ may have, and usually do have, different probabilities depending on baseline information such as covariate or block structure. Most commonly, the randomization is balanced with each treatment level occurring with equal frequency in each block.

Since the probability model is registered at baseline, i.e., pre-randomization, the model specifies the joint distribution for treatment T and response Y. The joint distribution implies a marginal distribution for treatment assignments, and a conditional distribution $\mathbf{t} \mapsto F(\cdot \mid T)$, which associates with each assignment vector $\mathbf{t}$ a conditional distribution for the response. Randomization subsequently produces a particular treatment configuration, and nearly every subsequent probability computation uses that value. In general, the conditional probability $F(A \mid T = \mathbf{t})$ of the event $Y \in A$ may depend on any and all registered baseline information. Every variable measured post-baseline, such as T, is regarded as the outcome of a random process, and, as such, is formally a part of the response.

Baseline need not mean a fixed point in calendar time. In studies of cell development, the baseline would ordinarily be set at a key developmental stage such as fertilization, which is a point in calendar time that may vary from cell

to cell. Similar remarks apply to clinical trials where the baseline is usually set at recruitment, which varies from one patient to another on the calendar scale.

11.2.2 Observational unit

The *observational units* are the objects $u \in \mathcal{U}$ on which variables are defined and measurements may be made. Usually measurements are made only on a small subset of observational units (the sample), so the phrase *measurements may be made* does not imply that measurements have been made or that plans are afoot to make such measurements.

The statistical universe almost always includes infinitely many extra-sample units, notional or otherwise, for which probabilistic prediction may be required. Sometimes each unit is a physical object such as a plot, a patient, a rat, a tree, or a M-F pair of fruit flies. Sometimes the units are less tangible, such as time points or time intervals for an economic series, or spatio-temporal points or intervals for a meteorological variable such as temperature or rainfall. Very often, the set of observational units is a Cartesian product set such as

$$\{\text{mice}\} \times \{\text{front, rear}\} \times \{\text{left, right}\} \times \{\text{day0, day1, day2}\}$$

which contains 12 observational units for each mouse. As an index set, time is structured cyclically in a similar way:

$$\{\textit{clock times}\} \times \{\textit{?? days}\} \quad \text{or} \quad \{\textit{365 calendar dates}\} \times \{\textit{?? years}\}$$

The index set may be stuctured in other ways such as pupils within classrooms within schools, which is a nested or hierachical structure defined by one or more relationships $R(u, u')$ on the units.

11.2.3 Population

The *population* $\mathcal{U}$ is the set of observational units, which is typically infinite; the *sample* is the finite subset of observational units that occurs in the study. Where necessary, the sample may be extended to include units for which observations are unavailable but response predictions are requested. In a meteorological context, the observational units are all points in the plane or sphere, or points in the spatio-temporal product space, so the population is uncountably infinite. For a spatial process, the units may be either points in the plane, or subsets of the plane, or less tangible objects such as signed measures on the plane or planar contrasts. The sample is the finite set of points at which measurements (sample values) are planned or available or desired.

The mathematical population is the *index set* on which the response (yield, health, weather,...) is defined as a stochastic process. As is often the case in mathematics, the mathematical index set is made sufficiently large that it encompasses every conceivable situation that might arise, and many more besides. For a clinical trial in which the experimental units are human patients,

the mathematical index set need not be finite, and in fact the mathematical subset of units having a specific sex, age and body-mass-index may also be infinite. A non-mathematician might object to the fact that the mathematical index set contains more points than there are real physical or biological entities, or atoms in the universe. Such objections are not to be entertained seriously; they are on a par with rejecting the real number system for engineering or accounting purposes on the grounds that it contains infinitely many 'useless' values that are not needed for billing purposes.

A non-trivial stochastic theory requires the sample to be a proper subset of the population, but it does not require $\mathcal{U}$ to be infinite. There are *bona fide* applications that call for a finite population, so we do not insist that all populations be infinite. However, we shall not encounter such applications in these notes.

Statistical colloquialisms. When one talks of a 'Normal population' or a 'Cauchy sample', the reference is not to the population or sample *per se*, but to the population values or sample values or their distribution, usually understood to have independent values for distinct units.

11.2.4 Biological populations

Every biological population evolves by a process of birth and death. Tomorrow's population is not the same as today's population or yesterday's population, but all three are finite. Mathematically speaking, the population is said to be locally finite in time. It is immaterial whether the entire population is globally finite or globally infinite. What is important is that only the current population is accessible or available for sample inclusion.

For some short-term social policy matters, voting and other political activities, the relevant population for inference is determined by democratic principles. Only the current population has a vote, so past and future generations are not counted in the population. Such populations are finite.

For medical and pharmaceutical studies, it is preferable to take a broader view, particularly if there are any plans to use the drug or therapy for future patients. However, this broader perspective means that not all individuals in the population are accessible and available for inclusion in a sample today.

In a clinical trial for a Covid-19 vaccine, the units available for recruitment are individuals who are alive and of a suitable age at the crucial time. It appears that the Covid-relevant population is finite. However, there are at least two reasons to reject the finiteness argument. The first is that the current population is very large. It is difficult to put a precise figure on on it, say 7.5–8.0 billion, and it is even more difficult to explain why this number is biologically or mathematically relevant for the assessment of drug safety or efficacy. The second argument is that the Covid-19 relevant population is not restricted to the present, but also includes at least one future generation. Given that some units are inaccessible, it is sufficient to take $\mathcal{U}$ to be infinite, so that the mathematical set is large enough to accommodate every conceivable demand, even beyond what is epidemiologically plausible.

11.2.5 Samples and sub-samples

The *sample* $S \subset \mathcal{U}$ is the finite subset of observational units on which the response and other variables is recorded. Technically, S is a finite ordered list of units, usually but not necessarily distinct, and the recorded response $Y[S]$ is the list of Y-values for $u \in S$ in the same order.

To be clear, the word 'sample' in these notes denotes a finite ordered subset of units. It does not imply a random sample, let alone a simple random sample. Two samples consisting of the same units taken in a different order are different, but statistically equivalent for most purposes.

In settings where prediction or interpolation is involved, it is necessary to consider an extended sample S', which includes S as a sub-sample. Each $u \in S' \setminus S$ is called an extra-sample unit. Only the restriction $Y[S]$ is actually observed. Prediction refers to the conditional distribution of $Y[S']$ given $Y[S]$; point prediction refers to the conditional expected value.

11.2.6 Illustrations

In the discussion of Example 1, it was asserted that each observational unit is a site on a rat, i.e., a (rat, site) pair, and the response is a real number, i.e., the state space is the real numbers. However, one could argue that each rat is one observational unit, and the state space is $\mathbb{R}^5$. At first glance, these appear to be equivalent.

What makes one choice more appropriate than the other is the nature of the five observations on each rat. If these were five otherwise unrelated variables such as pulse rate, temperature, weight and blood pressure, each rat would be one observational unit, and the state space would be $\mathbb{R}^5$. However, the observation consists of one biological variable measured at five sites. Although we do not necessarily expect the five measurements on one rat to be exchangeable or even to have the same expectation, the nature of the observation process—using the same instrument for each site—confers additional symmetry.

For one rat, either choice leads to a response distribution on $\mathbb{R}^5$. The difference is that the second version has more natural symmetries than the first. These symmetries arise from notionally permuting the units in various ways. For example, the model used in Example 1 has equal variances for all sites, and equal covariances for each pair of sites, which comes implicitly from assumptions about permuting sites. If we choose the rat as the observational unit, there is no possibility to permute sites, so these symmetries do not emerge as a consequence of permutation of units.

In Example 3, each observational unit was taken initially to be a mating event. But this was subsequently shown to be inappropriate for the design, and misleading for the analysis. Instead, it was deemed preferable to take one mating well as the observational unit.

For the daily temperature series, each observational unit for the analysis in Chapter 6 is a point in calendar time, consisting of a year and a date within the year. Date is a number in the range 1–365 having cyclic structure, i.e., a real

number with addition modulo 365.

For the frequency analysis in Chapter 7, each observational unit is a Fourier frequency. These also come with harmonic structure such that frequency ω is associated with its harmonics $\{\omega, 2\omega, \ldots\}$.

In Examples 1–7, one observational units is (i) a site on a rat; (ii) a (log, saw) or (log, team) pair; (iii) a mating well; (iv) a (plant, date) pair; (v) a louse; (vi) a point in calendar time; (vii) a frequency; (viii) a language. The population is some set of observational units, and there is usually little reason to restrict the mathematical population to a finite set. Many of these are all relatively straightforward from the definition given, but it is clear in several instances that other choices are possible.

11.3 Response and other variables

11.3.1 Variable

A *variable* is a function on the observational units, both sample units and extra-sample units. Everyday examples include 'weight in kg.', 'atmospheric pressure in cm. Hg.', and 'length in cubits'. In principle, the variable name includes the physical units of measurement so that the value $x_i \equiv x(i)$ of the variable x for unit i is a number, not an expression such as '184.5 cm'. Mathematically speaking, weight in kg. and weight in lbs are different variables; in practice, descriptive terms such as weight, height and temperature are used flexibly in everyday speech without specified units. Flexibility is good, but ambiguity can be costly—such as the loss by NASA in 1999 of a Mars orbiter at a cost of $125M because of a mix-up of distance units by Lockheed-Martin.

Qualitative variables include *sex* taking values in $\{M, F\}$; or *occupation* taking values in a suitable set of occupations. This set of values or levels must be exhaustive, so one of the values may be *'none of the above'*.

Operations: If u, v are two variables, the ordered pair (u, v) is also a variable: the value of (u, v) for unit i is $(u, v)(i) = (u_i, v_i)$, which is a point in the Cartesian product space. Each variable is defined on the population and recorded on the sample.

Feature is a synonym for variable or attribute—a function on the units. The feature vector takes values in the feature space.

In certain settings, the response on one unit is a vector, and each feature is one component; the primary response is a class or characteristic of the unit, and the goal is to classify each unit by computing the conditional distribution over the set of classes given the features.

11.3.2 Quantitative variable

A real-valued function on the observational units is called a *quantitative* variable. More generally, a quantitative variable is a function taking values in a vector space. Dose (of fertilizer or medication in suitable units) is a typical quantitative

variable whose values are non-negative. Blood pressure (systolic, diastolic) in mm. Hg. is a quantitative variable taking values in $\mathbb{R}^2$. This statement means that every realizable value of blood pressure can be found somewhere in $\mathbb{R}^2$; it does not mean that every point in $\mathbb{R}^2$ is realizable as a blood-pressure value for a live human subject. Values that are in conflict with hydrostatic and hydraulic theories are deliberately not excluded by the definition.

Operations: If x, z are two quantitative variables taking values in the same vector space, so also is the linear combination $3x + 4z$. If x, z are real-valued variables, so also is the unit-wise products xz. Consequently x^2, z^3 and other monomials such as x^2z are also quantitative variables.

11.3.3 Qualitative variable

A *qualitative* variable, also called a *classification factor*, is a function on the observational units taking values in a finite or countable set, called the *factor levels*. Examples include *sex*, *occupation*, *socioeconomic class*, and variables such as *genetic variant* with values 'wild type' and 'mutant'. Often, one level is designated as a reference level. A qualitative variable is sometimes called an *attribute* or a feature.

Ordered pairs: If u is the qualitative variable representing COVID vaccine with four levels *Pfizer, Moderna, AstraZeneca, Janssen*, and v is the dose count with three integer values $\{0, 1, 2\}$, the product set contains twelve ordered pairs that are mathematically distinct. Operationally, however, the four pairs associated with zero dose are not distinguishable, so the number of distinguishable ordered pairs is only nine.

11.3.4 External variable

Any variable measured post-baseline is regarded as a random variable whose probability distribution is specified at baseline. The randomization outcome becomes available post-baseline, so it is a component of the response in the sense that its distribution is specified at baseline. Usually, the randomization outcome is not of scientific interest in itself, so the focus of the investigation lies elsewhere.

Apart from the randomization, there may be other post-baseline variables that are relevant and must be considered, but are not themselves of scientific interest. An external or endogenous variable is one that is usually not independent of the primary response, but whose temporal evolution is independent of the primary response. For a definition of *independent evolution*, see section 11.2.10. Independent evolution is an asymmetric relation between two temporal processes, so this concept arises primarily in longitudinal designs or in time series analysis. Louse sex in Example 5 is a simple example of a post-baseline variable that is external in the sense of the definition.

11.3.5 Response

The *response*, usually denoted by Y, is the variable of primary interest, the variable that is measured or recorded on the sample units, e.g., yield in kg. per unit area, or time to failure in a reliability study, or stage of disease, or severity of pain, or death in a 5-year period following surgery. There may be secondary or intermediate response variables such as compliance with protocol in a pharmaceutical trial, which are also part of the response. Synonyms and euphemisms include *yield*, *outcome* and *end point*.

In statistical work, the response is regarded as the realized value of a random variable, or process $u \mapsto Y_u$ taking values in the *state space* $Y_u \in \mathcal{S}$. For an observational study, the distribution is denoted by F; for a randomized study $F(\cdot, \mathbf{t})$ is the joint distribution of the response and treatment assignment.

To be clear, the response is not some conceptualized or notional variable that we would like to measure but are unable to measure on the sample units. By definition, the response is the variable that is actually measured on the sampled units, i.e., the value recorded by a blood pressure instrument or a treadmill task at a particular time, or by a questionnaire for a psychiatric evaluation, not some notional 'true' state of health. Likewise, the probability model is a probability distribution for the process corresponding to the variable measured, including the procedure or instrument used to measure it.

Many of the stochastic models considered in these notes are built from simpler processes, for example, by addition of a smooth process plus white noise, or by using a latent smooth process as the intensity for a Bernoulli process or Poisson process. Some authors are then inclined to refer to the unobserved smooth process as the 'true value', suggesting that the observation is the false or corrupted value. Provided that the descriptive term 'true value' is understood in the non-pejorative pure-mathematical sense, this terminology causes no difficulty. But it can lead to linguistic awkwardness in instances where the true state of health is normal even after the patient has died.

11.3.6 Covariate

A *covariate* x is a baseline function on the observational units that is used in a probability model to permit the outcome distribution for one unit to differ from that of another unit. Ordinarily, if $x_i = x_j$, the events $Y_i \in A$ and $Y_j \in A$ are presumed to have the same probability; otherwise, if $x_i \neq x_j$, the probabilities may be different. For this to make operational sense, the covariate must be registered at baseline. Typical examples include patient age, sex of mouse, type of soil or soil pH (pre-planting).

If the set of observational units is a Cartesian product set $\mathcal{U} = \mathcal{U}_0 \times \mathcal{U}_1$, each marginal component $u \mapsto u_0$ or $u \mapsto u_1$ is a baseline variable. In Example 1, each unit u is a $(rat, site)$ pair, so the function $u \mapsto site(u)$ is a covariate. The function $u \mapsto rat(u)$ is also a baseline variable, but it is used as a block factor. In Example 5, each observational unit (louse) is associated with an ordered pair $(aviary_u, time_u)$, so aviary and time are baseline variables.

Operationally, a covariate is used in a randomized experiment to reduce 'unexplained' variation and thereby to increase the precision of treatment effect estimates. In an analysis of variance, the total sum of squares for the response is partitioned into various parts, one part associated with registered covariates and block factors, a second part associated with treatment, the remainder being 'unexplained' or residual variation. The part associated with covariates and block factors, the between-blocks variation, is said to be 'eliminated', and the more variation that can be eliminated, the less there is to contaminate the estimates of treatment contrasts. A covariate or block factor is said to be effective for this purpose if the associated mean square is substantially larger than the mean squared residual. This means that the response variation within blocks, the intra-block mean square, should be appreciably smaller than the response variation between blocks, the inter-block mean square.

In practice, it may be acceptable to fudge matters by using as a covariate, a variable measured post-baseline before the effect of treatment has had time to develop, or an external variable whose temporal evolution is known to be independent of treatment assignment for the system under study. Louse sex in Example 5 is a simple, uncontroversial, example of a post-baseline variable, which is not statistically independent of the response (louse size), but whose *evolution* is 'known to be' independent of both treatment assignment and louse size.

At a minimum, it is necessary first to check that the variable in question is indeed unrelated to treatment assignment; otherwise its use as a covariate could be counterproductive. It is well to remember that while measurement pre-baseline is strong positive evidence that no statistical dependence on treatment assignment exists, the most that can be expected of a post-baseline measurement is absence of evidence. For a variable of dubious status, absence of evidence is considerably better than its complement, but it does not provide the same positive assurance as evidence of absence. A concomitant variable of this sort is not counted as a covariate in these notes. It is formally regarded as a component of the response whose dependence on treatment assignment is to be specified as a part of the statistical model. The dependence may be null, but that alone does not give it the status of a covariate.

As always, a probability model F allows us to compute whatever conditional distribution is needed for inferential purposes. That includes the conditional distribution given any concomitant or intermediate outcome or the conditional distribution of health values given that the patient is alive, or the conditional distribution of the cholesterol level given that the patient has complied with the protocol, or even the probability of compliance given the cholesterol level. Whether these are the relevant distributions for the purpose at hand is an entirely different matter to be determined by the user.

11.3.7 Treatment

Treament is a function T: *sample units* $\rightarrow$ *levels* taking values in the set of treatment levels. Treatment is not a covariate because it is not a property of

the observational units that is registered at baseline; it is an *intervention* that changes the status quo for the sampled units only. Usually, treatment is a random variable whose value is the outcome of a *randomization scheme*. The components of T for distinct observational units, or even for distinct experimental units, are usually identically distributed, but seldom independent.

In computational work, the observed treatment configuration $(T_u)_{u\in S}$ is called the treatment factor. Although T is defined only for sample units, we must bear in mind that the sample can always be extended indefinitely, at least in principle, so the restriction to S is not a major part of the distinction between a classification factor and a treatment factor. The important distinction is that a pre-baseline variable is a property of the units, whereas treatment level is assigned to units at baseline.

11.3.8 Relationship

A *relationship* is a function on *pairs of units* that may be used in the statistical model to distinguish the joint outcome distribution for one pair of units versus another pair. For this to be feasible, the values must be registered at baseline. If each unit is a point in a metric space, or is associated with such a point, the metric $d(u, u')$ is a non-negative symmetric relationship among them. Experimental units are defined by a Boolean relationship: $E(u, u') = 1$ if u, u' belong to the same experimental unit. Other examples include genetic, familial, neighbour, and adjacency relationships.

Ordinarily, the relationship is defined on the population and recorded for the sample pre-baseline. In Example 5, however, *Aviary* is a block factor generated by randomization, and defined on the sample. Since the randomization may have been accomplished in two waves that were not necessarily synchronous, it is difficult to say whether this block factor is pre-baseline or post-baseline.

11.3.9 Block factor

A *block factor* is a Boolean function on pairs of observational units that is reflexive, symmetric and transitive—an equivalence relation registered at baseline. Each block factor (such as the experimental unit factor) partitions the set of observational units into disjoint non-empty subsets called blocks. The identity function on $\mathcal{U}$ is a block factor whose blocks are all singletons; at the other extreme, the function J such that $J_{u,u'} = 1$ for every pair, has exactly one block.

To each variable or factor x there corresponds a block factor B defined by

$$B_{ij} = 1 \text{ if and only if } x(i) = x(j).$$

Regardless of how the information is stored in an electronic device, the chief mathematical difference between B and x is that the x-blocks are labelled by x-levels, whereas the blocks of B are unlabelled. The x-block $x^{-1}(x(1)) = \{j \in S\colon x(j) = x(1)\}$, i.e., the subset of sample units having the same x-value as

unit 1, has the label $x(1)$. Since the blocks of B are unlabelled, a block factor has no reference level or reference block.

At the risk of over-simplification, covariates typically occur in the model for the mean response; block factors and other relationships occur in the model for covariances.

11.3.10 Independent evolution

Let the response be a two-component temporal process, so that (Z_t, Y_t) is the value at time t. For notational simplicity, time is discrete. Independent evolution is an asymmetric relation between the two processes. We say that Z *evolves independently of* Y if, for each t, future Z-values are conditionally independent of past Y-values given past Z-values. In particular, for every t,

$$Z_t \perp\!\!\!\perp Y^{(t-1)} \mid Z^{(t-1)} \tag{11.1}$$

where $Y^{(t)} = Y[\ldots, t]$ is the restriction to past values.

Independent evolution does not imply that the two processes are statistically independent, nor does it imply that Y evolves independently of Z. It is an asymmetric relationship between temporal processes, which simplifies the sequential factorization of the joint density

$$\begin{aligned} p(z_t, y_t \mid Z^{(t-1)}, Y^{(t-1)}) &= p(z_t \mid Z^{(t-1)}, Y^{(t-1)})\, p(y_t \mid Z^{(t)}, Y^{(t-1)}) \\ &= p(z_t \mid Z^{(t-1)}) \times p(y_t \mid Z^{(t)}, Y^{(t-1)}). \end{aligned}$$

When the focus is on Y as the primary response, an auxiliary process satisfying (11.1) is sometimes called *external* or *exogenous*.

In circumstances where Z is exogenous, the evolution of Y may be governed by synchronous Z-values only, in which case we have

$$Y_t \perp\!\!\!\perp Z^{(t-1)} \mid Z_t, Y^{(t-1)} \tag{11.2}$$

in addition to (11.1). The joint density then factors as

$$p(z^{(T)}) \times \prod_{t=1}^{T} p(y_t \mid Z_t, Y^{(t-1)}),$$

where the focus is usually on the second factor.

The much stronger conditional independence condition

$$Y_t \perp\!\!\!\perp Z^{(t-1)}, Y^{(t-1)} \mid Z_t \tag{11.3}$$

severely limits the nature of the temporal dependence in Y. In this case the second factor in the joint density simplifies further to

$$p(y \mid Z) = \prod_t p(y_t \mid Z_t).$$

The occurrence of Z is similar to the occurrence of a covariate, as if Z were recorded at baseline.

Example 5 shows that pigeon lice are sexually dimorphic, so size and sex are strongly dependent. Nonetheless, louse sex is a good example of a post-baseline variable that evolves according to Mendelian laws, independently of the main response (louse size). It is also clear on general grounds that only synchronous sex-values matter, so (11.2) is satisfied. However, Brownian evolution processes do not satisfy the stronger condition (11.3), so the simpler density factorization fails.

The health of an asthmatic patient may depend on recent local weather, but the evolution of weather patterns is, to an adequate approximation, independent of the health of patients. It is obvious in this setting that only local weather patterns matter, and recent is more important than not-so-recent, but it is less obvious that only synchronous weather matters, so (11.2) is dubious. Certainly, one would not expect (11.3) to hold for values measured at moderate to high frequency. Similar remarks could be made regarding investors in the stock market.

11.4 Comparative studies

11.4.1 Randomization

The *randomization scheme* is a probabilistic protocol for the assignment of treatment levels to sample units, often uniformly at random subject to design constraints. For a completely randomized design with 12 sample units and four treatment levels, a balanced randomization scheme is a function $T\colon [12] \to [4]$ (from sample units to treatment levels) chosen [uniformly] at random from the set of $12!/(3!^4) = 369600$ functions having treatment blocks $T^{-1}(1), \ldots, T^{-1}(4)$ of equal size. In the randomized blocks setting, each sample unit is an experimental unit.

Usually, the randomization probabilities depend on the block structure and covariate configuration occurring in the sample units. For a typical randomized blocks design, the joint probability that the pair (u, u') is assigned treatment levels (t, t') depends on whether the units belong to the same block or to different blocks. More generally, the probability $\mathrm{pr}(u \mapsto t; S)$ that treatment level t is assigned to unit u may depend not only on x_u but also on $x_{u'}$ for all other units $u' \in S$. Unless otherwise specified, we assume in these notes that the assignment probabilities $\mathrm{pr}(u \mapsto t; S) > 0$ are strictly positive for every unit and every treatment level. For an exception in which the menu of treatment options may be covariate-restricted, see Example 2.6.

In cases where the components of $\mathbf{t}$ are independent, the randomization probability $\mathrm{pr}(T_u = t_u)$ may depend on baseline covariates or classification variables such as sex. For example, a two-level treatment may be assigned in the ratio 1:2 for males and 2:1 for females. Ordinarily, a deliberately unbalanced design of this sort causes no problems in the analysis, except perhaps for a reduction

in efficiency. But there is one important exception to this rule. Randomization probabilities are invariably assumed to be independent of initial values; see section 11.4.5.

11.4.2 Experimental unit

The *experimental units* are the objects to which treatment is assigned, i.e., two distinct experimental units may be assigned different treatment levels. Or, to say the same thing in a different way, two distinct experimental units are assigned different treatment levels with strictly positive probability. Each experimental unit consists of one or more observational units, e.g., one mouse consisting of four legs, or one classroom consisting of 20–40 students in the preceding example.

Two observational units u, u' belong to the same experimental unit if the randomization scheme necessarily assigns them to the same treatment level. In mathematical terms, $R(u, u') = 1$ if and only if $T(u) = T(u')$ with probability one. By construction, R is an equivalence relation, which partitions the sample units into disjoint blocks. Each block of R is one experimental unit.

A/B testing: This phrase, which originates in commercial internet activity, refers to a treatment having two levels A, B, which may be connected with options for on-screen presentation of internet search results. Each search is an observational unit, the response being click/no click. The experimental units may be searches or users or IP addresses, depending on the circumstances.

11.4.3 Covariate and treatment effects

In standard probability language, the phrase 'X is independent of Y' is not a statement about the random variables as measurable functions or the pair of outcomes (X_u, Y_u) as numerical values for a particular unit, as it is a statement about probabilities: the joint probability for each product event $(X, Y) \in A \times B$ is multiplicative. Likewise, when we talk of a statistical effect in a context such as 'the effect of treatment on longevity' or 'the effect of variety on yield', the effect referred to is not a numerical difference of two survival times or two yields, but a difference of two probabilities or a difference between two probability distributions.

For example, if the probability model asserts that the yield in kg/Ha on plot u is distributed as $N(\mu, \sigma^2)$ for variety I and $N(\mu, 2\sigma^2)$ for variety II, the effect of variety (II versus I) is implicitly to double the yield variance. The effect of variety on [the probability of] a particular event $Y_u \in A$ is the difference $N(A; \mu, 2\sigma^2) - N(A; \mu, \sigma^2)$ between two conditional probabilities, which depends on both parameters. Similar remarks apply to the effect on linear and non-linear functionals such as means, medians or quartiles of the yield distribution.

Apart from treatment effects, there are other effects of a different nature, such as the difference in survival distributions for males versus females, or the effect of aging on mobility or cognitive function. These are covariate effects.

Every treatment effect in these notes is modelled as a group action on probability distributions, which is not necessarily the case for covariate effects.

The effect of a 10-year age gap on the probability of event A is the difference between two probabilities $P_u(A)$ and $P_{u'}(A)$ for two units such that $age(u') = age(u) + 10$; this covariate difference implies $u \neq u'$. The effect of treatment on the probability of A is the difference between conditional probabilities $P_u(A \mid T = 1)$ and $P_{u'}(A \mid T = 0)$ for two units having the same covariate value. Although no unit receives more than one treatment, this difference is defined and can be evaluated for $u = u'$. However, $x(u) = x(u')$ plus exchangeability implies

$$P_u(A \mid T = 0) = P_{u'}(A \mid T = 0) \quad \text{and} \quad P_u(A \mid T = 1) = P_{u'}(A \mid T = 1),$$

so the treatment effect is the same for every pair such that $x(u) = x(u')$, whether $u = u'$ or not.

11.4.4 Additivity

Additivity refers to additivity of effects associated with classification factors, block factors and treatment factors. For a two-factor model with factors A, B, the mean response is additive if

$$E(Y_u) = \alpha_{A(u)} + \beta_{B(u)}.$$

For non-Gaussian responses, and even for Gaussian models, it may be necessary first to apply a transformation to achieve additivity. For example, if $Y_u \sim \text{Ber}(\pi_u)$ is a Bernoulli variable, the logistic model

$$\text{logit}\, E(Y_u) = \alpha_{A(u)} + \beta_{B(u)}$$

exhibits additivity on the logistic scale.

Additivity usually refers to the mean model, but it can can also refer to random-effects models. For example if A is a treatment factor and B is a block factor, the Gaussian model

$$Y \sim N_n(\alpha_{A(u)},\ \sigma_0^2 I_n + \sigma_1^2 B)$$

exhibits additive treatment and block effects. If the effects are not additive, i.e., if the treatment effect for one level of B is different from the treatment effect at some other level, we say that interaction is present. Interaction and non-additivity are effectively synonymous terms; synergy is also used for non-additivity, particularly if the treatment effect is boosted by an increase in the level of the second variable.

11.4.5 Initial values

Every variable that is recorded at baseline is available for use as a covariate that modifies the distribution of the process being studied. If the response is a

vital variable such as blood pressure in a study of hypertension, the value varies over time and the initial value is invariably recorded at baseline. The present discussion makes a distinction between the initial value Y_{i0} of the process being studied, and other baseline variables such as sex, age, weight, marital status, and so on, whether these are constant in time or not. Treatment assignment is randomized according to a declared protocol, and the goal is to study its effect on blood pressure after six months.

Assume for simplicity that observations are made at exactly two points in time, the same time points $t = 0$ and $t = 1$ for every subject. The simplest Gaussian framework, which assumes independent responses for distinct patients, admits two slightly different ways of handling initial values. The simplest way is to regard the initial value as a covariate on an equal footing with all others. In the absence of interaction, the conditional expected value given treatment is assumed to be linear:

$$E(Y_1 \mid \mathbf{x}, \mathbf{t}, Y_0) = X\beta + Y_0\rho + \mathbf{t}\tau. \tag{11.4}$$

Ordinary least squares is used to estimate the treatment effect τ on the assumption of conditional independence and constant conditional variance. If we take the response to be the *change* in blood pressure, the expression becomes

$$E(Y_1 - Y_0 \mid \mathbf{x}, \mathbf{t}, Y_0) = X\beta + Y_0(\rho - 1) + \mathbf{t}\tau. \tag{11.5}$$

Whether we opt to work with the final value or the difference as response, both versions yield the same point estimate and standard error for the treatment effect. Ordinarily, we should expect the correlation ρ to be large and positive, say $\rho \simeq 0.75$, so the coefficient in the second version should be small and negative. Leaving aside the possibility for interactions such as unequal effects for males and females, this is a reasonably accurate description of recommended practice (Senn, ????).

The second method, which is illustrated for a more complex setting in section 5.2.3, is to regard each response pair (Y_{i0}, Y_{i1}) as bivariate Gaussian with variances σ_0^2, σ_1^2, correlation ρ, and to consider the natural linear model for a bivariate response. Under the assumption made at the end of section 11.4.1, randomization implies no treatment effect at baseline, so the two mean vectors are

$$E(Y_0 \mid \mathbf{x}, \mathbf{t}) = X\beta_0; \quad E(Y_1 \mid \mathbf{x}, \mathbf{t}) = X\beta_1 + \mathbf{t}\tau, \tag{11.6}$$

where β_0 is the regression coefficient for variables measured synchronously, and β_1 is the coefficient for asynchronous variables. The second part implies

$$E(Y_{i1} \mid \mathbf{x}, \mathbf{t}, Y_0) = \mathbf{x}_i'\beta_1 + \tau t_i + \gamma(Y_{i0} - \mathbf{x}_i'\beta_0) = \mathbf{x}_i'(\beta_1 - \gamma\beta_0) + \gamma Y_{i0} + \tau t_i,$$

with $\gamma = \rho\sigma_1/\sigma_0$. Provided that the constraint $-1 \leq \rho \leq 1$ is overlooked, this is equivalent to (11.4) with $\beta = \beta_1 - \gamma\beta_0$ and γ in place of ρ.

The terminology and notation leading to (11.4) and (11.5) imply $\sigma_1 = \sigma_2$, and this assumption (local stationarity) distinguishes the initial value from other baseline variables. If equality of variances is assumed, the covariance matrix in

(11.6) is a linear combination of the identity and the block matrix for pairs. In that case, the REML estimate of τ is not the same as the ordinary least squares estimate in (11.4). If equality of variances is not assumed, the covariance matrix in (11.6) is a linear combination of two diagonal matrices and one block matrix for pairs. For that case, the point estimate of τ is the same as the ordinary least squares estimate in (11.4), but the standard error as computed by REML is not the same as the least-squares standard error.

The following numerical example illustrates the magnitude of the differences in a sample of 12 patients:

y_0	4.06	7.63	5.89	4.33	7.74	5.35	2.75	5.60	4.76	6.76	6.87	5.13
y_1	2.19	5.84	4.86	1.45	7.05	5.85	2.54	6.53	4.68	5.56	5.79	3.29
$\mathbf{t}$	0	0	0	0	0	0	1	1	1	1	1	1

The three Gaussian models described above produce the following estimates for the treatment effect. In each case, the standard REML procedure was used for the estimation of variances and covariances.

	$\hat{\tau}$	s.e.($\hat{\tau}$)	s^2
(11.4): OLS	0.7259	0.6671	1.2912
(11.6): $\sigma_0 = \sigma_1$	0.6084	0.5885	1.1621
(11.6): $\sigma_0 \neq \sigma_1$	0.7259	0.6224	1.0391

The final column is the estimate of the conditional variance $\text{var}(Y_1 \mid Y_0)$, either computed directly from the residual mean square, or computed indirectly as a function of the fitted 2×2 covariance matrix.

Even if the response process is stationary, so that $\sigma_0 = \sigma_1$ is satisfied, the initial value is commonly used to determine patient eligibility. In such cases, the response process for eligible patients is not expected to be stationary; the regression phenomenon implies that the mean is not constant, and the variance may be expected to increase over time. For this setting, ordinary least squares based on (11.4) is fully efficient.

Apart from efficiency, computational convenience is the real reason for preferring (11.4) over the bivariate model (11.6). For a longitudinal study where each patient is measured at several points post-baseline, the principled approach of modelling the entire temporal process for each subject is greatly preferred. If it is needed, the conditional distribution given baseline values can be derived from the covariance function of the process.

In general, the only condition on treatment assignment probabilities is that they be fully specified as part of the protocol. Thus, treatment assignment probabilities may depend on any and all baseline covariates, including block factors. While it might not be efficient to do so, it is entirely legitimate for treatment levels to be assigned in the ratio 1:2 for males and 2:1 for females, However, (11.6) implies explicitly that Y_0 is independent of $\mathbf{t}$, at least componentwise given $\mathbf{x}$, and this assumption is implicit in (11.4). In this respect, the initial value is not treated on an equal footing with other baseline variables.

11.4.6 Design

The word *design* refers to the arrangement of the sample units by blocks, by covariates, and by restrictions on treatment assignment. Very often, it is helpful to distinguish between two aspects of the design, the *structure of the units*, meaning relationships among them, and the *treatment structure*, which is imposed on them. In a crossover design, where the same physical object occurs as a distinct experimental unit on several successive occasions, the structure of the units includes not only the temporal sequence, but also a block factor whose blocks are the distinct physical objects. In a field experiment, the structure of the units includes the geometric shape of each plot, their physical arrangement in space, and the width of access paths or guard strips separating neighbouring plots.

11.4.7 Replication

Replication means repeating the experiment independently for different experimental units under essentially identical circumstances in order to gauge the variation in response distribution. Independence is crucial. In an animal-behaviour study, it is easy to partition a one-hour observation interval into six consecutive ten-minute intervals, and to report behaviour counts for each interval. The number of animals, or pairs of animals, is unchanged, but the number of observations is immediately increased by a factor of six. Although the experimental settings may stay the same for each sub-interval, these values, sometimes called pseudo-replicates, are not independent. For a good example of an incorrect analysis for pseudo-replicates, see the *Drosophila* courtship experiment reported in section 3.5.

11.4.8 Independence

In the simplest class of statistical models, the responses on distinct observational units are assumed to be distributed independently given the treatment assignment, i.e., $Y(u_1), \ldots, Y(u_n)$ are independent given $\mathbf{t}$; In more complicated situations such as agricultural field experiments or crossover designs or studies involving infectious diseases, the responses on distinct observational or experimental units cannot reasonably be assumed to be conditionally independent given the treatment. For example, geographic or temporal or familial relationships may induce correlations that are detectable in the data and must be accommodated in the probability model. Most of the examples illustrated in this book exhibit non-trivial correlations.

As a general rule, lack of independence is not a serious problem provided that it is recognized, and steps are taken to make accommodations in the analysis. Ordinarily, this means that block factors and other relevant relationships are recorded at baseline and used in the model to accommodate correlations.

11.4.9 Interference

If the response Y_u for one experimental unit is statistically independent of the treatment applied to other units, we say there is no interference, or no pairwise interference. Lack of interference is a conditional independence assumption $Y_u \perp\!\!\!\perp \mathbf{t} \mid t_u$; it does not imply independence of components, nor does independence imply lack of interference.

Unless the experiment is deliberately designed to study it, interference is best avoided by design. A typical field experiment uses guard strips to separate adjacent plots; guard strips reduce interference from root competition and fertilizer seepage, but they seldom eliminate spatial correlation.

The more general definition of no interference $Y[U'] \perp\!\!\!\perp \mathbf{t} \mid \mathbf{t}[U']$ for each $U' \subset U$ requires the distribution of each restriction $Y[U']$ to depend only on the treatment restricted to U'. Independence and lack of interference are not so much statements of fact or fiction as they are mathematical restrictions on probability distributions. But both have implications for model formulation and analysis.

11.4.10 State space

In a statistical model, the response is regarded as a random variable, a function $u \mapsto Y(u)$ on the observational or experimental units taking values in the *state space* $\mathcal{S}$, (often the real numbers). In certain settings, particularly in observational studies where all variables are regarded as responses on an equal footing, the synonym *feature space* may be used. Usually the feature space is $\mathbb{R}^k$ for some fixed k.

It is important that the state space contain a point for every possible response-related post-baseline event that could possibly be recorded. In a pharmaceutical trial for cholesterol reduction, individual patients give informed consent and agree to abide by the protocol. However, subsequent participation is ultimately voluntary, and not all patients comply by taking their medications on the prescribed schedule. If it is recorded, compliance or the degree of compliance is a response variable, and failure to comply is one component of the response. The probability model is a probability distribution on the state space, which specifies the compliance probability, the conditional distribution given compliance, and the probability of compliance given the cholesterol levels past and future.

In all cases, the state space is a fixed measurable set, the same set for every unit, either observational unit or experimental unit, regardless of covariates. However, this restriction may lead to mathematical contortions. Consider an animal breeding study where each experimental unit is a family, and the response is measured on individual family members (offspring only) at age six weeks. Suppose that family size x is a covariate recorded at baseline, in which case the response Y_u for a family of size $x(u)$ is a point in $\mathbb{R}^{x(u)}$. The variation of the state space from one experimental unit to another depending on the covariate $x(u)$ appears to violate the definition of state space as a fixed set. But this violation is a mathematical illusion. We can simply re-define the state space to be the

disjoint union $\mathcal{S} = \cup_{k\geq 0}\mathbb{R}^k$, and construct the probability distribution on $\mathcal{S}$ in such a way that all of the probability mass for unit u resides in the component $x(u)$ of the state space,

$$\mathrm{pr}(Y_u \in \mathbb{R}^k) = \begin{cases} 1 & x(u) = k \\ 0 & \text{otherwise.} \end{cases}$$

Note that x is not a random variable, so we have not written this as a conditional probability statement.

If the measurements were weights at birth rather than later at six weeks, the baseline would necessarily have to be pre-natal, implying that family size X is a part of the response, not a covariate recorded at baseline. In that setting the response Y is a random variable taking values in $\mathcal{S}$, and the response distribution F determines the distribution of X by $\mathrm{pr}(X = k) = F(\mathbb{R}^k)$ (including $k = 0$). The conditional distribution given X is a function that associates with each integer $k \geq 0$ a probability distribution $F(\cdot \mid X = k)$ such that $F(\mathbb{R}^k \mid X = k) = 1$.

11.4.11 Censoring and state-space evolution

In a study of survival times following surgery, each patient is one unit, and the response $Y_u > 0$ is, *prima facie* at least, a point in $\mathbb{R}^+$, the positive real line. Only the most persnickety mathematician would bother to add a point at infinity to cover the remote possibility of immortality, which cannot be ruled out solely on mathematical grounds. However, the response $Y_u^{(t)}$ as it exists today or at the time of analysis, say $t = 1273$ days post-recruitment, is either a failure time in the interval $t^- = (0, t]$, or a not-yet-failure corresponding to the 'point' t^+, which is required to exist as a point in the state space for today. In other words, $\mathcal{S}^{(t)} = t^- \cup \{t^+\}$, the union of a bounded interval and a topologically isolated 'point' exceeding each number in the interval. The limit $\mathcal{S}^{(\infty)} = \mathbb{R}^+ \cup \{\infty\}$ differs from $\mathbb{R}^+$ by one isolated point that exceeds every real number.

To say the same thing in another way, the state space is a filtration, which evolves as an increasing σ-field in calendar time.

Every probability distribution F on $\mathcal{S}^{(\infty)}$ is determined by its hazard measure Λ on $\mathbb{R}^+$ and its survivor function $F(t^+) = \exp\bigl(-\Lambda(t^-)\bigr)$, which is decreasing as a function of t. If the total hazard $\Lambda(\mathbb{R}^+)$ is finite, the atom of immortality $F(\{\infty\}) = \exp(-\Lambda(\mathbb{R}^+))$ is strictly positive; otherwise the atom is zero. With respect to the state of information at time t, the probability density at $y \in \mathcal{S}^{(t)}$ is $\Lambda(dy)\exp\bigl(-\Lambda(y^-)\bigr)$ for $0 < y \leq t$, and $\exp\bigl(-\Lambda(y^-)\bigr)$ for $y = t^+$. In particular, if Λ is proportional to Lebesgue measure on $\mathbb{R}^+$, the density is $\lambda e^{-\lambda s}\,ds$ for $0 < s \leq t$ with an atom $e^{-\lambda t}$ at t^+.

Being alive at the time of analysis is one unavoidable form of censoring. In practice, some patients disappear off the radar screen at a certain point $t > 0$, and their subsequent survival beyond that time cannot be ascertained. These also are typically regarded as censored at the last time they were known to be alive.

11.4.12 Longitudinal study

In a longitudinal study, also called a panel study, each physical unit is measured at a sequence of time points. Growth studies, of plants or of animals, are of this type, the response $Y(i,t)$ being height or weight of unit i at time t. Usually the design calls for measurements to be made at regular intervals, but in practice the intervals tend to be irregular to some degree, particularly for studies involving human subjects.

A typical longitudinal design has a large number of subjects measured on a relatively small number of occasions. The first of these measurements is made at or pre-baseline. If the experiment has a randomized treatment assignment, the first measurement is ordinarily pre-randomization before the treatment is decided, and certainly before it can have had an effect. In the modelling and analysis, it may be necessary to include a null treatment level to denote pre-randomization status; this level is in addition to the control and active post-baseline levels.

11.4.13 Cemetery state

A situation arises in geriatric and other medical studies where, beginning at recruitment, measurements on physical or mental capacity are made annually on patients—but only while they are alive. All patients ultimately die, and the number k_i of measurements on patient i is a major part of the response, which is closely connected with survival time. In this setting, each patient may be regarded as an observational unit, in which case the response $Y_i = (Y_i(0), \ldots, Y_i(k_i - 1))$ is a point in the state space $\cup_{k \geq 0} \mathbb{R}^k$ implying death before time k_i. Alternatively, if each patient-time combination is regarded as one observational unit, it is necessary to add to the real numbers an absorbing state, such that $Y_i(t) = \flat$ implies that patient i is dead at time t. The state space for one observational unit is $\mathbb{R} \cup \{\flat\}$; the state space for one experimental unit (patient) is $\mathcal{S}^{(\infty)} = (\mathbb{R} \cup \{\flat\})^\infty$, each sequence $\flat$-padded on the right where needed.

As always, the state space at calendar time s includes only those events observed or observable up to that time; the state space is censored by the calendar, not by the death of patients.

11.5 Non-comparative studies

11.5.1 Examples

An experiment designed to measure the speed of light *in vacuo* is not comparative; the goal is not to estimate the ratio of the speed *in vacuo* relative to that in some other medium, but to estimate the absolute speed in km/s for a particular medium. A survey whose goal is to estimate the prevalence of COVID-19 antibodies in Santa Clara County in April 2020 is not comparative; the goal is not to estimate the prevalence in Santa Clara relative to that in San Mateo, but

to estimate the absolute prevalence as a percentage of the county population. An opinion poll with the aim of predicting the outcome of a plebiscite or general election is not comparative; the goal is to predict the outcome of the election on a particular day.

The avoidance of bias or systematic errors is important in all branches of science, but it is especially important in non-comparative studies. The next few sections consider the effect of response heterogeneity in a stratified population, finite or infinite.

11.5.2 Stratified population

A function $x : \mathcal{U} \to [k]$ taking values in the finite set $[k] = \{1, \ldots, k\}$ determines a partition of the units into k disjoint subsets, $\mathcal{U}_1, \ldots, \mathcal{U}_k$ called strata or blocks:

$$\mathcal{U}_r = \{u : x(u) = r\}.$$

In general, $\mathcal{U}$ may be finite or infinite; if $\mathcal{U}$ is not finite, at least one stratum is also not finite. In practice, if $\mathcal{U}$ is infinite, all of the strata are also infinite.

For current-population sampling applications, $\mathcal{U}$ is finite; to a close approximation x is known from the preceding census, so the strata sizes are also known in the same sense. Every classification variable such as *sex* determines a stratification; every pair of variables such as *(sex, location)* determines a finer stratification, and so on. For example, *location* might have levels *rural, suburban, urban*. The classification variables that are available for survey-sampling are mostly restricted to those recorded in the census.

11.5.3 Heterogeneity

Heterogeneity means that the distribution of response values in one stratum is not the same as the distribution in another stratum, or at least similarity is not to be assumed. The implication, ironically, is that the values within each stratum can be taken as exchangeable—infinitely exchangeable in the case of infinite strata, or finitely exchangeable otherwise. Exchangeability is either an explicit assumption, or it is forced as a consequence of random sampling.

11.5.4 Random sample

A *simple random sample* of size n taken from a finite population of size N is a random subset uniformly selected from the set of all subsets of size n. More correctly, a simple random sample is a function φ chosen uniformly at random from the set of 1–1 functions $[n] \to [N]$. The sample $(\varphi_1, \ldots, \varphi_n)$ is an ordered subset consisting of n distinct units taken from the population, and the sample value is $(Y_{\varphi(1)}, \ldots, Y_{\varphi(n)})$.

Operationally speaking, we first arrange the population units $1, \ldots, N$ in uniform random order $\sigma(1), \ldots, \sigma(N)$ by a uniform random permutation σ. By definition, the permuted values $(Y_{\sigma(1)}, \ldots, Y_{\sigma(N)})$ are finitely exchangeable in the usual sense that the distribution is unaffected by permutation. The

leading subset $\varphi = (\sigma(1), \ldots, \sigma(n))$ is a simple random sample, and the sample value is $(Y_{\sigma(1)}, \ldots, Y_{\sigma(n)})$. In other words, a simple random sample is a fixed sample taken from the randomized population. Simple random sampling is the guarantor of exchangeability.

11.5.5 Stratified random sample

A stratified random sample with sizes $n_1, \ldots, n_k$ consists of k simple random samples, one independent sample from each stratum.

11.5.6 Accessibility

It is possible to select a finite random sample from an infinite population. But simple random sampling and stratified sampling are possible only for finite populations. In practice, any form of random sampling is feasible only for the sub-population that is currently accessible. For example, a population consisting of a lineage of breeding flies that evolves in time is only partly accessible in any bounded temporal window.

11.5.7 Population averages

The mean for stratum r

$$\mu_r = E\big(Y_u : x(u) = r\big)$$

is either a finite average if $\mathcal{U}_r$ is finite, or a distributional mean of exchangeable random variables otherwise. In a finite or infinite population with strata fractions $(\pi_1, \ldots, \pi_k)$ adding to one, the weighted linear combination

$$\mu_\pi = \pi_1\mu_1 + \cdots + \pi_k\mu_k$$

is called the population average.

In the case of a locally finite population consisting of $N_t = \#\mathcal{U}_t$ units at time t, $N_r(t)$ is the stratum total, $\pi_r(t) = N_r(t)/N_t$ is the stratum fraction, and the democratic average $\mu_{\pi(t)} = \sum_{u\in\mathcal{U}_t} Y_u/N_t$ is the arithmetic mean in the current population.

11.5.8 Target of estimation I

In a stratified population, the target of estimation is usually the stratum mean vector $\mu = (\mu_1, \ldots, \mu_k)$. However, there are various applications, particularly related to marketing, opinion polling and voting, where the democratic average plays an outsize role. In the run-up to a crucial plebiscite such as the Brexit referendum, the democratic average of voter preferences looms so large that between-stratum variation is of little consequence.

11.5.9 Inverse probability weighting

Consider a stratified population consisting of 12m voters, 5m urban, 4m suburban and 3m rural. In a stratified random sample in Oct 2020, 500 voters out of 1000 declared that they would vote for candidate T; the breakdown by strata was as follows.

	Urban	Suburban	Rural	Total
Stratum size	5m	4m	3m	12m
π	5/12	4/12	3/12	1
Sample size	400	300	300	1000
Candidate T	110	140	250	500
$\bar{y}$	0.275	0.467	0.833	0.500

Note that the stratum relative proportions 5m:4m:3m are close to the sample fractions 4:3:3, but not exactly the same. The stratum averages for this sample are $\bar{y} = (0.275, 0.467, 0.833)$, and the population-weighted linear combination of stratum averages is

$$\hat{\mu}_\pi = 0.275 \times 5/12 + 0.467 \times 4/12 + 0.833 \times 3/12 = 0.4785$$

which is less than the equally-weighted poll average 500/1000.

The preceding calculation is an instance of a weighted linear combination of sample values,

$$\hat{\mu}_\pi = \sum_{i \in S} w_i Y_i \Big/ \sum_{i \in S} w_i,$$

where the weights are inversely proportional to the first-order sample inclusion probabilities (Horvitz and Thompson, 194?). Each urban voter has a sample inclusion probability 400/5m, so $w_i = 5/400$; each suburban voter has inclusion probability 300/4m, so $w_i = 4/300$; and each rural voter has inclusion probability 300/3m, so $w_i = 3/300$. The sum of these weights is 12, and the linear combination is displayed in the preceding paragraph.

11.5.10 Target of estimation II

The calculation illustrated in the preceding section is as obvious as it is uncontroversial. It is obvious as a matter of arithmetic, and it is uncontroversial because of the political setting used for its illustration. But inverse-probability weighting is not something to be taken for granted in other settings that might appear superficially similar.

Consider the COVID-19 antibody prevalence study for Santa Clara County in April 2020. The main controversy in the Stanford study centered correctly on the false-positive rate of the antibody test, which was of a magnitude similar to the reported prevalence. See the online blog by Gelman (????). For present purposes, we set that matter aside and suppose optimistically that the false-positive rate is zero.

Suppose that a similar set of numbers—suitably scaled to represent plausible prevalences—had arisen in the COVID-19 antibody prevalence study.

	Urban	Suburban	Rural	Total
Stratum size	0.5m	0.4m	0.3m	1.2m
π	5/12	4/12	3/12	1
Sample size	400	300	300	1000
Antibody cases	4	8	13	25
$\bar{y}$	0.010	0.027	0.043	0.050

Would it be appropriate to use the same weighted procedure

$$\hat{\mu}_\pi = 0.010 \times 5/12 + 0.027 \times 4/12 + 0.043 \times 3/12 = 0.024$$

and report only the county-wide antibody prevalence at 2.4%? I should hope not!

The crucial difference is not the numbers but the setting. For the political poll, the current-population average is the natural target mandated by democratic principles and supported by the force of law. In the epidemiological setting, the democratic average or prevalence is a natural summary, but it does not carry an equivalent epidemiological or legal mandate. Nor is it necessarily the most interesting summary or the most striking feature to emerge from such a study. In the table shown above, the observed prevalence in the rural community is more than four times that in the urban community. Admittedly, the case numbers are small, so the ratio in the population might not be so extreme. But a risk ratio or prevalence ratio as large as 3–4 calls out for an explanation, and that finding could be more interesting epidemiologically than the particular value of the county-wide prevalence.

The main point is that the overwhelming focus on prevalence is a distraction that has the potential to divert attention away from features that are epidemiologically more interesting. Any epidemiologist who reported only the prevalence of 2.4% would be derelict in his duty to draw attention to the extreme variation in rates for urban versus rural communities. To conclude, inverse-probability weighting is satisfactory as a summary statistic for a stratified population in two circumstances only: either the democratic average is mandated by law; or the degree of heterogeneity is moderate. In the latter case, the choice of weights matters little.

11.5.11 DATE

Consider a randomized-blocks design in which block r is a finite sample from stratum r. Treatment is assigned by randomization within each block, and the sample mean difference T_r for block r is a natural estimate of the effect of treatment for units in stratum r. The weighted linear combination

$$T_\pi = \pi_1 T_1 + \cdots + \pi_k T_k$$

is called the democratic average treatment effect (DATE).

For the present discussion, it is immaterial whether the strata are finite or infinite. The response is a random process Y that is exchangeable within strata only. Each block of the design is a sample $U_r \subset \mathcal{U}_r$ from that stratum. It is immaterial whether the sample is random or fixed; but if it is random, it must be independent of Y.

The remarks in the previous section are meant to draw attention to the limitations of every combination of this sort in a situation where there is appreciable between-stratum inhomogeneity. At worst, an excessive emphasis on DATE could conceal a qualitative interaction where the treatment effect for males is the same as that for females, but opposite in sign. It is a strategic error of judgement to disregard inter-stratum heterogeneity and to focus attention solely on a particular average. A combined estimate is a reasonable summary only if the inter-stratum heterogeneity is acceptably small.

11.6 Statistical pprinciples

11.6.1 Guiding principles

11.6.2 Likelihood principle

The likelihood principle is concerned with parametric inferences, i.e., inferential statements or conclusions about the parameter given the data. Suppose that the response density is $f(\cdot;\theta)$. Two points $y^{(1)}, y^{(2)}$ in the observation space give rise to the same likelihood function if the density ratio $f(y^{(1)},\theta)/f(y^{(2)},\theta)$ is constant in θ. According to the weak version of the likelihood principle, two points that determine the same likelihood function must lead to identical conclusions about θ.

The likelihood principle is confined to statements about parameters. It says nothing about other sorts of inferences involving statements about the observation space. In particular, $y^{(1)}$ and $y^{(2)}$ do not ordinarily lead to identical predictions.

The principle can be illustrated by two observations on a Bernoulli sequence $Y_i \sim \text{Ber}(\theta)$. For a sample of size $n = 20$, the density function and the likelihood function are

$$f(y;\theta) = \theta^s(1-\theta)^{20-s},$$

where the number of successes $s = y_\bullet$ is the sufficient statistic. Suppose that the two sequences are

$$y^{(1)} = (0,1,0,1,0,1,0,1,0,1,0,1,0,1,0,1,0,1,0,1)$$
$$y^{(2)} = (0,0,0,0,0,0,0,0,0,0,1,1,1,1,1,1,1,1,1,1).$$

Both sequences have $s = 10$, so the likelihood functions are equal. According to the weak likelihood principle, both observations must lead to the same conclusion about θ.

From the present viewpoint, it is immaterial whether we adopt a Bayesian-type beta-binomial model or we attempt to construct a confidence interval. Neither inferential approach would be satisfactory given either sequence.

11.6.3 Lesson of the likelihood principle

The premiss of the likelihood principle is that the statistician buys into the Bernoulli model exactly as stated, with no probabilistic reserve in the form of an opt-out clause to cover buyer's remorse. The lesson of experience is simply to avoid being sandbagged. The likelihood principle is not rejected, but a cautious applied statistician invariably adopts the stated model provisionally, with adequate reserves to cover mistakes, misunderstandings or unanticipated events. To do otherwise would be a serious error of professional judgement.

In effect, a consulting statistician using the Bernoulli model proceeds as follows. With probability 0.65 the sequence is Bernoulli with constant parameter θ; with probability 0.10 the sequence is Bernoulli with non-constant parameter; with probability 0.10, the sequence has some temporal dependence, possibly Markov; with probability 0.10, the design has some other feature that might lead in a different direction. The weights shown here may be varied to match the incidental information relevant to the context, but their sum is strictly less than one. After observing either sequence $y^{(1)}$ or $y^{(2)}$ the first weight component is drastically reduced.

Chapter 12

Probability distributions

12.1 Exchangeable processes

12.1.1 Unconditional exchangeability

Recall that a process with state space $\mathcal{S}$ associates with each sample S consisting of finitely many distinct units taken in a specified order, a probability distribution P_S on the observation space $\mathcal{S}^S$. Thus $P_S(A)$ is the probability of the event

$$(Y_{u_1}, \ldots, Y_{u_n}) \in A.$$

The process is said to be *unconditionally exchangeable* if two samples of the same size have the same joint distribution. In other words, the process is exchangeable if $\#S = \#S' = n$ implies $P_S(A) = P_{S'}(A)$ for every event $A \subset \mathcal{S}^n$. In particular, two samples consisting of the same distinct units taken in different orders have the same distribution. In that case, the n-dimensional joint distribution is usually denoted by P_n.

Unconditional exchangeability is a very demanding property that is seldom satisfied in scientific work, where experiments are almost invariably comparative. The goal is usually to study *differences* between one distribution P_u and another $P_{u'}$ that are related to covariate effects or treatment effects for pairs such that $x(u) \neq x(u')$. Nonetheless, a version of exchangeability is needed in order to make progress in situations where inhomogeneities associated with baseline covariates are anticipated.

12.1.2 Regression processes

Let $u \mapsto x_u$ be a covariate defined as a function $\mathcal{U} \to \mathcal{X}$ for every unit in the population. It is assumed implicitly that the only baseline relations are the identity function $\delta_{u,u'}$, which tells us whether or not two units are the same, and the one-block constant function $J_{u,u'} = 1$ for all pairs.

To each sample S there corresponds a covariate configuration $\mathbf{x}[S]$, which is usually encoded as a model matrix X whose rows are indexed by units $u \in$

S. The manner in which dose levels or classification factors are encoded is immaterial. The process is said to be regression-exchangeable if two samples of distinct units having the same covariate configuration automatically have the same joint distribution. In other words, $\mathbf{x}[S] = \mathbf{x}[S']$ implies $P_S = P_{S'}$.

This form of exchangeability is usually taken for granted in applied work, so much so that it is rarely judged to be worth even a brief comment. For instance, the great majority of generalized linear models are regression-exchangeable in this sense.

Planar white noise and Poisson processes are less obvious examples. In both cases, each unit is a planar subset and the covariate $x_u = \Lambda(u)$ is planar Lebesgue measure. For any collection of disjoint subsets, the white-noise values $Y(u)$ are independent zero-mean Gaussian variables with variance x_u. The Poisson-process values are independent Poisson variables with mean x_u. Disjointness of subsets is not part of either definition; it is needed here only to comply with the assumption that there are no relationships among the units other than the identity.

12.1.3 Block-exchangeability

Recall that a block factor is an equivalence relation $B\colon \mathcal{U}^2 \to \{0,1\}$ on the units, which partitions the population into disjoint non-empty subsets called blocks. The restriction of B to a finite sample S consisting of n distinct units is a symmetric binary matrix of order n, which partitions the sample into disjoint blocks. If the units are arranged in suitable order, $B[S]$ is block-diagonal.

The process is said to be block-exchangeable if two samples having the same block structure also have the same response distribution, i.e., $B[S] = B[S']$ implies $P_S = P_{S'}$. For samples of size one, $B[S] = B[S']$ is automatic, so all one-dimensional distributions are equal. For samples of size two with $u_1 \neq u_2$, either $B(u_1, u_2) = 0$ or $B(u_1, u_2) = 1$, so there are two distinct two-dimensional distributions.

A Gaussian process is block-exchangeable if and only if the mean vector is constant $\mu \in \mathbf{1}$, and the covariance matrix for a sample S is a non-negative linear combination of the three matrices

$$\operatorname{cov}(Y[S]) = \sigma_0^2 I_n + \sigma_1^2 B[S] + \sigma_2^2 J_n,$$

where $J_n(u, u') = 1$. In particular, $E(Y_u) = E(Y_{u'})$ for every pair, regardless of the block sizes, and the covariances are also independent of the block sizes.

The block-exchangeability assumption that P_S depends only on $B[S]$ is most natural if all population blocks are equal in size, which usually means infinite. If some or all of the population blocks are finite we can associate with each unit u the number $x(u)$, which is the size of the block in the population to which u belongs. When this is done, x is a covariate, and the implications of exchangeability are drastically different because P_S may depend on $x[S]$ in addition to $B[S]$.

12.1.4 Stationarity

Let $\mathcal{U} = \mathbb{R}$ be the index set. No covariates are registered, and the temporal difference $R(t, t') = t' - t$ is the only registered relationship. The restriction of R to a sample is a square matrix $R[S]$ of signed temporal differences; two samples are called congruent or structurally equivalent if $R[S] = R[S']$. Congruence is an equivalence relation on samples, denoted by $S \cong S'$; in this setting, it implies $S' = S + h$ for some real number h.

A process with distributions P_S is said to be stationary, or invariant with respect to temporal translation, if $S \cong S'$ implies $P_S = P_{S'}$. In particular, stationarity implies that all singletons have the same distribution $Y_t \sim Y_{t'}$.

Any transformation of R, such as the absolute value R^+, is also a relationship on the units; R^+ is said to be a coarser relationship than R because the partition defined by R is a sub-partition, or a finer partition, of that defined by R^+. In particular, $R^+(t, t') = R^+(t', t)$ is symmetric whereas R is not. If $R^+[S] = R^+[S']$ implies $P_S = P_{S'}$, the process is not only stationary but also reversible.

12.1.5 Exchangeability

The general principle of exchangeability is easy to understand and straightforward to state. Two samples are said to be congruent if they have the same structure; this is understood broadly to include not only covariates but also pairwise and higher-order relationships among units. Congruence, denoted by $S \cong S'$, is an equivalence relation among samples. It implies that the samples are of equal size, $S = (u_1, \ldots, u_n)$ and $S' = (u'_1, \ldots, u'_n)$; it implies that the covariate values are equal $x(u_i) = x(u'_i)$; it implies that all pairwise relationships are equal $R(u_i, u_j) = R(u'_i, u'_j)$, and so on. In particular, $u_i = u_j$ if and only if $u'_i = u'_j$.

Exchangeability is nothing more than the statement that congruent samples are required to have the same response distribution, i.e., $S \cong S'$ implies $P_S = P_{S'}$. For singletons, $x(u) = x(u')$ implies $P_u = P_{u'}$; for pairs $\big(x(u_1), x(u_2)\big) = \big(x(u'_1), x(u'_2)\big)$ and $R(u_1, u_2) = R(u'_1, u'_2)$ together imply $P_{u_1,u_2} = P_{u'_1,u'_2}$.

Exchangeability is not a statement of biological, medical or scientific fact. It is a mathematical statement of equity or equality, corresponding roughly to fairness or or even-handedness, which implies that probabilistic statements are based only on facts that are registered at baseline. By supposition, all relevant facts are encoded in $\mathbf{x}$. Without a symmetry condition of this sort, conveniently selected alternative facts are no less compelling than recorded facts. Such a world view may be acceptable in politics, but it is an impediment to progress in science.

12.1.6 Axiomatic point

Block exchangeability is defined in section 12.1.3 by the statement $B[S] = B[S']$ implies $P_S = P_{S'}$. This matrix equality $B[S] = B[S']$ makes sense only if both samples are ordered, which is the convention adopted throughout these notes

although it is not the standard statistical convention. However, if $B(u, u') = 1$ and $u \neq u'$, the ordered samples $S = (u, u)$ and $S' = (u, u')$ satisfy $B[S] = B[S']$. Despite the statement, block exchangeability does not imply $(Y_u, Y_{u'}) \sim (Y_u, Y_u)$ for pairs belonging to the same block. Why not?

In the first paragraph of section 12.1.3, the samples were required to consist of distinct units, so the difficulty was eliminated by this restriction. The real reason for the restriction is a more fundamental consequence of standard set theory, namely that the identity function is axiomatically a registered relationship for every set. Structural equivalence of samples implies both $I[S] = I[S']$ for the identity, and $B[S] = B[S']$ for the block factor. In the case of a regression process, structural equivalence implies $I[S] = I[S']$ and $\mathbf{x}[S] = \mathbf{x}[S']$. With this understanding the restriction to distinct units in section 12.1.1–12.1.3 is not needed.

12.1.7 Block randomization

Let B be a given partition of the finite set $[n]$ into blocks, and let $\mathcal{G}$ be the group of permutations that preserves the partition. In other words, $\mathcal{G}$ is the set of permutations $\sigma\colon [n] \to [n]$ such that $B_{\sigma(i),\sigma(j)} = B_{i,j}$.

To any vector $y = (y_1, \ldots, y_n)$ in $\mathcal{S}^n$ there corresponds a randomized vector $Y = y\sigma$ whose components $Y = (y_{\sigma(1)}, \ldots, y_{\sigma(n)})$ are obtained by composing the given vector with a random permutation σ uniformly distributed over the group. Randomization defines a process with state space $\mathcal{S}$ and finite index set $[n]$. By definition, for each $\tau \in \mathcal{G}$, the group product $\sigma\tau$ is also uniformly distributed, so the permuted random vector $Y\tau = y\sigma\tau$ has the same distribution as Y. The distribution is invariant with respect to the natural sub-group of permutations associated with the block factor.

If all blocks of B are of equal size, the distribution of Y is block-exchangeable in the sense of section 12.1.3. Otherwise, if there are blocks of different sizes, Y is not block-exchangeable. For example, if $n = 7$, and $B = 1|2|34|567$ is a partition into four blocks, the group contains $2\times2\times6 = 24$ elements. Block randomization implies $Y_1 \sim Y_2$ because the transposition $1\leftrightarrow2$ is a group element; it also implies $Y_3 \sim Y_4$ and $Y_5 \sim Y_6 \sim Y_7$ for similar reasons. But it does not imply $Y_1 \sim Y_3$ or $Y_3 \sim Y_5$ because there is no group element such that $\sigma(1) = 3$ or $\sigma(3) = 5$.

12.2 Families with independent components

12.2.1 Parametric models

A parametric statistical model associates with each parameter point $\theta \in \Theta$ a probability distribution P_θ. In general, a distribution is a process which associates with each sample $S \subset \mathcal{U}$ a probability distribution $P_{\theta,S}$ on the observation space $\mathcal{S}^S$. If the process has independent components, the description can be simplified to a great extent by focusing on the one-dimensional marginal distributions.

The following examples illustrate a range of possibilities.

12.2.2 IID model I

The parameter space is the set of probability distributions defined on the state space, say $\mathcal{S} = \mathbb{R}$ with Borel subsets. In other words, the sequence Y_u for $u \in \mathcal{U}$ has independent and identically distributed components $Y_u \sim \theta$.

Properties of a statistical model are often gauged by their behaviour under the action of a suitable group or semi-group of measurable transformations $g\colon \mathcal{S} \to \mathcal{S}$. If $Y_1, \dots$ are independent and identically distributed with distribution $\theta \in \Theta$, then the transformed variables $g(Y_1), g(Y_2), \dots$ are independent and identically distributed with parameter $g\theta \in \Theta$, where

$$g\theta(A) = P_\theta(gY \in A) = P_\theta(Y \in g^{-1}A) = \theta(g^{-1}A)$$

for $A \subset \mathcal{S}$. Since Θ is the set of Borel distributions on $\mathcal{S}$, the transformed distribution is simply another point $\theta' = g\theta$ in the same parameter space. Provided that g is invertible, the two specifications

$$Y_1, Y_2, \dots \overset{\text{iid}}{\sim} \theta, \quad \text{and} \quad gY_1, gY_2, \dots \overset{\text{iid}}{\sim} g\theta$$

are mathematically equivalent. Equivalence, or equi-variance, implies that any inferential statement about θ after observing y must be tied to inferential statements about $g\theta$ after observing gy.

To each observation point $y = (y_1, \dots, y_n)$ there corresponds an empirical distribution function

$$\hat{\theta}(A; y) = n^{-1} \sum_{i=1}^{n} \delta_{y_i}(A) = n^{-1} \#\{i \in [n] : y_i \in A\}.$$

The function $y \mapsto \hat{\theta}(y)$ is equi-variant in the sense that $\hat{\theta}(gy) = g\hat{\theta}(y)$. The transformation $y \mapsto gy$ acts component-wise $\mathcal{S}^n \to \mathcal{S}^n$, while $\theta \mapsto g\theta$ is the induced transformation on distributions on $\mathcal{S}$. For invertible transformations, this means $\hat{\theta}(y) = g^{-1}\hat{\theta}(gy)$. The empirical distribution is sometimes called the nonparametric maximum-likelihood estimate, or the bootstrap estimate.

12.2.3 IID model II

The parameter space is the set of Gaussian distributions on the real line. Since the Gaussian distribution is determined by its mean and variance, this statement means one of the following:

$$\begin{aligned} \Theta = \mathbb{R}^2; &\quad P_\theta = N(\theta_1, \theta_2^2); \\ \Theta = \mathbb{R} \times (0, \infty); &\quad P_\theta = N(\theta_1, \theta_2^2); \\ \Theta = \mathbb{R}^2; &\quad P_\theta = N(\theta_1, e^{\theta_2}). \end{aligned}$$

Given a parameter point θ, the components are independent and identically distributed $Y_u \sim P_\theta$ on the real line.

The three versions are not mathematically equivalent. In version one, the two distinct points $(\theta_1, \pm\theta_2)$ define the same distribution, so the parameter is not identifiable. In addition, the boundary subset of Dirac distributions $N(\theta_1, 0)$ is included in the first version, but not in the other two. Versions two and three are equivalent in the sense that they contain the same set of non-degenerate distributions. For the most part, differences of this sort are not of great importance, and are usually overlooked in applications.

All three versions are affine equi-variant in the sense that $Y_u \sim P_\theta$ implies $gY_u \sim P_{g\theta}$ for affine transformations $y \mapsto gy = g_0 + g_1 y$ with $g_1 > 0$. The induced transformation on the parameter space is group composition

$$(\theta_1, \theta_2) \mapsto (g_0 + g_1\theta_1, g_2\theta_2)$$

for versions one and two, and

$$(\theta_1, \theta_2) \mapsto (g_0 + g_1\theta_1, \theta_2 + 2\log|g_2|)$$

in the third version where θ_2 is the log variance.

For a sample of size $n \geq 2$ and an observation $y \in \mathbb{R}^n$, the usual estimate of the parameters for version 1 or 2 is

$$\hat{\theta}_1 = \bar{y}_n; \quad \hat{\theta}_2^2 = s_n^2 = \sum (y_i - \bar{y}_n)^2/(n-1).$$

This estimator is affine equivariant in the sense that $\hat{\theta}(gy) = g\hat{\theta}(y)$ for affine transformations $y \mapsto gy$ acting component-wise. For this purpose, the divisor $n-1$ could be replaced by n. Similar remarks with minor modifications apply to version 3.

The Cauchy distribution $C(\theta)$ with median θ_1 and probable error $|\theta_2|$ has a density

$$\frac{|\theta_2|\, dy}{\pi\, |y - \theta|^2},$$

where $\theta = \theta_1 + i\theta_2$ is a complex number, and $|y-\theta|^2$ is the squared modulus. Apart from the specific formulae for parameter estimates, all of the preceding remarks apply equally to the Cauchy family and every symmetric location-scale family on the real line.

12.3 Non-i.d. models

12.3.1 Classification factor

We consider a simple model for a process in which each individual is classified as male or female. All values are independent, and they are identically distributed for individuals of the same sex. Each model is defined by a parameter space Θ, a function $\theta \mapsto P_\theta$ for males, i.e., for all u such that $x(u) = M$, and a function $\theta \mapsto Q_\theta$ for all u such that $x(u) = F$. The symbol $\mathbb{R}_+$ may be interpreted as either the set of non-negative numbers or the set of strictly positive numbers.

Table 12.1 Parameterization of statistical models for a classification factor

	θ	Θ	P_θ	Q_θ	Eqv?
(i)	(π_0, π_1)	$(0,1)^2$	$\text{Ber}(\pi_0)$	$\text{Ber}(\pi_1)$	✓
(ii)	(μ_0, μ_1, σ)	$\mathbb{R}^2 \times \mathbb{R}_+$	$N(\mu_0, \sigma^2)$	$N(\mu_1, \sigma^2)$	✓
(iii)	$(\mu, \sigma_0, \sigma_1)$	$\mathbb{R} \times \mathbb{R}_+^2$	$N(\mu, \sigma_0^2)$	$N(\mu, \sigma_1^2)$	✓
(iv)	$(\mu_0, \mu_1, \sigma_0, \sigma_1)$	$\mathbb{R}^2 \times \mathbb{R}_+^2$	$N(\mu_0, \sigma_0^2)$	$N(\mu_1, \sigma_1^2)$	✓
(v)	(μ_0, μ_1, σ)	$\mathbb{R}^2 \times \mathbb{R}_+$	$N(\mu_0, \sigma^2)$	$N(\mu_1, 2\sigma^2)$	
(vi)	$(\mu_0, \mu_1, \sigma_0, \sigma_1)$	$\mathbb{R}^2 \times \mathbb{R}_+^2$	$N(\mu_0, \sigma_0^2)$	$C(\mu_1, \sigma_1)$	
(vii)	(θ_0, θ_1)	$\mathcal{D}(\mathbb{R}) \times \mathcal{D}(\mathbb{R})$	θ_0	θ_1	✓

In the first model, the values are independent Bernoulli with success rates π_0 for males and π_1 for females. The parameter space may be extended to include the boundary points if so desired. There is nothing exceptional in this or in the second model, which is Gaussian with sex-dependent mean and constant variance. The third model has a similar structure with constant mean and a sex-dependent variance, while both parameters are sex-dependent in the fourth model. In (vii), the distributions are arbitrary; $Y_u \sim \theta_0$ for males and $Y_u \sim \theta_1$ for females.

Most readers whose experience lies in applied work would blanch at the penultimate suggestion in which male values are Gaussian while female values are distributed as Cauchy. The reasons for this have nothing to do with Cauchy versus Gauss as individuals, or with male variability versus female variability, or with the suitability of this model for any specific application. Instead, they are anchored in the well-established legal principle of 'equality under the law', a desire to avoid overt bias related to visible factors such as race, sex and religion that are, by common agreement, incidental under law.

One mathematical statement of those principles is equi-variance under label-switching. In the present setting, the permutation σ that transposes M with F also switches P with Q. Equi-variance means that to each transposition of factor labels there corresponds a permutation of parameter components such that $P_\theta(A) = Q_{\sigma\theta}(A)$ for every event A. All of the models listed above are equi-variant except for (v) and (vi).

Equi-variance does not imply that the distribution for males is the same as the distribution for females, but it does imply that the set of distributions under consideration is the same for both. Each sex gets to pick one distribution from the same set, so there is equality of opportunity in that sense. However, the Gaussian model in the fifth row of Table 12.1 shows that equality of the sets $\{P_\theta : \theta \in \Theta\}$ and $\{Q_\theta : \theta \in \Theta\}$ is not sufficient for equi-variance.

Equi-variance is not a fundamental principle on a par with Kolmogorov consistency for a stochastic process. It is not even on a par with the principle of exchangeability for individuals having the same covariate value. Equi-variance is reasonably compelling in many circumstances and is a natural default for any factor whose levels are unordered or otherwise unstructured. For example, *occupation* is a classification factor, but the set of levels is not devoid of struc-

Table 12.2 Examples of statistical models for a treatment factor

	Θ_0	$\mathcal{G}$	$g\theta$	P_θ	$P_{g\theta}$
(i)	$\theta \in \mathbb{R}$	$\mathbb{R}$	$\theta + g$	$\mathrm{Ber}\left(\frac{e^\theta}{1+e^\theta}\right)$	$\mathrm{Ber}\left(\frac{e^{\theta+g}}{1+e^{\theta+g}}\right)$
(ii)	$\theta \in \mathbb{R}$	$\mathbb{R}$	$\theta + g$	$\mathrm{Ber}(\Phi(\theta))$	$\mathrm{Ber}(\Phi(\theta + g))$
(iii)	(μ, σ)	$\mathbb{R}$	$(\mu + g, \sigma)$	$N(\mu, \sigma^2)$	$N(\mu + g, \sigma^2)$
(iv)	(μ, σ)	$\mathbb{R}$	$(\mu, \sigma e^g)$	$N(\mu, \sigma^2)$	$N(\mu, \sigma^2 e^{2g})$
(v)	(μ, σ)	$\mathbb{R}^2$	$(\mu + g_1, \sigma e^{g_2})$	$N(\mu, \sigma^2)$	$N(\mu + g_1,\ \sigma^2 e^{2g_2})$
(vi)	(μ, σ)	$\mathrm{Aff}(\mathbb{R})$	$(g_0 + g_1\mu, \sigma g_1)$	$N(\mu, \sigma^2)$	$N(g_0 + g_1\mu,\ \sigma^2 g_1^2)$
(vii)	$\mathcal{P}(\mathbb{R})$	Bic	$\theta \circ g^{-1}$	θ	θg^{-1}

ture. In a survey with limited options, one level might be *none of the above*. Equi-variance is a mathematical solution to the problem of avoiding a form of bias associated with sexist or racist elements in the model.

12.3.2 Treatment factor

A treatment factor and a classification factor are accommodated in a statistical model in very different ways. The distinction is seldom emphasized and it is often not readily apparent. We begin by assuming homogeneity in the sense that no covariate is defined on the units. The first step is to specify the set of reference-level distributions $P_\theta : \theta \in \Theta_0$, each of which is interpreted as a conditional distribution given $T = 0$

$$P(Y_u \in A \mid T = 0; \theta) = P_\theta(A).$$

Treatment has an *effect*, so the second step focuses on the set of possible treatment effects $g \in \mathcal{G}$, and on how each reference-level distribution is modulated by those effects. Each treatment modulation is an action on the parameter space $\theta \mapsto g\theta$ which sends P_θ to $P_{g\theta}$. The interpretation of the action by g is as follows: If the conditional distribution given $T = 0$ is P_θ, and g is the treatment effect, the conditional distribution given $T = 1$ shall be $P_{g\theta}$. To make sense of this, it is necessary that $\mathcal{G}$ be a group acting on Θ_0; the group identity corresponds to a null treatment effect.

The overall parameter space is the product set $\Theta_0 \times \mathcal{G}$. By definition of group action, each transformation $g\colon \Theta_0 \to \Theta_0$ is invertible, so $g\Theta_0 = \Theta_0$. Thus, whatever the treatment effect may be, the set of conditional distributions given $T = 1$ is the same as the set of distributions given $T = 0$. Since the action is a group homomorphism, it is immaterial which level of T is used as the reference level. This condition immediately excludes the fifth and sixth models in Table 12.1 as possibilities for modelling a treatment effect.

For this setting, where there is a single treatment factor and no covariate, the Bernoulli logistic and probit models, are equivalent, and both are equivalent to the Bernoulli model in Table 12.1. The distributions are in 1–1 correspondence, and the only differences are in the parameterizations.

In the first three Gaussian models, the group acts additively on the parameter, sending $(\mu, \log\sigma)$ to $(\mu + g, \log\sigma)$ in example (iii), to $(\mu, g + \log\sigma)$ in example (iv), and $(\mu + g_1, g_2 + \log\sigma)$ in example (v). Each is a reparameterization of one of the Gaussian models listed in Table 12.1. Examples three and four are different reparameterizations of the same set of distributions. They are equivalent in exactly the same sense that the two Bernoulli models are equivalent.

In the fourth Gaussian model, $\theta = (\mu, \sigma)$ and (g_0, g_1) are two points in the group of affine transformations $\mathbb{R} \to \mathbb{R}$, and $g\theta$ is the group composition, which is not additive, i.e., $g\theta \neq \theta g$. The treatment effect is equivalent in distribution to the state-space transformation $Y \mapsto gY$, so that $gY \sim N(g_0 + g_1\mu, \sigma^2 g_1^2) = P_{g\theta}$. Most of the treatment effects exhibited in Table 12.2 are not induced by an action on the state space.

In the last example, Θ_0 is the set of probability distributions on the real line, so the control distribution is an arbitrary distribution defined on Borel subsets. For this setting, the treatment effect can be modelled using any group acting on distributions, whether it is finite-dimensional or infinite-dimensional. The suggestion $\mathcal{G} = \mathrm{Bic}(\mathbb{R})$, meaning bi-continuous transformations $\mathbb{R} \to \mathbb{R}$ having an inverse that is also continuous, is topologically natural. But there are many other possibilities such as the group of Borel-measurable transformations preserving Lebesgue measure. Regardless of the group, the product set $\Theta_0 \times \mathcal{G}$ is not in 1–1 correspondence with the product set $\mathcal{P}(\mathbb{R}) \times \mathcal{P}(\mathbb{R})$ in Table 12.1, so this pair of models is not equivalent for any group.

The group of transformations on distributions may be induced by a transformation $\mathcal{S} \to \mathcal{S}$ on the state space. The set of bi-continuous transformations provides an example of that type, as does $N(\mu, \sigma^2) \mapsto N(\mu + g, \sigma^2)$, which is induced by translation. However, the Bernoulli state space is $\mathcal{S} = \{0, 1\}$ and the group $\mathcal{G} = \mathbb{R}$ does not act on $\mathcal{S}$, so neither Bernoulli transformation is associated with a transformation $\mathcal{S} \to \mathcal{S}$. The second and third Gaussian examples are also not associated with a state-space transformation.

12.3.3 Classification factor plus treatment

Let $x\colon \mathcal{U} \to [k]$ be a k-level classification factor, so that x_u is the class of unit u. We assume that the levels are unordered and otherwise unstructured. For the logistic version of the Bernoulli model, the parameter space Θ_0 consists of k real numbers $\theta_1, \ldots, \theta_k$, where

$$\operatorname{logit} P_\theta(Y_u = 1 \mid T_u = 0) = \theta_{x(u)}.$$

is the conditional log odds of success for unit u, and for every unit in this class. The treatment effect is a group action on the space $\Theta_0 = \mathbb{R}^k$, which sends θ to $g\theta$. In the absence of additional structure (such as an inner product) there are two principal options for the group and its action; either $\mathcal{G} = \mathbb{R}$ and $g\theta = (\theta_1 + g, \ldots, \theta_k + g)$ or $\mathcal{G} = \mathbb{R}^k$ and $g\theta = (\theta_1 + g_1, \ldots, \theta_k + g_k)$.

The first option means that

$$\operatorname{logit} P_\theta(Y_u = 1 \mid T_u = 1) = \theta_{x(u)} + g,$$

so the conditional odds of success for unit u satisfy

$$\log\left(\frac{\operatorname{odds}(Y_u = 1 \mid T_u = 1)}{\operatorname{odds}(Y_u = 1 \mid T_u = 0)}\right) = g.$$

By this odds-ratio yardstick, the effect of treatment is the same number g for unit in every class. No interaction between treatment and the class means that the treatment effect on some specified scale is the same for every class, so this group action implies no interaction on the logistic scale.

The second option means that $g \in \mathbb{R}^k$ acts additively

$$\operatorname{logit} P_\theta(Y_u = 1 \mid T_u = 1) = \theta_{x(u)} + g_{x(u)},$$

so the conditional odds of success for unit u satisfy

$$\log\left(\frac{\operatorname{odds}(Y_u = 1 \mid T_u = 1)}{\operatorname{odds}(Y_u = 1 \mid T_u = 0)}\right) = g_{x(u)}.$$

Unless $g \in \mathbf{1}_k \subset \mathbb{R}^k$, the effect of treatment as measured by the odds ratio is different for each class. This group action implies interaction on the logistic scale. Generally speaking, no interaction on the logistic scale implies interaction on the probit scale, and vice-versa.

12.3.4 Quantitative covariate plus treatment

If a quantitative covariate $x\colon \mathcal{U} \to \mathcal{X}$ is defined on the units, the baseline parameter is a function $x \mapsto \theta(x)$ from $\mathcal{X}$ into some space such as $\mathbb{R}$ or $\mathbb{R}^2$, and Θ_0 is a suitable set of such functions on which the treatment group acts. The statement that x is a quantitative covariate implies that $\mathcal{X}$ is a vector space and that the topology is relevant, so each function $x \mapsto \theta(x)$ in Θ_0 is required to be continuous. Ordinarily, Θ_0 contains the one-dimensional space of constant functions plus the k-dimensional space of linear functionals, where $k = \dim(\mathcal{X})$.

For both Bernoulli models in Table 12.2, $\theta(x)$ is the logit or probit value, which is a real number. In thelogistic version with a constant treatment effect, an individual u whose covariate value is x_u, has log odds of success either $\theta(x_u)$ if $T = 0$, or $\theta(x_u) + g$ if $T = 1$. In the probit version, $\theta(x_u)$ and $\theta(x_u) + g$ are the values on the probit scale. Other link functions such as the complementary log-log operate in the same way. In general, the treatment effect is a group action, $\theta(x) \mapsto \theta(x) + g(x)$ by addition of functions.

In a setting where Θ_0 contains the $k+1$-dimensional space of affine functionals $\theta(x) = \theta_0 + \theta_1(x)$, the simplest options available to accommodate treatment

effects with or without interaction are the following:

$$\begin{aligned}
&\operatorname{logit} P_\theta(Y_u = 1 \mid T_u = 0) = \theta(x_u) \\
&\operatorname{logit} P_\theta(Y_u = 1 \mid T_u = 1) = \theta(x_u) + g_0 \\
&\operatorname{logit} P_\theta(Y_u = 1 \mid T_u = 1) = \theta(x_u) + g_1(x_u) \\
&\operatorname{logit} P_\theta(Y_u = 1 \mid T_u = 1) = \theta(x_u) + g_0 + g_1(x_u).
\end{aligned}$$

Here $g_0 \in \mathbb{R}$, and g_1 is a linear functional $\mathcal{X} \to \mathbb{R}$, so $g_1(0) = 0$. The third version implies that treatment has no effect on the set of units for which $x_u = 0$. This situation is not common, but it does occur if x represents time, and $x = 0$ is the baseline either immediately pre-treatment, or immediately post-treatment before the treatment has had time to take effect. For an illustration, see Example 5.

It is crucial that the space of functions Θ_0 be closed under the group. For example, if $\mathcal{X}$ is a vector space and every $\theta \in \Theta_0$ is a linear functional $\mathcal{X} \to \mathbb{R}$, then $x \mapsto \theta(x) + g$ sends zero to g, which is not a linear functional on $\mathcal{X}$. The standard choices for statistical practice are either the space $\Theta_0 = \mathbf{1}$ of constant functions on $\mathcal{X}$, or the space of affine functions $\mathcal{X} \to \mathbb{R}$, not the space of linear functionals. Other options include the space of inhomogeneous polynomial functions of degree $\leq k$.

12.4 Examples of treatment effects

12.4.1 Simple Gaussian model without interaction

Let $\mathcal{D}_n$ be the space of Gaussian distributions $N_n(\mu, \Sigma)$ indexed by $\mu \in \mathbb{R}^n$ and Σ in the space of positive-definite $n \times n$ matrices. The outcome of randomization is a treatment assignment vector $T \in \mathbb{R}^n$ with components in $\{0, 1\}$. The joint distribution of T is known, and specified in the protocol.

Given $T = \mathbf{t}$, the effect of treatment is an action $\mathcal{D}_n \to \mathcal{D}_n$ on distributions by some group $\mathcal{G}$ of treatment effects. In the simplest case, $\mathcal{G} = \mathbb{R}$, and the action is additive on the mean

$$N_n(\mu, \Sigma) \overset{g}{\longmapsto} N_n(\mu + \mathbf{t}g, \Sigma), \tag{12.1}$$

keeping the covariances fixed.

Consider a standard linear model that is typical of what might be encountered in a scientific experiment, where $i \mapsto x_i$ is the covariate, and $(i, j) \mapsto V_{ij}$ is a non-identity relation that is also positive semi-definite. In the absence of treatment, i.e., if $\mathbf{t} = 0$, the response distribution is some point in the subset $\Theta_0 \subset \mathcal{P}_n$

$$N(X\beta,\, \sigma_0^2 I_n + \sigma_1^2 V)$$

indexed by $\beta \in \mathbb{R}^p$, with two variance components $\sigma_0^2, \sigma_1^2 > 0$. Given $T = \mathbf{t}$, the group action generates an orbit

$$\mathcal{O}(\mathbf{t}) = \{N_n(X\beta + \mathbf{t}g,\, \sigma_0^2 I_n + \sigma_1^2 V) : g \in \mathcal{G}\}$$

consisting of Gaussian distributions indexed by $\beta \in \mathbb{R}^p, g \in \mathbb{R}$, plus $\sigma_0^2, \sigma_1^2 > 0$. Provided that span$(X)$ includes the one-dimensional subspace of constant functions, the complementary treatment vectors $\mathbf{t}$ and $\bar{\mathbf{t}} = 1 - \mathbf{t}$ generate the same orbit.

This is the standard Gaussian model for a treatment effect that is constant and additive for all units, regardless of the covariate value. In general, however, the treatment effect need not be additive on the mean; moreover, if it is additive it need not be the same constant for every unit.

12.4.2 Additive interaction

Loosely speaking, interaction means that the effect of treatment for one unit is not the same as the effect for another unit. In order for this to be the case, we must have $x(u) \neq x(u')$, so the treatment action depends on x. In the simplest setting, x is binary, $\mathcal{G} = \mathbb{R}^2$, and the group action (12.1) becomes

$$N_n(\mu, \Sigma) \overset{g}{\longmapsto} N_n(\mu + \mathbf{t}g_0 + \mathbf{t}\cdot x g_1, \Sigma). \tag{12.2}$$

For units at the reference level such that $x_u = 0$, the treatment effect is an additive increase in the mean by g_0; for units such that $x_u = 1$, the treatment effect is additive by $g_0 + g_1$. The difference g_1 is called the interaction.

In (5.2), the treatment effect is a differential drift, whose magnitude is directly proportional to time-since-baseline. That means that the action of the group element $g \in \mathbb{R}$ is an additive function of the product $g \times \textit{time}$. The effect on the mean is not the same for every unit. Nonetheless, $\mathcal{G} = \mathbb{R}$, so it is not entirely clear whether this should be counted as interaction.

A similar effect can be generated artificially by restriction of (12.2) to the one-dimensional sub-group $g_0 = g_1$.

12.4.3 Dispersion effects

Let $\mathcal{D}_n$ be the space of Gaussian distributions $N_n(\mu, \Sigma)$ on the observation space $\mathbb{R}^n$. We regard $\mathbb{R}^n$ as a Hilbert space with inner product $W = \Sigma^{-1}$. To each subspace $\mathcal{X}$ there corresponds a W-orthogonal projection $P_{\mathcal{X}}$ whose image is $\mathcal{X}$, and complementary orthogonal projection $Q_{\mathcal{X}}$. Both depend on Σ.

Given the treatment assignment vector $T = \mathbf{t}$, the effect of treatment is an action $\mathcal{D}_n \to \mathcal{D}_n$ on Gaussian distributions by the additive group $\mathcal{G} = \mathbb{R}$ as follows:

$$N(\mu, \Sigma) \overset{g}{\longmapsto} N_n\big(\mu,\, Q_{\mathbf{t}}\Sigma + e^g P_{\mathbf{t}}\Sigma\big). \tag{12.3}$$

In this setting, $P_{\mathbf{t}}$ is the W-orthogonal projection onto the subspace span$(\mathbf{1}_n, \mathbf{t})$, or, equivalently, span$(\mathbf{t}, \bar{\mathbf{t}})$, which is ordinarily two-dimensional. The action on the parameter space leaves μ fixed, and sends Σ to $g(\Sigma) = Q_{\mathbf{t}}\Sigma + e^g P_{\mathbf{t}}\Sigma$ For fixed $\mathbf{t}$, the map $\Sigma \mapsto g(\Sigma)$ is a group action on positive definite matrices in the sense that $g = 0$ is the identity transformation, and composition

$$\mathcal{D}_n \xrightarrow{g} \mathcal{D}_n \xrightarrow{g'} \mathcal{D}_n$$

satisfies $g'(g(\Sigma)) = (g'+g)(\Sigma)$. In general, the orthogonal projection $P_{\mathbf{t}}$ depends on Σ, but it is constant on orbits.

In typical applications involving dispersion effects, the null model has $\Sigma \propto I_n$, and either $\mu \in \mathbf{1}_n$ or $\mu = X\beta$ belonging to span(X). Then $P_t\Sigma$ Thus, treatment has a multiplicative effect on variances, but no effect on means.

It is possible to enlarge the group and to combine this group action with either (12.1) or (12.2).

12.4.4 Binary models with correlation

The maximal parameter space for the standard logistic model with Gaussian random effects, consists of ordered pairs $\theta = (\alpha, \Sigma)$ with $\alpha \in \mathbb{R}^n$ and Σ positive definite of order n. Let $\mathcal{D}_n = \{P_\theta : \theta \in \Theta\}$ be the set of distributions on $\{0,1\}^n$ such that $P_\theta(y)$ is the Gaussian integral

$$P_\theta(y) = \int_{\mathbb{R}^n} \prod_{i=1}^{n} \frac{e^{(\alpha_i+\eta_i)y_i}}{1+e^{\alpha_i+\eta_i}}\, \phi_\Sigma(\eta)\, d\eta,$$

where $\phi_\Sigma(\eta) \propto \exp(-\eta'\Sigma^{-1}\eta/2)$ is the normal density on $\mathbb{R}^n$ with covariance Σ.

Examples of group actions associated with a treatment assignment $\mathbf{t}$ are

$$\begin{aligned}
(\alpha, \Sigma) &\overset{g}{\longmapsto} (\alpha + g\mathbf{t},\ \Sigma),\\
(\alpha, \Sigma) &\overset{g}{\longmapsto} (\alpha,\ Q_{\mathbf{t}}\Sigma + e^{g}P_{\mathbf{t}}\Sigma),\\
(\alpha, \Sigma) &\overset{(g_0,g_1)}{\longmapsto} (\alpha + g_0\mathbf{t},\ Q_{\mathbf{t}}\Sigma + e^{g_1}P_{\mathbf{t}}\Sigma),\\
(\alpha, \Sigma) &\overset{g}{\longmapsto} (\alpha + g\mathbf{t},\ Q_{\mathbf{t}}\Sigma + e^{g}P_{\mathbf{t}}\Sigma).
\end{aligned}$$

The last example shows that not all group actions are equally interesting or practically useful.

12.4.5 Survival models

One further example may help to illustrate the options available for group action on distributions. Let Θ_0 be the set of non-negative measures on $\mathbb{R}^+ = (0, \infty)$. To each $\theta \in \Theta_0$ there corresponds a probability distribution on $\mathcal{S} = \mathbb{R}^+ \cup \{\infty\}$ defined by the survivor function

$$P_\theta(Y > t) = \exp\bigl(-\theta((0,t])\bigr)$$

which implies $P_\theta(Y > 0) = 1$ and $P_\theta(Y = \infty) = e^{-\theta(\mathbb{R}^+)}$. A distribution on $\mathcal{S}$ is called a survival distribution; every distribution P on $\mathbb{R}^+$ is regarded as a survival distribution such that $P(\{\infty\}) = 0$. In this setting, θ is called the hazard measure, and the set of survival distributions is in 1–1 correspondence with hazard measures on $\mathbb{R}^+$.

Hazard multiplication

Consider now the group $\mathcal{G} = \mathbb{R}^+$ of positive scalars acting on Θ_0 by scalar multiplication

$$\theta(dt) \overset{g}{\longmapsto} g \times \theta(dt).$$

Each group element is a treatment effect, which is an invertible transformation $g\colon \Theta_0 \to \Theta_0$, or equivalently $P_\theta \mapsto P_{g\theta}$, by scalar multiplication of the hazard measure. In the absence of covariates, the parameter space is $\Theta_0 \times \mathcal{G}$.

The proportional-hazards model states that each individual has a conditional hazard given treatment, one for $T = 0$ and one for $T = 1$; if $g > 0$ is the treatment effect, the two hazard measures are $\theta_u(dt)$ and $g\theta_u(dt)$. As always, these are subject to exchangeability: $x(u) = x(u')$ implies $\theta_u = \theta_{u'}$.

The group does not act transitively on the parameter space, which means that there is more than one orbit—infinitely many in fact. For example, the subset $\theta(dt) \propto \Lambda(dt)$ consisting of measures having a constant strictly positive density with respect to Lebesgue measure is an orbit, corresponding to the set of exponential distributions. For each real α, the subset $\theta(dt) \propto \Lambda(dt)t^\alpha$ is an orbit, and the union of such orbits for $\alpha > -1$ is the family associated with the set of Weibull distributions. For $\alpha < -1$ each orbit consists of measures that have finite total mass; these do not correspond to Weibull distributions. The distribution P_θ on $\mathcal{S}$ exists, but $P_\theta(\{\infty\}) = e^{-\theta(\mathbb{R}^+)} > 0$ implies that the restriction to $\mathbb{R}^+$ is not a probability distribution.

Temporal dilation

Consider now the group $\mathcal{G} = \mathbb{R}^+$ of positive scalars acting on the space of hazard measures by the usual rules for the transformation of distributions by temporal dilation. For present purposes, dilation means that $(g\theta)(A) = \theta(gA)$ for $A \subset \mathbb{R}^+$. Each group element is a treatment effect, which is an invertible transformation $g\colon \Theta_0 \to \Theta_0$, or equivalently $P_\theta \mapsto P_{g\theta}$, by scalar dilation, either of the hazard measure or the distribution itself.

The accelerated-failure model states that each individual has a conditional hazard given treatment, one for $T = 0$ and one for $T = 1$; if $g > 0$ is the treatment effect, the two hazard densities are $\theta'_u(t)$ and $g\theta'_u(gt)$. As always, these are subject to exchangeability: $x(u) = x(u')$ implies $\theta_u = \theta_{u'}$.

Non-constant hazard multiplication

The group actions illustrated above are the ones most commonly encountered in survival analysis. It is evident that there are many other possibilities for group action, most of which have limited potential for applied work, either because they are implausible in one way or another, or because they lead to intractable computations. Nonetheless, it may be helpful to describe a few. In the first two examples, the group is $\mathbb{R}^2$ with addition, and the action on hazards is

Table 12.3. Seven examples of class-plus-treatment survival models

	Male		Female			
	C	T	C	T	$\mathcal{G}$	Eqv?
(i)	θ	$g\theta$	θ	$g\theta$	$\mathbb{R}^+$	✓
(ii)	θ	$g\theta$	θ	$g^{-1}\theta$	$\mathbb{R}^+$	
(iii)	θ	$g_1\theta$	θ	$g_2\theta$	$\mathbb{R}^+ \times \mathbb{R}^+$	
(iv)	θ_1	$g\theta_1$	θ_2	$g\theta_2$	$\mathbb{R}^+$	✓
(v)	θ_1	$g\theta_1$	θ_2	$2g\theta_2$	$\mathbb{R}^+$	NA
(vi)	θ_1	$g_1\theta_1$	θ_2	$g_2\theta_2$	$\mathbb{R}^+ \times \mathbb{R}^+$	✓
(vii)	θ_1	$g_1\theta_1$	θ_2	$g_1g_2\theta_2$	$\mathbb{R}^+ \times \mathbb{R}^+$	✓

multiplicative but not constant

$$\theta(dt) \overset{g}{\longmapsto} e^{g_1+g_2 t}\,\theta(dt)$$
$$\theta(dt) \overset{g}{\longmapsto} e^{g_1+g_2\log(t)}\,\theta(dt).$$

Exponentiation is used to convert the additive group $\mathbb{R}$ or $\mathbb{R}^2$ into the multiplicative group $\mathbb{R}^+$ or $\mathbb{R}^+ \times \mathbb{R}^+$.

There are numerous variations on this theme in which Θ_0 is replaced with some subset that is closed under the group. Ordinarily, the group action should be chosen to be compatible with temporal dilation.

Classification factor plus treatment

Let x be a binary classification factor such as sex. Exchangeability plus independence implies that the values are independent and identically distributed for each sex. Suppose that the hazard measure for males in the control group is a point $\theta \in \Theta_M$. The effect of treatment on males is an action $\Theta_M \to \Theta_M$ in which the group element $g \in \mathcal{G}$ sends P_θ to $P_{g\theta}$. Thus, θ and $g\theta$ both belong to Θ_M. The first two columns of Table 12.3 illustrate this action for males, while columns 3–4 illustrate a similar action for females. In each case, $\Theta_M = \Theta_F$ is the same set of hazard measures, which is closed under scalar multiplication.

Example (v) is not a group action or group homomorphism because the group identity $g = 1$ is not associated with the identity map $\Theta \to \Theta$ on hazard measures. The set of treatment effects, i.e., the set of maps $(\theta_1, \theta_2) \mapsto (g\theta_1, 2g\theta_2)$ in (v), does not have a null element or identity map corresponding to no effect. This is strictly forbidden. Example (ii) is quirky, but it is a group action, and it is equi-variant.

Equivariance for males and females implies not only that $\Theta_M = \Theta_F$, but also that the effect of treatment is the same action either by the same group or by a second copy of the same group. If the treatment effect is an action by the same group element, we say that there is no interaction. Otherwise, the effect is sex-dependent.

In the case of the proportional-hazards model, the two hazard measures for males are θ and $g\theta$. In the absence of interaction, $\mathcal{G} = \mathbb{R}^+$ and the two hazards

for females are $\theta', g\theta'$ with the same proportionality constant. If interaction is present, the group $\mathcal{G} = \mathbb{R}^+ \times \mathbb{R}^+$ consists of pairs (g, g') in which g is the hazard multiplier for males and g' is the multiplier for females.

12.5 Exercises

12.1 Let $\mathbf{t}$ be the treatment assignment vector, and let $B_{\mathbf{t}}$ be the associated block factor, i.e., $B_{\mathbf{t}}(i,j) = 1$ if $t_i = t_j$ and zero otherwise. For $g \in \mathbb{R}$, consider the transformations

$$\Sigma \overset{g}{\longmapsto} \Sigma + g^2 B_{\mathbf{t}}$$

for Σ in the space of positive definite matrices. Discuss whether these transforms determine a group action or group homomorphism (preserving identity and composition). If not, is it a semi-group homomorphism in a suitable sense? Maybe after changing g^2 to e^g or $|g|$ to maintain positivity?

12.2 This exercise is concerned with a possible action of the additive group of real numbers on the space of positive definite matrices of order n. Let $\mathcal{X} \subset \mathbb{R}^n$ be a given subspace. To each Σ and $W = \Sigma^{-1}$ there corresponds a W-orthogonal projection P_W whose image is $\mathcal{X}$, and a complementary projection $Q_W = I - P_{\mathcal{X}}$. In matrix notation, $P_W = X(X'WX)^{-1}X'W$ depends on Σ. For $g \in \mathbb{R}$, show that the transformations

$$\Sigma \overset{g}{\longmapsto} Q_W \Sigma + e^g P_W \Sigma = \Sigma + (e^g - 1)X(X'\Sigma^{-1}X)^{-1}X'$$

determine a group homomorphism by linear transformations on the space of positive-definite matrices.

12.3 Let $\mathbf{t}$ be the treatment assignment vector, and let P_W be the W-orthogonal projection onto the subspace span$(\mathbf{1}, \mathbf{t})$. Show that the transformation

$$N_n(\mu, \Sigma) \overset{g}{\longmapsto} N_n(\mu + g_0\mathbf{t},\, Q_{\mathcal{X}}\Sigma + e^{g_1}P_W\Sigma)$$

is an action of the additive group $\mathbb{R}^2$ on the space of Gaussian distributions. Describe the orbit of the distribution $N_n(\mathbf{1}, I_n)$.

12.4 Under what conditions does the treatment model in the preceding exercise satisfy the lack of interference condition?

Chapter 13

Gaussian distributions

13.1 Real Gaussian distribution

13.1.1 Density and moments

The standard Gaussian distribution has a density

$$\Phi(dy) = \phi(y)\,dy = \frac{1}{\sqrt{2\pi}} e^{-y^2/2}\,dy$$

with respect to Lebesgue measure on the real line. It is symmetric with finite moments of all orders. The moment generating function is

$$M_0(t) = \int e^{ty}\phi(y)\,dy = e^{t^2/2} = \sum_{r=0}^{\infty} \mu_r t^r / r!,$$

from which the odd moments are zero, and the even moments are

$$\mu_{2r} = \frac{(2r)!}{2^r\, r!} = 1 \cdot 3 \cdots (2r-1).$$

The cumulant generating function is

$$K_0(t) = \log M_0(t) = \sum \kappa_r t^r / r! = t^2/2,$$

so all of the cumulants are zero except for the variance $\kappa_2 = 1$.

If ε is a standard normal variable, and (μ, σ) is any pair of real numbers with $\sigma > 0$, the affine transformation $Y = \mu + \sigma\varepsilon$ is distributed according to the Gaussian distribution with mean μ and variance σ^2. The density function of the transformed variable at y is

$$\frac{1}{\sigma}\phi\Big(\frac{y-\mu}{\sigma}\Big) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(y-\mu)^2/(2\sigma^2)}.$$

The moment and cumulant generating functions are

$$M_{\mu,\sigma}(t) = E(e^{tY}) = e^{t\mu} E(e^{t\sigma\varepsilon}) = e^{t\mu + t^2\sigma^2/2}$$
$$K_{\mu,\sigma}(t) = \log M_{\mu,\sigma}(t) = t\mu + t^2\sigma^2/2.$$

The mean is μ, the variance is $\kappa_2 = \sigma^2$, and all other cumulants are zero.

For $x > 0$, the ratio of the right tail probability $1 - \Phi(x)$ to the density $\phi(x)$ is called Mills's ratio. The asymptotic expansion is

$$\frac{1 - \Phi(x)}{\phi(x)} = \frac{1}{x} - \frac{1}{x^3} + O(x^{-5}).$$

This stands in sharp contrast with heavy-tailed distributions for which the corresponding ratio is increasing in x; in the case of the Cauchy distribution the ratio is asymptotically x.

13.1.2 Gaussian distribution on $\mathbb{R}^n$

Let $X = (X_1, \ldots, X_n)$ be a random vector in $\mathbb{R}^n$ whose components are independent and identically distributed $N(0,1)$ variables. Independence implies that the joint density function with respect to Lebesgue measure at $x \in \mathbb{R}^n$ is the product of the marginal density functions, which is

$$\Phi_n(dx) = \phi_n(x)\, dx = (2\pi)^{-n/2} e^{-\|x\|^2/2}\, dx,$$

where $\|x\|^2 = x_1^2 + \cdots + x_n^2$ is the standard Euclidean squared norm.

This is called the standard normal distribution on $\mathbb{R}^n$, and is denoted by $N_n(0, I_n)$. The joint moment generating function is the product of the marginal generating functions

$$M_0(t) = \int_{\mathbb{R}^n} e^{t_1 x_1 + \cdots + t_n x_n} \phi_n(x)\, dx = e^{\|t\|^2/2},$$

and the cumulant generating function is $\|t\|^2/2$, which is quadratic and radially symmetric as a function of t. All of the joint cumulants are zero except for the variances, which are $\operatorname{cov}(X_i, X_j) = \delta_{ij}$, i.e., one for $i = j$ and zero otherwise.

Let L be a linear transformation $\mathbb{R}^n \to \mathbb{R}^n$, so that the matrix L is of order $n \times n$. The moment generating function of the transformed variable $Y = LX$ is

$$E\big(e^{t'Y}\big) = \int_{\mathbb{R}^n} e^{t'Lx} \phi_n(x)\, dx = M_0(L't) = e^{\|L't\|^2/2},$$

so the cumulant generating function $\|L't\|^2/2 = t'LL't/2$ is quadratic in t but not radially symmetric. All of the cumulants are zero except for the variances and covariances, which are the components of the matrix

$$\Sigma = \operatorname{cov}(Y) = \operatorname{cov}(LX) = LL',$$

which is symmetric and positive semi-definite. The random variable Y has the normal distribution in $\mathbb{R}^n$ with mean zero and covariance Σ, which is denoted by $N_n(0, \Sigma)$.

If L is invertible, the covariance matrix $\Sigma = LL'$ is also invertible with inverse $W = L'^{-1}L^{-1}$. In that case, the Jacobian of the transformation is the absolute value of the determinant of the transform matrix

$$dy = |\det(L)|\, dx = \det^{1/2}(\Sigma)\, dx.$$

The joint density of the transformed variable is

$$(2\pi)^{-n/2}|W|^{1/2}e^{-y'Wy/2}\, dy, \tag{13.1}$$

which is the density at y of the Gaussian distribution $N_n(0, \Sigma)$.

In general, the linear transformation L need not be invertible. In that case the subspaces $\mathrm{Im}(L) = \mathrm{Im}(\Sigma)$ and $\ker(L') = \ker(\Sigma)$ are complementary of dimensions $n - k$ and k respectively, and are also orthogonal with respect to the standard inner product in $\mathbb{R}^n$. With probability one, $Y = LX$ belongs to $\mathrm{Im}(\Sigma)$, so the distribution $N_n(0, \Sigma)$ necessarily puts mass one on this subspace. If $k > 0$, the distribution $N_n(0, \Sigma)$ is singular and does not have a density with respect to Lebesgue measure on $\mathbb{R}^n$.

The translation $Y \mapsto Y + \mu$ sends the distribution $N_n(0, \Sigma)$ to $N_n(\mu, \Sigma)$, which is supported on the coset, or affine subspace, $\mu + \mathrm{Im}(\Sigma)$. The cumulant generating function is $t'\mu + t'\Sigma t/2$, so the mean vector is μ and the covariance matrix is Σ. If Σ is invertible, then $\mathrm{Im}(\Sigma) = \mathbb{R}^n$, and the distribution has a density

$$(2\pi)^{-n/2}|W|^{1/2}e^{-(y-\mu)'W(y-\mu)/2}\, dy. \tag{13.2}$$

13.2 Complex Gaussian distribution

13.2.1 One-dimensional distribution

The one-dimensional Gaussian distribution on the complex plane is nothing more than a two-dimensional Gaussian distribution on $\mathbb{R}^2$ that is also rotationally symmetric. The zero-mean complex Gaussian distribution with variance σ^2 has a density

$$\phi(z) = \frac{e^{-|z|^2/\sigma^2}}{\pi\sigma^2}$$

with respect to two-dimensional Lebesgue measure. The real part and the imaginary part of $Z \sim \mathbb{C}N(0, 1)$ are independent zero-mean real gaussian variables with variance $\sigma^2/2$ each. The argument of Z is uniformly distributed on $[0, 2\pi)$, and independent of $\|Z\|^2$, which is exponentially distributed with mean σ^2.

Rotational symmetry implies not only that $E(Z) = 0$ but also that complex powers satisfy $E(Z^k) = 0 = E(\bar{Z}^k)$ for every integer $k \geq 1$. The only non-zero integer moments are $E(\|Z\|^{2k}) = k!\sigma^{2k}$ in which Z and $\bar{Z}$ occur an equal number of times in the product. The kth order cumulant is $\mathrm{cum}_k(\|Z\|^2) = (k-1)!\sigma^{2k}$.

13.2.2 Gaussian distribution on $\mathbb{C}^n$

Let $\varepsilon_1, \ldots, \varepsilon_n$ be independent and identically distributed $\mathbb{C}N(0,1)$ random variables, so that the joint density is

$$\pi^{-n} \prod_{r=1}^{n} e^{-|\varepsilon_r|^2} = \pi^{-n} e^{-\varepsilon^* \varepsilon} = \pi^{-n} e^{\|\varepsilon\|^2}$$

with respect to $2n$-dimensional Lebesgue measure. Let $Z = L\varepsilon$, where L is a full-rank complex matrix of order n, and let $\Sigma = LL^*$ be positive-definite Hermitian. The derivative matrix of the linear transformation $L\colon \mathbb{C}^n \to \mathbb{C}^n$ is L, but the Jacobian of the linear transformation $\mathbb{R}^{2n} \to \mathbb{R}^{2n}$ is $\det(LL^*) = \det(\Sigma)$. Thus, the density of the transformed vector is

$$\pi^{-n} \det(\Sigma)^{-1} e^{-z^* \Sigma^{-1} z}$$

with respect to $2n$-dimensional Lebesgue measure at $z \in \mathbb{C}^n$. This distribution is denoted by $Z \sim \mathbb{C}N_n(0, \Sigma)$, where $E(ZZ^*) = E(L\varepsilon\varepsilon^* L^*) = LL^* = \Sigma$.

As a reminder, Hermitian symmetry means that the real part of Σ is symmetric, and the imaginary part is anti-symmetric or skew-symmetric. Thus the conjugate is equal to the transpose $\bar{\Sigma} = \Sigma'$, while $\bar{\Sigma}' = \Sigma^* = \Sigma$. Strict positive definiteness means that every Hermitian quadratic form $\xi^* \Sigma \xi$ in complex vectors is strictly positive unless $\xi = 0$. If Σ is strictly positive definite, so also is the complex conjugate matrix $\bar{\Sigma}$, and the real part $\Re(\Sigma)$.

The conjugate vector is distributed as $\mathbb{C}N(0, \bar{\Sigma})$, and the unit complex multiple $e^{i\theta} Z$ has the same distribution as Z. The one-dimensional marginal distribution of Z_1 is complex Gaussian with variance Σ_{11}, and the marginal distribution of $Z_{i_1}, \ldots, Z_{i_k}$ is complex Gaussian with covariance $\Sigma[\mathbf{i}, \mathbf{i}]$ restricted to rows $\mathbf{i} = \{i_1, \ldots, i_k\}$, and the same columns. Note that the restriction is applied to the rows and columns of Σ, not to the rows and columns of the precision matrix Σ^{-1}.

Exercises 13.1–13.3 show that the Hermitian matrix $\Sigma = \Sigma_0 + i\Sigma_1$ can be associated with a $2n \times 2n$ real symmetric matrix in such a way that the pair of real vectors $\Re(Z), \Im(Z)$ is jointly Gaussian with covariance

$$\operatorname{cov} \begin{pmatrix} \Re(Z) \\ \Im(Z) \end{pmatrix} = \begin{pmatrix} \Sigma_0 & \Sigma_1 \\ \Sigma_1' & \Sigma_0 \end{pmatrix}.$$

where $\Sigma_1' = -\Sigma_1$. Any pair of identically distributed real Gaussian vectors X, Y defines a complex Gaussian vector $Z = X + iY$ if and only if the cross-covariances are anti-symmetric, $\operatorname{cov}(X, Y) = -\operatorname{cov}(Y, X)$.

13.2.3 Moments

Rotational symmetry with respect to complex unit scalar multiplication means that $E(Z_r) = E(e^{i\theta} Z_r)$ is necessarily zero. Likewise the product moment $E(Z_r Z_s) = E(e^{2i\theta} Z_r Z_s)$ is also zero. The only non-zero second-order moments

are $\mathrm{cov}(Z_r, \bar{Z}_s) = \Sigma_{rs}$. More generally, the only non-zero moments of degree $2k$ are those in which conjugated and non-conjugated components occur in equal number, such as the product $Z_{i_1} \cdots Z_{i_k} \bar{Z}_{j_1} \cdots \bar{Z}_{j_k}$.

The evaluation of Gaussian moments is a classical problem dating back to Isserlis (1918), in the case of real vectors. In the case of complex Gaussian vectors, the product moment is related to Wick's theorem (Wick, 1950), to Boson point processes (McCullagh and Møller, 2006), and to Feynman diagrams. The complex case is a little simpler than the real case, and the product moment is as follows. To each permutation $\pi\colon [k] \to [k]$ there corresponds a 1–1 matching $i_r \mapsto j_{\pi(r)}$ of conjugated with non-conjugated components. Each matching gives rise to a product of k covariances

$$E(Z_{i_1} \cdots Z_{i_k} \bar{Z}_{j_1} \cdots \bar{Z}_{j_k}) = \sum_{\pi} \prod_{r=1}^{k} \Sigma_{i_r, j_{\pi(r)}} = \mathrm{per}(\Sigma[\mathbf{i}, \mathbf{j}]),$$

which is the permanent of the $\mathbf{i} \times \mathbf{j}$ sub-matrix. Note that rows or columns may be repeated, so that $E(|Z_1|^{2k}) = \Sigma_{11}^k\, k!$, which are the moments of the exponential distribution. The permanent is the same as the determinant except that all $k!$ terms in the permutation expansion have coefficient $+1$.

Complex-valued random variables seldom occur in experimental research except in the setting of Fourier transformation for time series, as in Example 6. They are not used in the remainder of this chapter, but they do also arise in connection with stationary Gaussian processes, particularly space-time processes in chapter 14.

13.3 Gaussian Hilbert space

13.3.1 Euclidean structure

It is often convenient to associate with the Gaussian distribution $N_n(0, \Sigma)$ or $N_n(\mu, \Sigma)$ a vector space having very specific geometric properties that match the second moments of the distribution. In doing so, the mean vector is ignored, so 'second moments' refers to variances and covariances. For simplicity, we assume that $\Sigma = W^{-1}$ is invertible, so the domain or support of the distribution is the entire vector space $\mathbb{R}^n$. The Euclidean geometric properties (length, angle, orthogonality,...) are generated by the specific inner product $\langle x, y\rangle = \sum w_{ij} x_i y_j$ matching the norm in the exponent of the density (13.1) or (13.2). This inner-product space $\mathcal{H} = (\mathbb{R}^n, \langle \cdot, \cdot \rangle)$ is called the Gaussian Hilbert space. Apart from minor modifications of notation, the algebra for complex Gaussian spaces is essentially the same as that for real vector spaces, so the notation here uses real vector spaces.

The geometry associated with $\mathcal{H}$ is a special case of the geometry associated with the Rao-Fisher-information metric generated by a parametric model. In particular, the linear model $Y \sim N_n(X\beta, \Sigma)$ determines a subspace $\mathcal{X} \subset \mathcal{H}$, and a linear transformation $Y \mapsto \hat{\beta}$ that sends Y to the weighted least-squares

coefficient vector $\hat{\beta} = (X'WX)^{-1}X'WY$ in $\mathbb{R}^p$. This linear transformation also sends the distribution $N_n(X\beta, \Sigma)$ on $\mathbb{R}^n$ to $N_p\big(\beta, (X'WX)^{-1}\big)$, so it is a transformation $\mathcal{H} \to \mathcal{H}_p$, where $\mathcal{H}_p$ is p-dimensional Euclidean space with inner product matrix $X'WX$. The transformation $\mathcal{H}_p \to \mathcal{X} \subset \mathcal{H}$ that sends $\hat{\beta}$ to $\hat{\mu} = X\hat{\beta}$ is in fact a Euclidean isometry, so the geometric and distributional properties of $\hat{\beta} \in \mathcal{H}_p$ mirror exactly those of the orthogonal projection $\hat{\mu} = PY \in \mathcal{X} \subset \mathcal{H}$.

The main reasons for endowing the domain with Euclidean structure are as follows:

1. Orthogonality of subspaces is associated with independence of random variables;

2. The orthogonal projection having a given image is associated with maximum-likelihoodF and weighted least squares;

3. The orthogonal projection having a given kernel is associated with a number of statistically distinct operations such as least-squares residual, prediction, interpolation, smoothing and Kriging;

4. Cochran's theorem and much of the distribution-theory associated with linear regression and analysis of variance become more transparent.

13.3.2 Cautionary remarks

From the vantage of linear algebra, it is natural to specify the inner product directly through the inner-product matrix W, which is symmetric and strictly positive definite. The inner product in the dual space of linear functionals is the matrix inverse, $\Sigma = W^{-1}$.

The order of operations in statistical work is ordinarily reversed. A Gaussian process is defined by its covariance function, which is naturally subject to restrictions such as stationarity, isotropy, or exchangeability, depending on the structure of its domain. Consequently, the matrix Σ, which is the restriction of the covariance function to the sample points, is specified first. The inverse matrix then determines the inner product in $\mathcal{H}$ *for the particular sample selected.*

For a process sampled at points $u_1, \ldots, u_n$ in some domain $\mathcal{U}$, the matrix component $\Sigma_{ij} = \mathrm{cov}(Y(u_i), Y(u_j))$ depends on u_i, u_j only, and is independent of the configuration of the remaining sample points. By contrast w_{ij} depends on the entire configuration of sampled points. For example, if the process is stationary on the plane, $\mathcal{U} = \mathbb{R}^2$, and $u_i - u_j = u_{i'} - u_{j'}$ implies $\Sigma_{ij} = \Sigma_{i'j'}$. But this does not imply $w_{ij} = w_{i'j'}$.

Despite the substantial advantages listed in the preceding section, it is good to be aware of one additional limitation of associating a specific geometry with the Gaussian distribution. In statistical work it is often necessary to compare two candidate distributions on the same observation space, for example by computing the likelihood ratio. For example, the candidate distributions might be $N_n(0, \Sigma_0)$ and $N_n(0, \Sigma_1)$ for two given matrices. To compute a likelihood ratio,

it is essential to compare candidate distributions on the same space, so it could be a serious mistake to associate with each distribution its own geometry.

13.3.3 Projections

Specification by image

Let X be any matrix of order $n \times p$ whose columns span the subspace $\mathcal{X} \subset \mathcal{H}$ of dimension p. The transformation $P\colon \mathcal{H} \to \mathcal{H}$ whose matrix representation is

$$P = X(X'WX)^{-1}X'W \tag{13.3}$$

has the following properties.

1. $P^2 = P$;
2. For each $x \in \mathcal{H}$, Px belongs to $\mathcal{X}$;
3. For each $x \in \mathcal{X}$, $Px = x$;
4. For each $x, y \in \mathcal{H}$, $\langle x,\, Py\rangle = \langle Px,\, y\rangle$.

The first of these, called idempotence, is the definition of a projection, orthogonal or otherwise. The second says that the image of P is a subspace of $\mathcal{X}$; The third says that P acts as the identity on $\mathcal{X}$, so $\mathrm{Im}(P)$ contains $\mathcal{X}$; the second and third together imply $\mathrm{Im}(P) = \mathcal{X}$. The fourth is the self-adjointness condition, which implies that $\mathrm{Im}(P)$ and $\ker(P)$ are orthogonal subspaces in $\mathcal{H}$. It follows that P is the orthogonal projection $\mathcal{H} \to \mathcal{H}$ whose image is $\mathcal{X}$, and the complementary transformation $Q = I_n - P$ is the orthogonal projection whose kernel is $\mathcal{X}$.

Properties 1–3 hold for any strictly positive definite matrix W whether or not it coincides with the inner product in $\mathcal{H}$. In particular, $P_0 = X(X'X)^{-1}X'$ is a projection with image $\mathcal{X}$, and $Q_0 = I_n - P_0$ is the complementary projection with kernel $\mathcal{X}$, but neither projection is orthogonal in $\mathcal{H}$.

If L is $p \times p$ of full rank, then the matrices X and XL span the same space. If we replace X with XL in the definition of P, we obtain the same projection; likewise for P_0. In other words, P and P_0 are independent of the vectors selected to span the image subspace.

These are the most familiar versions of projection matrices that arise in statistical work, where the projection is usually targeted to have a particular image. But it is occasionally convenient to specify a projection directly by a linear transformation matrix having the desired kernel.

Specification by kernel

Let $\mathcal{K} \subset \mathcal{H}$ be a given subspace of dimension k, and let $K\colon \mathcal{H} \to \mathbb{R}^{n-p}$ be any matrix of order $n - k \times n$ with kernel $\mathcal{K}$. Then the matrix

$$Q^\dagger = \Sigma K'(K\Sigma K')^{-1}K \tag{13.4}$$

satisfies $Q^\dagger Q^\dagger = Q^\dagger$, so $Q^\dagger$ is a projection $\mathcal{H} \to \mathcal{H}$. It is easily verified that $\ker(Q^\dagger) = \mathcal{K}$. Symmetry of $WQ^\dagger$ implies self-adjointness, so $Q^\dagger$ is the orthogonal projection with kernel $\mathcal{K}$. In particular, if we choose the matrix K so that $\mathcal{K} = \mathcal{X}$, uniqueness implies that $Q^\dagger$ coincides with $Q = I_n - P$ as defined in (13.3). This identity is by no means obvious from the matrix algebra alone.

Self-adjointness identity

Let P be any orthogonal projection $\mathcal{H} \to \mathcal{H}$ such as (13.3) or (13.4). Idempotence implies $P^2 = P$, and self-adjointness implies that $WP = P'W$ is a symmetric matrix. It follows that $P'WP = WP = P'W$ is symmetric and positive semi-definite.

Mixed products

By definition, two projections such that $\mathrm{Im}(P_0) \subseteq \mathrm{Im}(P_1)$ satisfy $P_1P_0 = P_0$. If both projections have the same image, then

$$P_1P_0 = P_0; \quad P_0P_1 = P_1;$$

i.e., the first or rightmost projection prevails in the product. Mixed products having nested kernels exhibit the opposite behaviour; $\ker(Q_0) \subseteq \ker(Q_1)$ implies $Q_1Q_0 = Q_1$ in which the last, or leftmost, projection prevails.

In statistical work related to linear models, two linear transformations T, T' having the same kernel are statistically equivalent in the sense that there exists linear transformations L, L' such that $T = LT'$ and $T' = L'T$. One can be obtained from the other by a linear transformation; for projections, $L = T$ and $L' = T'$.

Trace and rank

Let P be any linear projection, not necessarily an orthogonal projection $\mathcal{H} \to \mathcal{H}$. The idempotence condition $P^2 = P$ means that the eigenvalues of P satisfy $\lambda^2 = \lambda$, which implies that λ is either zero or one. Consequently, the trace of P, which is the sum of the eigenvalues, is equal to the rank of P or the dimension of the image space.

Rank degeneracy

Suppose that $Y \sim N_n(0, \Sigma)$, where Σ has rank $n-p$. Let $K\colon \mathbb{R}^n \to \mathbb{R}^{n-p}$ be any linear transformation whose kernel coincides with the kernel of Σ, i.e,. $\ker(K) = \ker(\Sigma) = \mathcal{X}$. This means that K is a matrix of order $n-p \times n$ and rank $n-p$. With respect to the standard inner product in $\mathbb{R}^n$, $\mathrm{Im}(K')$ is complementary and orthogonal to $\ker(K)$. Symmetry of Σ implies $\mathrm{Im}(K') = \mathrm{Im}(\Sigma)$.

The relevant Gaussian Hilbert space associated with $N_n(0, \Sigma)$ is either the subspace $\mathrm{Im}(\Sigma)$, or the quotient space $\mathbb{R}^n/\mathcal{X}$. In either case, the dimension

is $n - p$. For either representation of $\mathcal{H}$, the inner product is a positive semi-definite quadratic form $\langle x, y\rangle = x'Wy$ in for $x, y \in \mathbb{R}^n$, where W is the $n \times n$ symmetric matrix

$$W = K'(K\Sigma K')^{-1}K, \tag{13.5}$$

which has the same image and kernel as Σ. Evidently, $W\Sigma K' = K'$ so $W\Sigma$ is the identity on $\mathrm{Im}(K') = \mathrm{Im}(\Sigma)$, and $W\Sigma$ is zero on $\ker(\Sigma)$, so $W\Sigma$ is a projection. Symmetry of $W\Sigma W = W$ implies that $W\Sigma$ is self-adjoint, so $W\Sigma$ is the orthogonal projection $\mathcal{H} \to \mathcal{H}$ whose image is $\mathrm{Im}(\Sigma)$, i.e., $W\Sigma$ is the identity $\mathcal{H} \to \mathcal{H}$.

The preceding algebra has another interpretation that is related to incomplete Gaussian distributions in which $N_n(0, \Sigma; \mathcal{X})$ is a Gaussian distribution on the quotient space $\mathbb{R}^n/\mathcal{X}$. In other words, $N_n(0, \Sigma; \mathcal{X})$ is the restriction of $N_n(0, \Sigma)$ to Borel subsets $A \subset \mathbb{R}^n$ such that $A + \mathcal{X} = A$. In this situation, it is necessary only that Σ be positive definite on $\mathcal{X}$-contrasts, which means that $K\Sigma K'$ is strictly positive definite. Two covariance matrices such that $K\Sigma_1 K' = K\Sigma_2 K'$ are equivalent on $\mathcal{X}$-contrasts, and determine the same distribution on $\mathbb{R}^n/\mathcal{X}$. In that case, the matrix W in (13.5) serves as the inner product in $\mathcal{H}$.

For a simple example of the latter, floating Brownian motion is a generalized Gaussian process on the real line whose covariance function is $-|u - u'|$. For any collection of points $u_1, \ldots, u_n$ in $\mathbb{R}$, the matrix whose components are $\Sigma_{ij} = -|u_i - u_j|$ is symmetric but clearly not positive definite or even semi-definite. However, if we take $\mathcal{X} = \mathbf{1}$, the subspace of constant functions, and $\ker(K) = \mathbf{1}$, it can be shown that $K\Sigma K'$ is positive definite. We say that $-|u-u'|$ is *positive-definite on simple contrasts.*

The Dirac difference measure $\delta_u(\cdot) - \delta_{u'}(\cdot)$ is an example of an elementary contrast, and the process takes a value $Y(\delta_u - \delta_{u'})$, conventionally written as an increment $Y(u) - Y(u')$, which is distributed as Gaussian with variance

$$(1, -1)\begin{pmatrix} 0 & -|u-u'| \\ -|u-u'| & 0 \end{pmatrix}\begin{pmatrix} 1 \\ -1 \end{pmatrix} = 2|u - u'|.$$

Despite the notation, $Y(u)$ or $Y(\delta_u)$ in isolation is not a Gaussian variable with finite variance. Provided that the phrase is understood informally as a limit, it is seldom misleading to regard $Y(u)$ as Gaussian with 'infinite' variance. Floating Brownian motion is not defined pointwise, but it is stationary on its domain of contrasts. Standard Brownian motion

$$B(u) = Y(u) - Y(0) \sim N(0, 2|u|)$$

is defined pointwise for $u \in \mathbb{R}$, but is not stationary. Realizations of either process are everywhere continuous but nowhere differentiable.

13.3.4 Dual space of linear combinations

Let $Y = (Y_1, \ldots, Y_n)$ be a random vector distributed as $N_n(0, \Sigma)$ on $\mathbb{R}^n$, where Σ is invertible. To each coefficient vector $\alpha = (\alpha_1, \ldots, \alpha_n)$ there corresponds a

linear combination

$$Y(\alpha) = \alpha_1 Y_1 + \cdots + \alpha_n Y_n.$$

Instead of indexing Y by the points $i \in [n]$, the preceding notation suggests that we use the space of linear combinations as an extended index set. Strictly speaking, this extension is unnecessary and superfluous. As a linear functional, the extension $Y(3\alpha + 4\beta) = 3Y(\alpha) + 4Y(\beta)$ is linear and additive, so all values are determined by the values on any basis.

The covariance of two linear combinations is bilinear:

$$\operatorname{cov}\big(Y(\alpha), Y(\beta)\big) = \langle \alpha, \beta \rangle = \sum \alpha_i \beta_j \Sigma_{ij}.$$

The Hilbert space $\mathcal{H}^*$ consisting of coefficient vectors, or linear functionals, with this inner product is the dual of $\mathcal{H}$. By definition, it is restricted to coefficient vectors α such that the linear combination $Y(\alpha)$ has finite variance $\|\alpha\|^2 < \infty$. The dual space arises most prominently in problems of prediction and computation of conditional distributions for spatial and temporal processes.

An observation on the process consists of a finite sample $\{x_1, \ldots, x_n\}$ of sites plus the site values $Y(x_1), \ldots, Y(x_n)$. The observation values serve as basis elements in the observation space $\mathcal{H}$, while the sample points serve as basis elements for the dual space of linear combinations $\alpha \in \mathcal{H}_0^*$. As a process, the space of samples is embedded in a larger Hilbert space $\mathcal{H}^* \supset \mathcal{H}_0^*$ associated with extended samples. For any $\beta \in \mathcal{H}^*$, the conditional distribution of $Y(\beta)$ given the sample values $\{Y(\alpha) : \alpha \in \mathcal{H}_0^*\}$ is Gaussian with moments

$$Y(\beta) \mid Y[\mathcal{H}_0^*] \sim N\big(Y(P\beta),\, \|Q\beta\|\big) \tag{13.6}$$

where P is the orthogonal projection $\mathcal{H}^* \to \mathcal{H}^*$ with image $\mathcal{H}_0^*$, and Q is the complementary projection.

13.4 Statistical interpretations

13.4.1 Canonical norm

In this section, $\mathcal{H}$ is the Hilbert space associated with the distribution $N_n(0, \Sigma)$. For simplicity of exposition, $W = \Sigma^{-1}$ is invertible and $\dim(\mathcal{H}) = n$.

The squared norm of a vector $x \in \mathcal{H}$ is $\|x\|^2 = x'Wx$. For $Y \sim N_0, \Sigma)$, the distribution of the scalar random variable $\|Y\|^2$, can be obtained from its moment generating function

$$\begin{aligned} E\big(e^{t\|Y\|^2}\big) &= (2\pi)^{-n/2} |W|^{1/2} \int_{\mathcal{H}} e^{ty'Wy - y'Wy/2}\, dy \\ &= (2\pi)^{-n/2} |W|^{1/2} \int_{\mathcal{H}} e^{(1-2t)\|y\|^2/2}\, dy = (1-2t)^{-n/2} \end{aligned}$$

provided that $t < 1/2$. The moment generating function of the χ_1^2-distribution is $(1-2t)^{-1/2}$, so $Y'WY$ is distributed as χ_n^2, which is the distribution of the sum $Z_1^2 + \cdots + Z_n^2$ of squares of n independent standard Gaussian variables.

The χ^2 density function is available in closed form, but is not especially important for either theory or applications. The cumulant generating function $-n\log(1-2t)/2$ implies that the rth cumulant is $\kappa_r = n\,(r-1)!2^{r-1}$. All cumulants are proportional to n, the mean and variance are n and $2n$, and the central limit theorem implies $\chi^2_n \simeq N(n, 2n)$ for large n. For numerical work, the cumulative distribution function is available in R using the syntax `pchisq(x, df=n)`, and simulated variables are available using `rchisq(..., df=n)`.

13.4.2 Independence

Two orthogonal projections $P, Q\colon \mathcal{H} \to \mathcal{H}$ are said to be *mutually orthogonal* if $PQ = QP = 0$, so the projections (13.3) and (13.4) are both complementary and mutually orthogonal. Orthogonality of projections implies that the random vectors PY, QY are independent. This can be verified directly from the matrix forms (13.3) or (13.4), which satisfy

$$P\Sigma P' = P\Sigma; \quad Q\Sigma Q' = Q\Sigma; \quad \text{and} \quad P\Sigma Q' = PQ\Sigma = 0.$$

More directly, the joint moment generating function

$$E\big(e^{t_1'PY + t_2'QY}\big) = e^{(t_1'P + t_2'Q)\Sigma(P't_1 + Q't_2)/2} = e^{t_1'P\Sigma P't_1 + t_2'Q\Sigma Q't_2}$$

is the product of the marginal generating functions.

Cochran's theorem

Let $P_1, \ldots, P_k$ be orthogonal projections $\mathcal{H} \to \mathcal{H}$ that are (i) mutually orthogonal in the sense $P_rP_s = 0$ for $r \neq s$, and (ii) complementary in the sense $P_1 + \cdots + P_k = I_n$. Since the rank and trace are equal, complementarity implies $n_1 + \cdots + n_k = n$, where $n_r = \mathrm{tr}(P_r) = \mathrm{rank}(P_r)$. Then, for every $Y \in \mathcal{H}$, we have the linear and Pythagorean identities

$$\begin{aligned} Y &= P_1Y + \cdots + P_kY; \\ \|Y\|^2 &= \|P_1Y\|^2 + \cdots + \|P_kY\|^2. \end{aligned}$$

By an obvious extension of the argument given above for $Y \sim N_n(0, \sigma^2 V)$, the projected random variables $P_rY \sim N(0, \sigma^2 P_rV)$ are mutually independent, and $\|P_rY\|^2 \sim \sigma^2\chi^2_{n_r}$ are also mutually independent. In the iterature on analysis of variance, this distributional decomposition is known as Cochran's theorem, or the Fisher-Cochran theorem.

For $r \neq s$, the ratio of mean squares

$$\frac{\|P_rY\|^2/n_r}{\|P_sY\|^2/n_s}$$

is distributed independently of σ according to Fisher's F distribution F_{n_r,n_s}.

The decomposition can be stated in an alternative way in terms of a sequence of strictly nested subspaces

$$\mathbf{0} = \mathcal{X}_0 \subset \mathcal{X}_1 \subset \cdots \subset \mathcal{X}_k \subset \mathcal{X}_{k+1} = \mathcal{H}$$

of dimensions $0 < n_1 < n_2 < n_k < n$. Let P_r be the orthogonal projection onto $\mathcal{X}_r$ so that $P_r P_s = P_{r\wedge s}$, and $Q_r Q_s = Q_{r\vee s}$ for the complementary projections. Then the increments $(\Delta P)_r = P_r - P_{r-1} = Q_{r-1} - Q_r$ are mutually orthogonal projections satisfying the conditions for Cochran's theorem. In particular, if $Y \sim N(X\beta, \sigma^2 V)$ satisfies the standard linear model assumption with non-zero mean such that $\mathcal{X}_1 = \text{span}(X)$, and $\mathcal{X}_2 = \text{span}(X, Z)$ is any proper subspace of $\mathcal{H}$ containing $\mathcal{X}$ as a proper subspace, then

$$\|Q_1 Y\|^2 = \|Q_1 - Q_2)Y\|^2 + \|Q_2 Y\|^2$$

is a decomposition of the residual sum of squares into independent $\sigma^2\chi^2$ components. Consequently, the ratio of mean squares

$$F = \frac{\|(Q_1 - Q_2)Y\|^2/(n_2 - n_1)}{\|Q_2 Y\|^2/(n - n_2)}$$

is distributed according to the F distribution.

13.4.3 Prediction and conditional expectation

Let $Y \sim N_n(0, \Sigma)$ on $\mathbb{R}^n$, and let $Z = KY$ be any linear transformation whose kernel is $\mathcal{K}$. Then the conditional expectation given Z is $E(Y \mid Z) = QY$, where Q is the orthogonal projection $\mathcal{H} \to \mathcal{H}$ whose kernel is $\mathcal{K}$. If K has full rank, the matrix form (13.4) makes it clear that the conditional expected value

$$QY = \Sigma K'(K\Sigma K')^{-1} KY = \Sigma K'(K\Sigma K')^{-1} Z$$

is indeed a function of the observation Z.

If we write $Y = PY + QY$ as the sum of complementary orthogonal projections, the proof is trivial because $KQ = K$ and $Z = KQY$ is independent of PY. Consequently,

$$\begin{aligned} E(Y \mid Z) &= E(PY + QY \mid KQY) = QY; \\ \text{cov}(Y \mid Z) &= \text{cov}(PY + QY \mid QY) = \text{cov}(PY) = P\Sigma. \end{aligned} \tag{13.7}$$

Thus, the conditional distribution of Y given Z is $N(QY, P\Sigma)$. These equations are dual to (13.6).

In standard probability terminology, prediction calls for the conditional distribution given the σ-field generated by the observation as a measurable transformation. By definition, the σ-field generated by a linear transformation with kernel $\mathcal{K} \subset \mathbb{R}^n$ is the Borel σ-field in $\mathbb{R}^n/\mathcal{K}$, i.e., all Borel subsets $A \subset \mathbb{R}^n$ such that $A + \mathcal{K} = A$. In that probabilistic sense, all linear transformations having the same kernel are equivalent.

Partitioned matrix representation

In applied work, it is often the case that $Y\colon U \to \mathbb{R}$ is a function on the units, and the observation $Z = Y[U_0]$ is the restriction of Y to a sub-sample $U_0 \subset U$

of size $n-k$. For that setting, it is convenient and computationally efficient to use partitioned-matrix notation in which $Y_0 = Y[U_0]$, and $Y_1 = Y[\bar{U}_0]$ is the restriction to the complementary subsample:

$$Y = \begin{bmatrix} Y_0 \\ Y_1 \end{bmatrix}; \quad K = [I_{n-k} : 0]; \quad \Sigma = \begin{bmatrix} \Sigma_{00} & \Sigma_{01} \\ \Sigma_{10} & \Sigma_{11} \end{bmatrix}; \quad \Sigma^{-1} = \begin{bmatrix} W_{00} & W_{01} \\ W_{10} & W_{11} \end{bmatrix};$$
$$Q = \begin{bmatrix} I_{n-k} & 0 \\ \Sigma_{10}\Sigma_{00}^{-1} & 0 \end{bmatrix}; \quad P = \begin{bmatrix} 0 & 0 \\ -\Sigma_{10}\Sigma_{00}^{-1} & I_k \end{bmatrix}; \quad P\Sigma = \begin{bmatrix} 0 & 0 \\ 0 & \Sigma_{11} - \Sigma_{10}\Sigma_{00}^{-1}\Sigma_{01} \end{bmatrix}.$$

The conditional distribution of Y_1 given Y_0 is Gaussian with moments

$$E(Y_1 \mid Y_0) = (QY)_1 = \Sigma_{10}\Sigma_{00}^{-1}Y_0;$$
$$\operatorname{cov}(Y_1 \mid Y_0) = (P\Sigma)_{11} = \Sigma_{11} - \Sigma_{10}\Sigma_{00}^{-1}\Sigma_{01} = W_{11}^{-1}.$$

This component-wise version of the conditional distribution speaks directly to the goal of predicting the values for extra-sample units, and it is computationally more efficient than (13.7) because it sets aside obvious degeneracies. But it does so at the cost of obscuring a crucial aspect of the geometry, namely that Gaussian point prediction is an orthogonal projection.

In applied work, the situation is typically a little more complicated because $\mu = E(Y)$ is never zero, in which case the conditional mean is

$$E(Y_1 \mid Y_0) = \mu_1 + \Sigma_{10}\Sigma_{00}^{-1}(Y_0 - \mu_0).$$

In practice, unknown parameters must be estimated before this can be computed.

Example: exchangeable Gaussian process

A zero-mean exchangeable Gaussian process has finite-dimensional distributions

$$Y[n] \sim N\big(0,\ \Sigma_n = \sigma_0^2 I_n + \sigma_1^2 J_n\big),$$

where $J_n(i,j) = 1$ is the $n \times n$ matrix whose components are all one. The inverse matrix is

$$\Sigma_n^{-1} = \sigma_0^{-2}\Big(I_n - \frac{\theta}{1+n\theta}J_n\Big),$$

where $\theta = \sigma_1^2/\sigma_0^2$ is the variance-component ratio.

Regardless of the variance parameters, $P_n = J_n/n$ is the orthogonal projection onto the subspace $\mathbf{1}_n$ of constant functions, and $Q_n = I_n - J_n$ is the complementary orthogonal projection. Thus, the projected random vectors

$$P_nY \sim N(0,\ n^{-1}J_n\Sigma) = N\big(0,\ (\sigma_0^2/n + \sigma_1^2)J_n\big)$$
$$Q_nY \sim N(0,\ Q_n\Sigma) = N\big(0,\ \sigma_0^2 Q_n\big)$$

are independent. To avoid confusion in statistical work where the covariance matrix is not completely known, it is best to fix the inner product in $\mathcal{H}$ rather

than having a parameter-dependent inner product: see the cautionary remarks in section 13.2.2. Most statistical work involving exchangeable or partially exchangeable processes uses the standard invariant inner product $\sum x_i y_i$. With this understanding, the squared norms

$$\|P_n Y\|^2 = n\bar{Y}_n^2 \sim (\sigma_0^2 + n\sigma_1^2)\chi_1^2,$$
$$\|Q_n Y\|^2 = (n-1)s_n^2 \sim \sigma_0^2 \chi_{n-1}^2$$

are independent χ^2 random variables with scale factors as indicated. Much of analysis of variance for balanced designs is based on extensions of this result to partially exchangeable arrays.

For $n \geq 1$, the partitioned-matrix formulae in the preceding subsection imply that the conditional distribution of $Y_{n+1}, \ldots, Y_{n+m}$ given $Y[n]$ is exchangeable Gaussian with moments

$$E(Y[n+1{:}m] \mid Y[n]) = \frac{n\theta \bar{Y}_n}{1+n\theta}\mathbf{1}_m,$$
$$\operatorname{cov}(Y[n+1{:}m] \mid Y[n]) = \sigma_0^2 I_m + \sigma_1^2 J_m/(1+n\theta).$$

The same partitioned-matrix formulae also imply that the the conditional distribution of the average $(Y_{n+1} + \cdots + Y_{n+m})/m$ given $Y[n]$ is Gaussian with moments

$$E\big(\bar{Y}_{n+1:m} \mid Y[n]\big) = \frac{n\theta \bar{Y}_n}{1+n\theta},$$
$$\operatorname{var}\big(\bar{Y}_{n+1:m} \mid Y[n]\big) = \sigma_0^2/m + \sigma_1^2/(1+n\theta).$$

This conditional distribution has a limit as $m \to \infty$ for fixed n, implying that the infinite average is a conditionally non-degenerate random variable such that

$$\bar{Y}_\infty \sim N\Big(\frac{n\theta \bar{Y}_n}{1+n\theta},\ \frac{\sigma_0^2 \theta}{1+n\theta}\Big).$$

The limit $\theta \to \infty$ gives $\bar{Y}_\infty - \bar{Y}_n \sim N(0, \sigma_0^2/n)$. For $n \geq 2$, the internally-standardized ratio

$$\frac{\sqrt{n}\,(\bar{Y}_\infty - \bar{Y}_n)}{s_n} \sim t_{n-1},$$

is distributed as Student's t on $n-1$ degrees of freedom.

13.4.4 Eddington's formula

Scalar signal estimation

Given a random signal $X \sim P$ and an observation $Y = X + \varepsilon$ contaminated by independent additive Gaussian noise, how do we estimate the signal? This version of the signal estimation problem was first posed in 1926 by the Astronomer Royal, Sir Frank Watson Dyson, in connection with parallax resolution problems in astronomical observations made at Greenwich.

In the astronomical setting, there is a large number of independent signals, all identically distributed according to some unknown signal distribution p, and the parallaxes $Y_i = X_i + \varepsilon_i$ are contaminated by independent additive Gaussian error with known variance. It is feasible to estimate the marginal density by smoothing, but it is not obvious how to estimate the signal distribution or how to adjust the observation to account for measurement error. Eddington provided a simple and elegant solution.

If the signal density is $p(\cdot)$, and the noise is standard Gaussian, the joint density of (Y, X) is $p(x)\phi(y-x)$, and the marginal density is

$$m(y) = \int p(x)\phi(y-x)\,dx = \phi(y)\int_{\mathbb{R}} p(x)e^{-x^2/2+xy}\,dx.$$

Thus the density ratio $m(y)/\phi(y)$ is the Laplace transform of the function $p(x)e^{-x^2/2}$. In addition, $p(x)e^{-x^2/2}\phi(0)/m(0)$ is a probability density whose cumulant-generating function is $\log\bigl(m(y)/\phi(y)\bigr)$. However it is phrased, the goal of signal estimation is to compute the conditional expected value of the signal given the data.

$$\begin{aligned}
E(e^{tX} \mid Y) &= \int e^{tx}p(x)\phi(y-x)\,dx \Big/ m(y) \\
&= \frac{\phi(y)}{m(y)}\int e^{-x^2/2+xy+tx}p(x)\,dx \\
&= \frac{m(y+t)}{\phi(y+t)}\Big/\frac{m(y)}{\phi(y)}; \\
E(X \mid Y) &= \frac{d}{dt}\log\left(\frac{m(y+t)}{\phi(y+t)}\right)_{t=0} \\
&= \frac{d}{dy}\log\left(\frac{m(y)}{\phi(y)}\right) = y + \frac{m'(y)}{m(y)}.
\end{aligned}$$

Eddington's solution was the additive adjustment $\sigma^2 m'(y)/m(y)$, scaled for the observation variance. Higher-order derivatives are the higher-order cumulants of the conditional distribution.

Dyson started out with a table or histogram of parallaxes of stars measured at Greenwich, from which he estimated the marginal density by smoothing. Knowing the observation variance from replicates, he estimated the adjustment and reported the adjusted values. He also noted that if the signal distribution happens to be normal, the correction reduces to $-y\sigma^2/(\sigma^2 + \sigma_x^2)$. Dyson's observed parallax distribution was strongly skewed in the positive direction, so his signal distribution was far from normal.

Eddington's formula is remarkable for two reasons. The most obvious is that it depends only on the marginal density of the observations. Less obvious is the fact that if the signals are restricted to an interval, say $-1 \le x \le 2$ or $x > 0$, then $E(X \mid Y)$ also lies in the interval. This aspect was crucial for Dyson's task because parallaxes are positive even if some observations are negative, and Dyson was understandably reluctant to report a negative value.

Isotropic vector signal estimation

Eddington's formula applies also to vector signals $X \in \mathbb{R}^d$ contaminated by additive Gaussian noise $\varepsilon \sim N_d(0, \Sigma)$. The conditional mean given $Y = X + \varepsilon$ is then

$$E(X \mid Y) = \Sigma \frac{d}{dy} \log\left(\frac{m(y)}{\phi(y)}\right) = y + \Sigma m'(y)/m(y),$$

where ϕ is the density of $N_d(0, \Sigma)$, and $m'(y)$ is the gradient vector.

For the vector formula to be useful in practical work, it is usually necessary to make further simplifying assumptions. Rotational symmetry for both the signal and the noise is usually the most natural. In that case $\varepsilon \sim N(0, I_d)$, the signal density satisfies $p(\sigma x) = p(x)$ for each orthogonal transformation $\sigma \colon \mathbb{R}^d \to \mathbb{R}^d$, and $m(\sigma y) = m(y)$ is also rotationally symmetric. The conditional expectation $H(y) = E(X \mid Y = y)$ then satisfies the commutativity condition

$$H(\sigma y) = \sigma H(y),$$

which means that the transformation $y \mapsto H(y)$ is radial. The direction is retained, so the unit vectors $y/\|y\|$ and $H(y)/\|H(y)\|$ are equal. Ordinarily, the modification to the norm is shrinkage towards the origin, but not necessarily so: see Exercise 13.10.

Isotropic matrix reconstruction

Suppose that the signal X is a random matrix of order $n \times p$, and that the components of ε are independent standard normal. Suppose also that the signal distribution is rotationally symmetric in the sense that the density is invariant with respect to left and right orthogonal transformation. In other words, $p(\sigma x \tau) = p(x)$ for all orthogonal matrices σ of order n and τ of order p. Since ε is rotationally symmetric, the convolution is also rotationally symmetric in the same sense, and the conditional expectation is equi-variant in the sense $H(\sigma y \tau) = \sigma H(y) \tau$. Equi-variance implies that H acts only on the singular values; rotational symmetry ensures that the left and right singular vectors are retained in the reconstruction.

Although the conditional expectation $y \mapsto H(y)$ is an action on singular values, the transformation does not necessarily act component-wise, nor is it necessarily a shrinkage. Under certain sparsity assumptions, it is possible to be more specific about the nature of the transformation, which is a shrinkage towards the origin applied component-wise to the singular values: see section 15.5.3.

13.4.5 Linear regression

Let $\mathcal{X} \subset \mathbb{R}^n$ be a subspace of dimension p spanned by the columns of the given matrix X of order $n \times p$, let V be a given strictly positive definite matrix with inverse $W = V^{-1}$, and let $\mathcal{H}$ be the Euclidean space with inner product $\langle x, y \rangle = x'Wy$.

Linear regression refers to the *family* of Gaussian distributions on $\mathbb{R}^n$

$$\{N_n(\mu, \sigma^2 V) \mid \mu = X\beta \in \mathcal{X},\ \sigma > 0\}$$

indexed by $\beta \in \mathbb{R}^p$ and $\sigma > 0$. For geometrical purposes, we want to regard each of these as a distribution on the same Hilbert space, so we choose the given matrix $W = V^{-1}$ rather than Σ^{-1} in order that all operations in $\mathcal{H}$ be computable.

We now suppose that, for some unspecified parameter point (β, σ^2), an observation $Y \sim N_n(X\beta, \sigma^2 V)$ is generated and the value $y \in \mathcal{H}$ is observed. To estimate the parameter, we use the log likelihood function, which is the log density

$$l(\beta, \sigma^2; y) = -\tfrac{1}{2}\|y - X\beta\|^2/\sigma^2 - n \log \sigma + \text{const.}$$

So far as the regression parameter is concerned, maximization of the log likelihood is equivalent to minimizing the Euclidean squared distance $\|y - X\beta\|^2$ over points $\mu = X\beta$ in $\mathcal{X}$. Regardless of σ^2, the minimum over $\mathcal{X}$ occurs at the Euclidean projection

$$\hat{\mu} = X\hat{\beta} = Py = X(X'WX)^{-1}X'Wy,$$

and the minimum value achieved is the residual quadratic form $\|(I - P)y\|^2 = \|Qy\|^2$.

The projection PY and its complement QY are independent Gaussian random vectors with distributions

$$PY \sim N(X\beta, \sigma^2 PV), \qquad QY \sim N(0, \sigma^2 QV).$$

It follows from section 13.4.1 that $\|QY\|^2/\sigma^2$ is distributed as χ^2_{n-p}, i.e., the weighted residual sum of squares $\|QY\|^2$ is distributed as $\sigma^2\chi^2_{n-p}$. The conventional estimate of σ^2 is the mean-squared residual

$$s^2 = \|Qy\|^2/(n - p),$$

which is unbiased and strictly larger than the maximum-likelihood estimate $\|Qy\|^2/n$.

Statistical computer packages invariably report the least-squares coefficient vector $\hat{\beta} = (X'WX)^{-1}X'Wy$ together with the standard errors, which are the square roots of the diagonal components of the matrix

$$\text{cov}(\hat{\beta}) = s^2(X'WX)^{-1}.$$

13.4.6 Linear regression and prediction

Exercises 13.1–13.5 give an outline of the argument for combining linear regression with prediction. The parametric family $Y \sim N(X\beta, \sigma^2 V)$ on $\mathcal{H} = (\mathbb{R}^n, W)$ determines a parametric family on the observation space, which is the image of a linear transformation $K\colon \mathbb{R}^n \to \mathbb{R}^{n-k}$. The regression parameter is estimated by

weighted least squares based on the observation $Z \sim N_{n-k}(KX\beta, \sigma^2 KVK')$, or equivalently, based on $QY \sim N_n(QX\beta, \sigma^2 QV)$, where Q is the orthogonal projection having the same kernel. Identifiability requires $p = \operatorname{rank}(KX) \leq n-k$, in which case the composite transformation $L\colon Y \mapsto QY \mapsto \hat{\mu}$ is a linear projection $\mathcal{H} \to \mathcal{H}$.

The conditional distribution of Y given $Z = KY$ is the same as the conditional distribution of the sum $PY + QY$ given QY. Independence of PY and QY implies that the conditional distribution is

$$N_n(QY + P\mu,\ \sigma^2 PV).$$

The least squares estimate is obtained by parameter substitution

$$N_n(QY + P\hat{\mu},\ s^2 PV) = N(QY + L_0 Y,\ \sigma^2 PV), \tag{13.8}$$

with σ^2 replaced by s^2 if needed.

The transformation $Y \mapsto \hat{\mu} = LY$ is a linear projection whose image is $\mathcal{X}$ and whose kernel includes $\mathcal{K}$. These are arbitrary non-overlapping subspaces so L is not orthogonal in $\mathcal{H}$. It is the sum of two transformations $L_1 = QL$, which is the orthogonal projection with image $Q\mathcal{X}$, and $L_0 = PL$ which is nilpotent, i.e., $L_0^2 = 0$, because $LP = 0$.

Notation for component-wise transformation

If the observation is a component-wise restriction of Y, it is convenient and efficient to express (13.8) in partitioned matrix notation such that $K = [I_{n-k}, 0]$, $KY = Y_0$ and $KX = X_0$, in which case

$$QY = \begin{pmatrix} Y_0 \\ V_{10}V_{00}^{-1}Y_0 \end{pmatrix}, \quad P\hat{\mu} = \begin{pmatrix} 0 \\ \hat{\mu}_1 - V_{10}V_{00}^{-1}\hat{\mu}_0 \end{pmatrix},$$
$$(PV)_{11} = V_{11} - V_{10}V_{00}^{-1}V_{01} = W_{11}^{-1}.$$

The least-squares estimate is $\hat{\beta} = (X_0' V_{00}^{-1} X_0)^{-1} X_0' V_{00}^{-1} Y_0$, giving the fitted mean with components $\hat{\mu}_0 = X_0\hat{\beta}$ and $\hat{\mu}_1 = X_1\hat{\beta}$. Given Y_0, the predictive distribution for Y_1 has moments

$$\hat{\mu}_1 + V_{10}V_{00}^{-1}(Y_0 - \hat{\mu}_0) \quad \text{and} \quad \sigma^2 W_{11}^{-1},$$

with σ^2 replaced by s^2 where needed. The predictive mean in this setting goes by various names—best linear predictor, fiducial predictor, Kriging estimate, smoothing spline—depending on the area of application.

Fiducial prediction

The least-squares predictive distribution (13.8) associates with each observation point $z = Ky$ in $\mathbb{R}^{n-k}$ a probability distribution on $\mathbb{R}^n$. It assigns probability one to the k-dimensional coset

$$y + \mathcal{K} = Qy + \mathcal{K} = \{y \in \mathbb{R}^n \mid Ky = z\}, \tag{13.9}$$

which is the subset of points that are consistent with the observed value.

Any distribution defined on Borel subsets of $\mathbb{R}^n$ can be restricted to a sub-σ-field if the need arises. In the linear-model setting $N_n(X\beta, \sigma^2 V)$ with $\mathcal{X} = \text{span}(X)$, an event $A \subset \mathbb{R}^n$ is said to be translation-invariant if $A + \mathcal{X} = A$. For historical reasons, the invariant events are also called fiducial events; the set of fiducial events is the Borel σ-field $\mathcal{B}(\mathbb{R}^n/\mathcal{X})$.

According to the fiducial argument, the set of distributions $N_n(X\beta, \sigma^2 V)$ indexed by $\beta \in \mathbb{R}^p$ for fixed σ is interpreted as a single distribution $N_n(0, \sigma^2 V)$ on fiducial events. For this purpose, two Gaussian distributions $N_n(\mu_0, \Sigma_0)$ and $N_n(\mu_1, \Sigma_1)$ are equivalent modulo $\mathcal{X}$, if, for any linear transformation T whose kernel includes $\mathcal{X}$, $T\mu_0 = T\mu_1$ and $T\Sigma_0 T' = T\Sigma_1 T'$. In particular, $\ker(Q_\mathcal{X}) = \mathcal{X}$ implies that the distributions $N_n(X\beta, V)$ and $N_n(0, Q_\mathcal{X} V)$ are fiducially equivalent. Thus, a single fiducial distribution has multiple covariance-matrix representations in $\mathbb{R}^n$.

Fiducially speaking, the response distribution is $N_n(0, \sigma^2 Q_\mathcal{X} V)$, the observation is a linear transformation $Q^\dagger$ with kernel $\mathcal{X} + \mathcal{K}$, and (13.7) implies that the conditional distribution given the observation is

$$\begin{aligned} N_n\big(Q^\dagger Y,\ \sigma^2(Q_\mathcal{X} - Q^\dagger)Q_\mathcal{X} V\big) &= N_n\big(Q^\dagger Y,\ \sigma^2(Q_\mathcal{X} - Q^\dagger)V\big) \\ &\cong N_n\big(Q^\dagger Y + \hat{\mu},\ \sigma^2(P^\dagger - P_\mathcal{X})V\big) \\ &\cong N_n\big(Q^\dagger Y + \hat{\mu},\ \sigma^2 P_\mathcal{K} V\big). \end{aligned}$$

The predictive distribution is restricted to fiducial events in $\mathbb{R}^n$, and these various expressions are equivalent when restricted to $B(\mathbb{R}^n/\mathcal{X})$. For example, the addition of $\hat{\mu}$ to the mean has no effect. The last expression is the [fiducal restriction of the] least-squares predictive distribution.

13.5 Additivity

13.5.1 1DOFNA algorithm

One degree of freedom for non-additivity is a technique introduced by Tukey (1949) to check the adequacy of the linear model $Y \sim N(X\beta, \sigma^2 V)$ by testing for deviations from additivity and/or linearity. Tukey was particularly concerned with additivity assumptions for factorial models used in randomized-blocks and Latin-square designs, but the technique is valid more broadly for simple linear regression and multiple linear regression.

The computational procedure goes as follows. First compute the least-squares fitted vector $\hat{\mu} = X\hat{\beta} = PY$ together with the residual sum of squares $\|Y - \hat{\mu}\|^2$ on $n - p$ degrees of freedom. Second, compute the derived vector z with components $z_i = \hat{\mu}_i^2$. Third, fit the extended linear model including both X and z, i.e., $E(Y) = X\beta + z\gamma$, and compute the least-squares fitted vector $\hat{\mu}_1 = P_1 Y$ by projection onto the subspace spanned by X and z.

The reduction in residual sum of squares $\|QY\|^2 - \|Q_1 Y\|^2$ is the one degree of freedom for non-additivity. According to Tukey, the null distribution is exactly

$\sigma^2\chi^2_1$, and the mean-square ratio

$$F = \frac{\|QY\|^2 - \|Q_1Y\|^2}{\|Q_1Y\|^2/(n-p-1)} \tag{13.10}$$

is distributed exactly as $F_{1,n-p-1}$ if the null assumption $Y \sim N(X\beta, \sigma^2 V)$ is correct. The 1DOFNA F-ratio provides an exact test of the null model, large values being interpreted as evidence of non-additivity. Alternatively, the least-squares coefficient of z can be used in the standard manner

$$T = \hat{\gamma}/\,\text{s.e.}(\hat{\gamma}),$$

and the value compared with the null distribution t_{n-p-1}. As always, $T^2 = F$, so the two approaches are effectively equivalent. If the F-ratio is large, the remedy suggested is to transform the response $Y \mapsto Y^\lambda$, and Tukey's suggested power transform is $\lambda = 1 - 2\hat{\gamma}\bar{Y}$.

13.5.2 1DOFNA theory

Although this description is entirely satisfactory as a computational procedure, the notation is misleading because the transformation that sends u to P_1u is neither a projection $\mathcal{H} \to \mathcal{H}$ nor a linear transformation. It does not satisfy $P_1(u+v) = P_1u + P_1v$ for vectors $u, v \in \mathcal{H}$, nor does it satisfy $P_1^2 = P_1$. Hence, Tukey's claim is not an immediate consequence of earlier remarks such as Cochran's theorem (section 13.3.2), which is concerned exclusively with linear transformations and orthogonal projections.

A correct argument proceeds as follows. First, the projected random vectors PY and QY are independent, so the conditional distribution of QY given $\hat{\mu} = PY$ is $N(0, \sigma^2 QV)$, and the conditional distribution of QY given z is also $N_n(0, \sigma^2 QV)$. It follows that $\gamma = 0$ if the null model holds. Given $\hat{\mu}$, least-squares estimate of γ is the regression coefficient of the residuals on z—or more correctly on Qz—which is conditionally linear

$$\hat{\gamma} = (z'WQz)^{-1}z'WQY,$$

and the conditional distribution given $\hat{\mu}$ is $N(0, \sigma^2(z'WQz)^{-1})$. Given $\hat{\mu}$, the residual vector is split additively into two orthogonal parts

$$QY = Qz\hat{\gamma} + Q(Y - z\hat{\gamma}) = Qz(z'WQz)^{-1}z'WQY + Q_1Y.$$

The residual sum of squares also splits into two parts

$$\|QY\|^2 = \gamma'(z'WQz)\gamma + \|Q_1Y\|^2 = Y'WQz(z'WQz)^{-1}z'WQY + Y'WQ_1Y.$$

According to section 13.3.2, these are conditionally independent given $\hat{\mu}$ with distributions $\sigma^2\chi^2_1$ and $\sigma^2\chi^2_{n-p-1}$ respectively. Thus, Tukey's distributional assertions are upheld conditionally on $\hat{\mu}$, and therefore unconditionally.

13.5.3 Scope and rationale

Apart from the condition $p \leq n-2$, which is needed to ensure $Q_1 \neq 0$, one additional condition related to the algebraic properties of the subspace $\mathcal{X}$ is needed. A subspace $\mathcal{X} \subset \mathbb{R}^n$ is said to be a commutative ring if, for each pair of vectors $u, v \in \mathcal{X}$, the component-wise product $uv = vu$ also belongs to $\mathcal{X}$. If $\mathcal{X}$ happens to be a ring, then $\hat{\mu} \in \mathcal{X}$ implies $z = \hat{\mu}^2 \in \mathcal{X}$, so that $Qz = 0$. Thus, if $\mathcal{X}$ is a ring under multiplication, no degree of freedom for non-additivity exists.

In the factorial-model setting with non-nested block or treatment factors $A, B, C, \ldots$, each of the factorial subspaces

$$\mathbf{0},\ \mathbf{1},\ A,\ B,\ C,\ AB,\ AC,\ BC,\ ABC$$

is also a commutative ring that is closed under multiplication. (Here, AB is the subspace denoted by either `A:B` or `A*B` in R notation.) In order for the 1DOFNA to be non-trivial, it is necessary that $\mathcal{X}$ not be a ring—in other words the set of points $x \in \mathcal{X}$ such that $x^2 \in \mathcal{X}$ must have measure zero. Apart from degenerate designs having completely aliased factors, the 1DOFNA is non-trivial for every other factorial subspace such as $A + B$ or $AB + C$ or $AB + BC$, that includes a non-trivial + operator. Note that `1 + A` is the same as A and `A+B+A*B` is the same as AB, both of which are rings.

In the case of simple linear regression with $E(Y_i) = \beta_0 + \beta_1 x_i$ for a quantitative variable x, the constructed variable $\hat{\mu}^2$ is a quadratic function of x. Provided that the design contains at least three distinct x-values, the vectors $1, x, x^2$ are linearly independent. With probability one $\hat{\beta}_1 \neq 0$, in which case the vectors $1, x, \hat{\mu}^2$ are also linearly independent. For that setting, the 1DOFNA is equivalent to the one degree of freedom for non-linearity, and specifically quadratic deviations from linearity.

Provided that the constructed variable is a function of $\hat{\mu}$, any component-wise non-linear transformation such as $z_i = \exp(\hat{\mu}_i)$, or any non-component-wise transformation $\mathcal{H} \to \mathcal{H}$, may be used in the algorithm. Subject to the condition $z \notin \mathcal{X}$ mentioned above, the distributional argument leading to the conclusion that the 1DOFNA F-ratio is distributed as $F_{1,n-p-1}$ is unaffected by the choice of transformation. For $\mathcal{X} \neq \mathbf{1}$, the neighbour average $z_i = \text{ave}_{j \in \text{nb}(i)}\, \hat{\mu}_j$ is an example of a linear non-component-wise transformation $\mathcal{H} \to \mathcal{H}$ that might arise in a spatial or graphical setting.

Note that the word transformation is used above in two distinct senses that are algebraically distinct. First, every statistical vector is a function $U \to \mathbb{R}$ on the units, and component-wise transformation $g\colon \mathbb{R} \to \mathbb{R}$ refers to composition $y \mapsto gy$ on the left as illustrated by the diagram $U \xrightarrow{y} \mathbb{R} \xrightarrow{g} \mathbb{R}$. Component-wise transformation exploits the fact that $\mathbb{R}^U$ is a commutative ring. Second, every statistical vector is also a point $y \in \mathcal{H}$, and a typical linear transformation $\mathcal{H} \to \mathcal{H}$ such as $y \mapsto \hat{\mu}$ or $y \mapsto Qy$ does not act component-wise.

13.6 Exercises

13.1 Show that the set of $2n \times 2n$ real matrices of the form

$$\begin{pmatrix} A & B \\ -B & A \end{pmatrix}$$

is closed under matrix addition and multiplication. Show also that the 'linear' mapping into the space of complex $n \times n$ matrices

$$\begin{pmatrix} A & B \\ -B & A \end{pmatrix} \mapsto A + iB$$

is an isomorphism preserving addition and multiplication.

13.2 Let $A + iB$ be a full-rank Hermitian matrix of order n. Show that the inverse matrix $C + iD$ is also Hermitian and satisfies the pair of equations

$$AD + BC = 0; \quad AC - BD = I_n.$$

Deduce that the $2n \times 2n$ real symmetric matrices

$$\begin{pmatrix} A & B \\ -B & A \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} C & D \\ -D & C \end{pmatrix}$$

are mutual inverses. What does this matrix isomorphism imply about the relation between complex Gaussian vectors and real Gaussian vectors?

13.3 By writing the complex vector z and the Hermitian matrix Γ as a linear combination of real and imaginary parts, show that the Hermitian quadratic form $z^*\Gamma z$ reduces to the following linear combination of real quadratic forms:

$$(x' - iy')\Gamma_0(x + iy) + i(x' - iy')\Gamma_1(x + iy) = x'\Gamma_0 x + y'\Gamma_0 y + y'\Gamma_1 x - x'\Gamma_1 y.$$

Hence deduce that the real and imaginary parts of $Z \sim \mathbb{C}N(0, \Sigma)$ are identically distributed Gaussian vectors $N(0, \Sigma_0)$ with covariances $\operatorname{cov}(X, Y) = -\operatorname{cov}(Y, X) = \Sigma_1$.

Gaussian linear prediction: The next five exercises are concerned with estimation and prediction in the Gaussian linear model $Y \sim N_n(\mu = X\beta, \sigma^2 V)$ in which the observation is the linear transformation $Z = KY$. The matrices X of order $n \times p$, K of order $n - k \times n$, and V of order $n \times n$ are given, while β, σ^2 are parameters to be estimated. All three matrices are of full rank, the product KX has rank $p \leq n - k$, while the Hilbert space $\mathcal{H}$ with inner-product matrix $W = V^{-1}$ determines the geometry.

13.4 Show that the maximum-likelihood estimate of β satisfies

$$[X'K'(KVK')^{-1}KX]\hat{\beta} = X'K'(KVK')^{-1}Z = X'WQY,$$

where $Q\colon \mathcal{H} \to \mathcal{H}$ is the orthogonal projection with kernel $\ker(K)$.

13.5 Deduce that the linear transformation $Y \mapsto LY = \hat{\mu} = X\hat{\beta}$ is a projection $\mathcal{H} \to \mathcal{H}$, but not an orthogonal projection unless $Q\mathcal{X} = \mathcal{X}$.

13.6 Deduce that the composite linear transformation $Y \mapsto L_1 Y = Q\hat{\mu}$ is is also a projection, and that it is the orthogonal projection whose image is the p-dimensional subspace $Q\mathcal{X}$.

13.7 For the complementary projection $P = I_n - Q$ whose image is $\mathcal{K}$, deduce that the composite linear transformation $Y \mapsto L_0 Y = P\hat{\mu}$ is nilpotent, i.e., that $L_0^2 = 0$. What does nilpotence imply about the image and kernel of L_0? Construct the multiplication table for L_0, L_1.

13.8 Show that the least-squares estimate of the conditional distribution of Y given Z is

$$N_n\big(QY + P\hat{\mu},\ s^2 PV\big)$$

for some scalar s^2. Show that that the least-squares estimate is singular and is supported on the k-dimensional coset $QY + \mathcal{K}$. Explain why self-consistency requires $KQ = K$.

13.9 Show that the zero-mean exchangeable Gaussian process in section 13.3.3 with covariances

$$\mathrm{cov}(Y_r, Y_s) = \sigma_0^2 \delta_{rs} + \sigma_1^2,$$

has a dynamic or sequential representation beginning with $Y_0 = 0$ followed by

$$Y_{n+1} = \frac{n\theta \bar{Y}_n}{1 + n\theta} + \sigma_0 \sqrt{1 + \theta/(1 + n\theta)}\, \epsilon_{n+1}$$

for $n \geq 0$. Here $\theta = \sigma_1^2/\sigma_0^2$ is the variance ratio, and $\epsilon_1, \ldots$ are independent standard normal variables.

13.10 Suppose that X is uniformly distributed on the surface of the unit sphere in $\mathbb{R}^d$, and that $Y \sim N(X, \sigma^2 I_d)$ is observed. Show that Eddington's formula reduces to the projection $E(X \mid Y) = Y/\|Y\|$.

13.11 Suppose that X is uniformly distributed on the interior of the unit sphere in $\mathbb{R}^d$, and that $Y \sim N(X, \sigma^2 I_d)$ is observed. Show that Eddington's formula is a radial shrinkage so that $E(X \mid Y)$ has norm strictly less than one.

Chapter 14

Space-time processes

14.1 Gaussian processes

Let $\mathcal{U}$ be an arbitrary index set, here identified with the domain. A Gaussian process associates with each u in the domain a random variable Z_u in such a way that for each sample $U = \{u_1, \ldots, u_n\}$, the random variable $Z[U] = (Z_{u_1}, \ldots, Z_{u_n})$ has a Gaussian distribution. It should be noted that the sample points are taken in a specific order, so U is an n-tuple of points from the domain, and the components of Z are taken in the same order. If U contains repeats, say $U = (u_1, u_1, u_2)$, then the first two components of $Z[U]$ are necessarily identical.

As a function on the index set, Z may be real-valued or complex-valued or $\mathbb{R}^k$-valued or $\mathbb{C}^k$-valued. This chapter focuses entirely on scalar processes, either real-valued or complex-valued, so Z is a function $\mathcal{U} \to \mathbb{R}$ or a function $\mathcal{U} \to \mathbb{C}$ into the space of scalars. Since the complex numbers are in 1–1 correspondence with ordered pairs of reals, every complex-valued process $Z = X + iY$ is also a $\mathbb{R}^2$-valued process (X, Y). Any reader who has made it this far has every right to ask why, in a book that professes to be concerned with scientific applications of statistical ideas, we should concern ourselves with a complex-valued process when a $\mathbb{R}^2$-valued process would serve the same purpose. However, there is a legitimate reason, which is central to the theme of this chapter. For reasons discussed below, an arbitrary $\mathbb{R}^2$-valued Gaussian process (X, Y) is not a complex Gaussian process in the algebraic sense. The algebra of the complex numbers is not irrelevant in the real world.

A Gaussian process is determined by its mean function $\mu(\cdot)$ and its covariance function $K(\cdot, \cdot)$. In the case of a real-valued process, μ is a function $\mathcal{U} \to \mathbb{R}$, and K is a symmetric function $\mathcal{U} \times \mathcal{U} \to \mathbb{R}$ that is also positive definite. In the case of a complex-valued process, μ is a function $\mathcal{U} \to \mathbb{C}$, and K is a positive-definite Hermitian function $\mathcal{U} \times \mathcal{U} \to \mathbb{C}$. Hermitian symmetry means that $K(u, u')$ is the complex conjugate of $K(u', u)$, so K is real and positive on the diagonal. The mean function is $\mu(u) = E(Z_u)$; the covariance function is $K(u, u') = \mathrm{cov}(Z_u, Z_{u'})$ for a real-valued process, and $K(u, u') = \mathrm{cov}(Z_u, \bar{Z}_{u'})$

for a complex-valued process. On the finite subset $U = (u_1, \ldots, u_n)$, the covariance matrix of $Z[U]$ is the finite restriction of K to ordered pairs (u_i, u_j). Positive definiteness means that the Hermitian form $\xi^* K \xi$ is non-negative for every complex n-vector ξ, and every finite restriction of K.

This chapter is concerned exclusively with variances and covariances, so $\mu(u) = 0$ throughout. In the case of a real-valued process the covariance function

$$K(u, u') = \operatorname{cov}\big(Z(u), Z(u')\big) = \operatorname{cov}\big(Z(u'), Z(u)\big) = K(u', u)$$

is necessarily real and symmetric. In the case of a complex-valued process $Z_u = X_u + iY_u$ is a pair of real-valued Gaussian processes with covariance functions K_X and K_Y respectively. For each pair of points u, u', not necessarily distinct, there are two [linearly independent] complex products and four real products whose means are as follows:

$$\begin{aligned} E(Z_u Z_{u'}) &= E(X_u X_{u'} - Y_u Y_{u'}) + iE(X_u Y_{u'} + X_{u'} Y_u) \\ &= K_X(u, u') - K_Y(u, u') + i\big(K_{XY}(u, u') + K_{XY}(u', u)\big) = 0; \\ E(Z_u \bar{Z}_{u'}) &= E(X_u X_{u'} + Y_u Y_{u'}) - iE(X_u Y_{u'} - X_{u'} Y_u) \\ &= K_X(u, u') + K_Y(u, u') - i\big(K_{XY}(u, u') - K_{XY}(u', u)\big) = K(u, u'). \end{aligned}$$

The first equation is the condition for a pair of real-valued Gaussian processes X, Y to determine a complex Gaussian process in the algebraic sense. The zero real part implies $K_X = K_Y$, so the real and imaginary parts of Z are two processes having the same distribution. The imaginary part of the first equation implies that the cross-covariances satisfy the mysterious skew-symmetry condition

$$\operatorname{cov}(Y_{u'}, X_u) = \operatorname{cov}(X_u, Y_{u'}) = -\operatorname{cov}(X_{u'}, Y_u) = -\operatorname{cov}(Y_u, X_{u'}).$$

As a consequence, all non-zero second moments of the complex-valued Gaussian process are encapsulated in the conjugated second moments $\operatorname{cov}(Z_u, \bar{Z}_{u'}) = K(u, u')$.

It is apparent from the preceding paragraph that if X, Y are independent real Gaussian processes having the same distribution with covariance function $K/2$, then $Z = X + iY$ is a complex Gaussian process whose covariance K is real and symmetric. In that sense, the only interesting complex Gaussian process are those whose covariance function has a non-zero imaginary part.

In most instances, $\mathcal{U}$ is a topological space such as the real line, the plane, the sphere or the torus, so the continuity or degree of smoothness of the function $u \mapsto Z(u)$ is of considerable interest. In principle, it is possible for the degree of smoothness to vary throughout the domain, in either a random manner or in a predetermined manner. But the processes described here are all well-behaved in the sense that they have the same behaviour throughout the domain.

14.2 Stationarity and isotropy

14.2.1 Definitions

Stationarity and isotropy are properties of a process that are associated with a group action on the domain. Stationarity is a symmetry or distributional invariance under domain translation; isotropy is an invariance under rotation or orthogonal transformation. For translation to make sense, the domain is necessarily a vector space or an affine space; for orthogonal transformation to make sense, the domain is necessarily a Euclidean space.

A stochastic process Z with domain $\mathcal{U}$ is said to be stationary if the following properties hold:

1. The domain is a vector space, either $\mathbb{R}^d$ or $\mathbb{C}^d$ for some $d \geq 0$;
2. Each $g \in \mathcal{U}$ acts on the domain by addition, sending u to $u + g$;
3. The action on the domain sends the original process to $Z^g(u) = Z(u + g)$ by composition, which is a translation by $-g$;
4. The process is stationary if each Z^g has the same distribution as Z.

Since the group acts transitively on the domain, stationarity implies that each Z_u has the same distribution as Z_0, i.e., all one-dimensional marginal distributions are equal. Since differences are invariant under translation, stationarity implies that $(Z_u, Z_{u'})$ has the same joint distribution as $(Z_v, Z_{v'})$ whenever $u - u' = v - v'$. Stationarity does not imply that the pair $(Z_u, Z_{u'})$ has the same distribution as the reverse pair $(Z_{u'}, Z_u)$.

Isotropy has a similar meaning in relation to a different group acting on the domain, which is necessarily a Euclidean space with an inner product:

1. The domain is Euclidean space, either $\mathbb{R}^d$ or $\mathbb{C}^d$ for some $d \geq 0$;
2. The orthogonal group [with positive determinant] acts on the domain, sending u to gu;
3. The action on the domain sends the original process to $Z^g(u) = Z(gu)$ by composition, which is a [reverse] rotation;
4. The process is isotropic if each Z^g has the same distribution as Z.

Sometimes it is necessary to ask for clarification whether the full orthogonal group, including reflections, is intended. Sometimes the domain may be a proper subset of Euclidean space on which the group acts, for example, the unit circle or the unit disk in the complex plane or the unit sphere in $\mathbb{R}^3$.

Ordinarily in applied work, the domain has no natural origin, so it is better described as an affine space. In such applications isotropy is not a natural requirement on its own, but the group of proper Euclidean motions (translation plus rotation) is very natural. Depending on the setting, reflections may or may not be included.

A zero-mean complex Gaussian process is stationary if and only if $K(u, u') = G(u - u')$ for some function G such that $G(-u) = \bar{G}(u)$. In the case of a real Gaussian process, G is real, and therefore symmetric. A zero-mean complex Gaussian process is stationary and isotropic if and only if $K(u, u') = G(\|u - u'\|)$ for some real-valued function G. Each function G is necessarily positive definite.

It follows that every stationary isotropic complex Gaussian process $Z = X + iY$ is a pair of independent isotropic real Gaussian processes $X \sim Y$ having the same distribution. Conversely, a pair $(X, Y) \mapsto X + iY$ of independent and identically distributed stationary isotropic real-valued Gaussian processes determines a complex Gaussian process. The situation for stationary non-isotropic processes is different.

14.3 Stationary Gaussian time series

14.3.1 Spectral representation

Any process whose domain is the set of real numbers is called a time series; points in the domain are denoted by t.

The Hermitian function

$$K_\omega(t, t') = e^{i\omega(t-t')},$$

which is the Hermitian outer product $\xi\xi^*$ of the vector $\xi(t) = e^{i\omega t}$ with itself, is positive definite of rank one. Accordingly, if μ is a non-negative measure on frequencies, the convex combination

$$K_\mu(t, t') = \int_{-\infty}^{\infty} e^{i\omega(t-t')}\, d\mu(\omega)$$

is positive definite Hermitian. As a function of $t - t'$, it is necessarily the covariance of a stationary Gaussian time series. Moreover, every stationary Gaussian process has an associated spectral measure.

In the algebra that follows it is helpful to split the spectral measure into symmetric and skew-symmetric parts $\mu = \mu_{\text{sym}} + \mu_{\text{alt}}$ as follows:

$$2\mu_{\text{sym}}(A) = \mu(A) + \mu(-A) \tag{14.1}$$

$$2\mu_{\text{alt}}(A) = \mu(A) - \mu(-A). \tag{14.2}$$

The symmetric part is non-negative. The alternating part is a signed measure such that $-1 \le (d\mu_{\text{alt}}/d\mu_{\text{sym}})(\omega) \le 1$ for every ω. If we write t for the temporal difference, the result of this decomposition of the measure is

$$\begin{aligned} K_\mu(t) &= \int_{-\infty}^{\infty} \cos(\omega t)\, d\mu_{\text{sym}}(\omega) + i \int_{-\infty}^{\infty} \sin(\omega t)\, d\mu_{\text{alt}}(\omega) \\ &= K_\mu^{\text{sym}}(t) + iK_\mu^{\text{alt}}(t), \end{aligned}$$

which is a decomposition of K into real and imaginary parts. Every pair of symmetric and alternating measures such that $|\mu_{\text{alt}}| \le \mu_{\text{sym}}$ gives rise to a positive definite Hermitian function, and vice-versa.

14.3.2 Matérn class

Let $\nu > 0$ be a fixed index, and $-1 \le a \le 1$ a given constant. The Matérn spectral measure is symmetric with density

$$d\mu_{\text{sym}}(\omega) = \frac{d\omega}{(1+\omega^2)^{\nu+1/2}},$$

so the Matérn covariance function is real and symmetric. To a certain extent, the choice of μ_{alt} is arbitrary, but one choice that pairs well is

$$d\mu_{\text{alt}}(\omega) = \frac{d\omega}{(1+\omega^2)^{\nu+1/2}} \times \frac{2b\omega}{1+\omega^2}.$$

The condition $-1 \le b \le 1$ implies that the skew factor $2bw/(1+w^2)$ lies in $[-1,1]$, so $|\mu_{\text{alt}}| \le \mu_{\text{sym}}$. Other possibilities for the skew factor include $a\omega/(1+\omega^2)^{1/2}$.

For the spectral measures shown above, the covariance function is proportional to

$$K_\mu(t) = |t|^\nu \mathcal{K}_\nu(|t|)\big(1 + ibt/(\nu+1/2)\big), \tag{14.3}$$

where $\mathcal{K}_\nu$ is the Bessel function of order ν. For a derivation of the real part, see Stein (1999, section 2.10) or Exercises 14.1–14.5. In applications, t is replaced with t/ρ for some temporal range ρ.

14.4 Stationary spatial process

14.4.1 Spectral decomposition

We assume in this section that the domain is $\mathbb{R}^d$ for some $d \ge 1$. In most examples, the domain is also assumed to be Euclidean with an inner product and a norm, so that the orthogonal group may act on it. To distinguish space from time, particularly in the case $d = 1$, points in the domain are called sites and are denoted by x.

The frequency vector $\omega = (\omega_1, \ldots, \omega_d)$ is a linear functional on the domain, so that the scalar product $\omega x \equiv \omega' x$ is the value at x. The Hermitian function

$$K_\omega(x, x') = e^{i\omega(x-x')},$$

which is the Hermitian outer product $\xi\xi^*$ of the function $\xi(x) = e^{i\omega x}$ with itself, is positive definite of rank one. For each non-negative measure μ on frequencies, the convex combination

$$K_\mu(x, x') = \int_{-\infty}^{\infty} e^{i\omega(x-x')}\, d\mu(\omega)$$

is positive definite Hermitian. As a function of $x - x'$, it is necessarily the covariance of a stationary Gaussian process. Moreover, every stationary Gaussian process has an associated spectral measure.

We decompose the measure into symmetric and alternating parts as defined in (14.1), so that

$$-1 \le \frac{d\mu_{\rm alt}}{d\mu_{\rm sym}}(\omega) = -\frac{d\mu_{\rm alt}}{d\mu_{\rm sym}}(-\omega) \le 1.$$

Since $\mu_{\rm sym}$ is even and $\mu_{\rm alt}$ is odd, the associated covariance function is

$$\begin{aligned} K_\mu(x) &= \int_{-\infty}^{\infty} \cos(\omega x)\, d\mu_{\rm sym}(\omega) + i \int_{-\infty}^{\infty} \sin(\omega x)\, d\mu_{\rm alt}(\omega) \\ &= K_\mu^{\rm sym}(x) + i K_\mu^{\rm alt}(x), \end{aligned}$$

where x is the spatial difference vector for two sites. By construction $K_\mu^{\rm sym}(-x) = K_\mu^{\rm sym}(x)$ is even, whereas $K_\mu^{\rm alt}(-x) = -K_\mu^{\rm alt}(x)$ is odd.

For a simple illustrative example, the first-order Taylor expansion about $a = 0$ of the shifted Matérn measure

$$\begin{aligned} \frac{d\omega}{(1 + \|x - a\|^2)^{d/2+\nu}} &= \frac{d\omega}{(1 + \|a\|^2 + \|\omega\|^2 - 2a\omega)^{d/2+\nu}} \\ &= \frac{d\omega}{(1 + \|\omega\|^2)^{d/2+\nu}} + \frac{(d/2+\nu) 2a\omega\, d\omega}{(1 + \|\omega\|^2)^{d/2+\nu+1}} + o(\|a\|) \end{aligned}$$

is a $\mu_{\rm sym} + \mu_{\rm alt}$ decomposition in which $\mu_{\rm sym}$ is also orthogonally invariant.

14.4.2 Matérn spatial class

For general $\nu > 0$, the Matérn spectral measure on $\mathbb{R}^d$ is finite and radially symmetric with density

$$d\mu_{\rm sym}(\omega) = \frac{\Gamma(\nu + d/2)\, d\omega}{\pi^{d/2}(1 + \|\omega\|^2)^{\nu+d/2}}. \tag{14.4}$$

The Matérn covariance function $\|x\|^\nu \mathcal{K}_\nu(\|x\|)$ is real symmetric, and the associated Gaussian process is isotropic in $\mathbb{R}^d$.

To a certain extent, the choice of $\mu_{\rm alt}$ is arbitrary, but there is one mathematically natural choice

$$d\mu_{\rm alt}(\omega) = d\mu_{\rm sym}(\omega) \times \frac{2a\omega}{1 + \|\omega\|^2},$$

where $a\omega$ is the scalar product of Euclidean vectors. The skew perturbation is the stereographic projection from the unit sphere in $\mathbb{R}^{d+1}$ into $\mathbb{R}^d$ of the spherical harmonic of degree one having polar vector $a \in \mathbb{R}^d$ in the equatorial plane. The polar condition $\|a\| \le 1$ is required for positivity of the density on the sphere, so $|\mu_{\rm alt}| \le \mu_{\rm sym}$.

The covariance function for the symmetric spectral measure is the standard Matérn function

$$M_\nu(\|x - x'\|) \propto \|x - x'\|^\nu \mathcal{K}_\nu(\|x - x'\|),$$

where $\mathcal{K}_\nu$ is the Bessel function of order ν. The Matérn model is isotropic with strictly positive covariances at every distance. For the spectral measure $\mu_{\mathrm{sym}} + \mu_{\mathrm{alt}}$, the covariance function is

$$M_\nu(\|x - x'\|)\big(1 + ia(x - x')/(\nu + d/2)\big). \tag{14.5}$$

In addition to the index and the range parameter which is not shown, this covariance also depend linearly on the polar vector.

The effect of the polar asymmetry in (14.5) can be understood from the covariance of sums with spatial differences

$$\mathrm{cov}\big(Z(x) + Z(x'),\, \bar{Z}(x) - \bar{Z}(x')\big) = 2M_\nu(\|x - x'\|)\frac{ia(x' - x)}{\nu + d/2}$$

so that the absolute covariance is maximized by spatial differences $x' - x$ in the polar direction.

Domain restriction

The reason for choosing the dimension-dependent power in the denominator of (14.4) is partly to guarantee integrability, but that reason is not sufficient to explain this particular choice. The real reason is that the index $\nu + d/2$ ensures that the measures $\mu_{\mathrm{sym},d}$ for different spaces are mutually compatible in the sense that $\mu_{\mathrm{sym},d+1}(A \times \mathbb{R}) = \mu_{\mathrm{sym},d}(A)$ for every d and arbitrary subsets $A \subset \mathbb{R}^d$. See Exercises 14.3–14.5.

Compatibility of spectral measures is more a matter of convenience than logical necessity. It implies that if Z is a Matérn process with index ν on $\mathbb{R}^{d+1}$, the restriction of Z to a lower-dimensional affine subspace is also a Matérn process having the same index and range parameter.

The situation for μ_{alt} and the covariance function (14.5) is more complicated because the effect on the polar vector of the subspace restriction must also be taken into account. Since a acts as a linear functional on the domain, it is natural to associate with each subspace restriction the corresponding orthogonal projection. As they are written, the measures $\mu_{\mathrm{alt},d}$ are not mutually compatible in that sense. Compatibility is restored if we set a finite maximum for d and replace a with $a' = a/(\nu + d/2)$. Otherwise, we can regard a as an infinitesimal generator for a perturbation; see section 14.6.3.

14.4.3 Illustration by simulation

Figures 14.1 shows two independent simulations of the zero-mean complex-valued Gaussian process using the isotropic Matérn covariance function with

index $\nu = 1$. At each point x in the 50×50 grid, the arrow shows the magnitude and direction of the field $Z(x)$ at that point. In fact, the plotted value is not $Z(x)$ but the deviation of $Z(x)$ from the sample average. This was done in order to reduce visual clutter and to focus attention on spatial variability.

This process with $\nu = 1$ is relatively smooth with one continuous spatial derivative (a 2×2 partial derivative matrix), so that streamlines can be traced visually. Both simulations are on a 10×10 grid, with range parameter $\rho = 5$ in the first and $\rho = 1$ in the second. The large range parameter means that the most distant pairs are moderately highly correlated with correlation 0.14. In the second plot, the distant pairs are essentially independent. The first plot can be viewed as a five-fold magnification of a part of the second process.

The simulations for $\nu = 0.5$ in Fig. 14.2 are in the same format. They are considerably rougher, and the streamlines more ragged. The correlation for the most distant pairs in the first plot is 0.06.

In the isotropic case, the covariance function is real, which means that the real and imaginary components of the field are independent and identically distributed. In that respect, the visual impression may be misleading.

Four relatively smooth anisotropic zero-mean processes are illustrated in Fig. 14.3–14.4. These have covariance function (14.5) with $\nu = 1$. The range parameter is $\rho = 1$, so the values at more distant points are essentially independent. The polar vector is the unit vector $(\cos\theta, \sin\theta)$ with arguments $\theta = 0, \pi/6, \pi/3, \pi/2$. The nature of the anisotropy is not easy to discern from simulations.

14.5 Covariance products

14.5.1 Hadamard product

Let $Z = (Z_1, \ldots, Z_n)$ and $W = (W_1, \ldots, Z_n)$ be zero-mean independent Gaussian vectors in $\mathbb{C}^n$ with covariance matrices K and C respectively. Then the covariances of the products are

$$\operatorname{cov}(Z_r W_r, Z_s, W_s) = 0; \qquad \operatorname{cov}(Z_r W_r, \bar{Z}_s, \bar{W}_s) = K_{rs} C_{rs}.$$

The non-zero covariance is the component-wise product of the two covariance matrices. In the case of real-valued random variables, where there is no distinction between Z and $\bar{Z}$, the conjugated version prevails even if the distributions are non-Gaussian.

The preceding derivation is the simplest proof of Schur's product theorem, which states that the Hadamard product of positive-definite matrices is itself positive definite. The most immediate consequence for Gaussian processes is that the functional product $K(x, x')C(x, x')$ of two covariance functions *on the same space* is also a covariance function. In particular, $C = K$ implies that the squared function $K^2(x, x')$ is positive-definite Hermitian, and $C = \bar{K}$ implies that the squared modulus $|K|^2(x, x')$ is real symmetric and positive-definite.

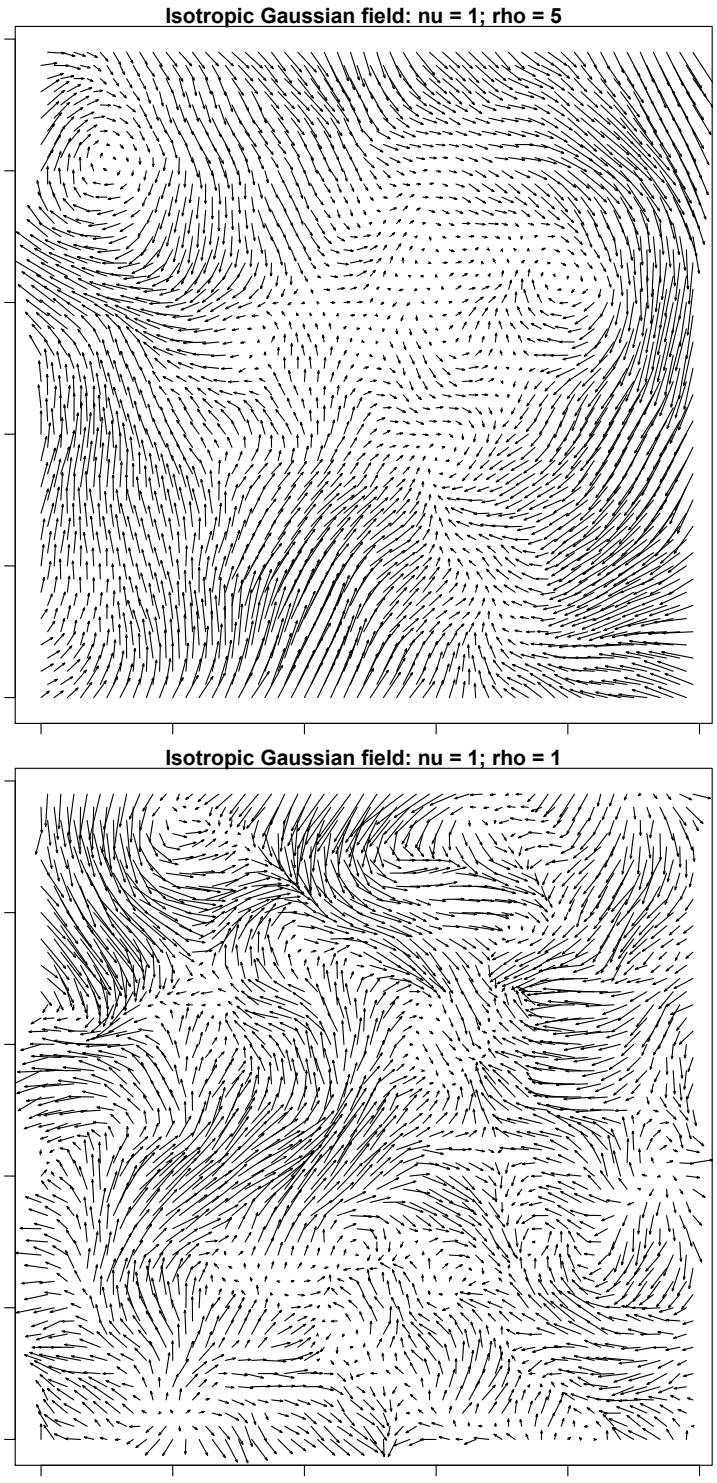

Figure 14.1: Two simulations of an isotropic complex Gaussian field

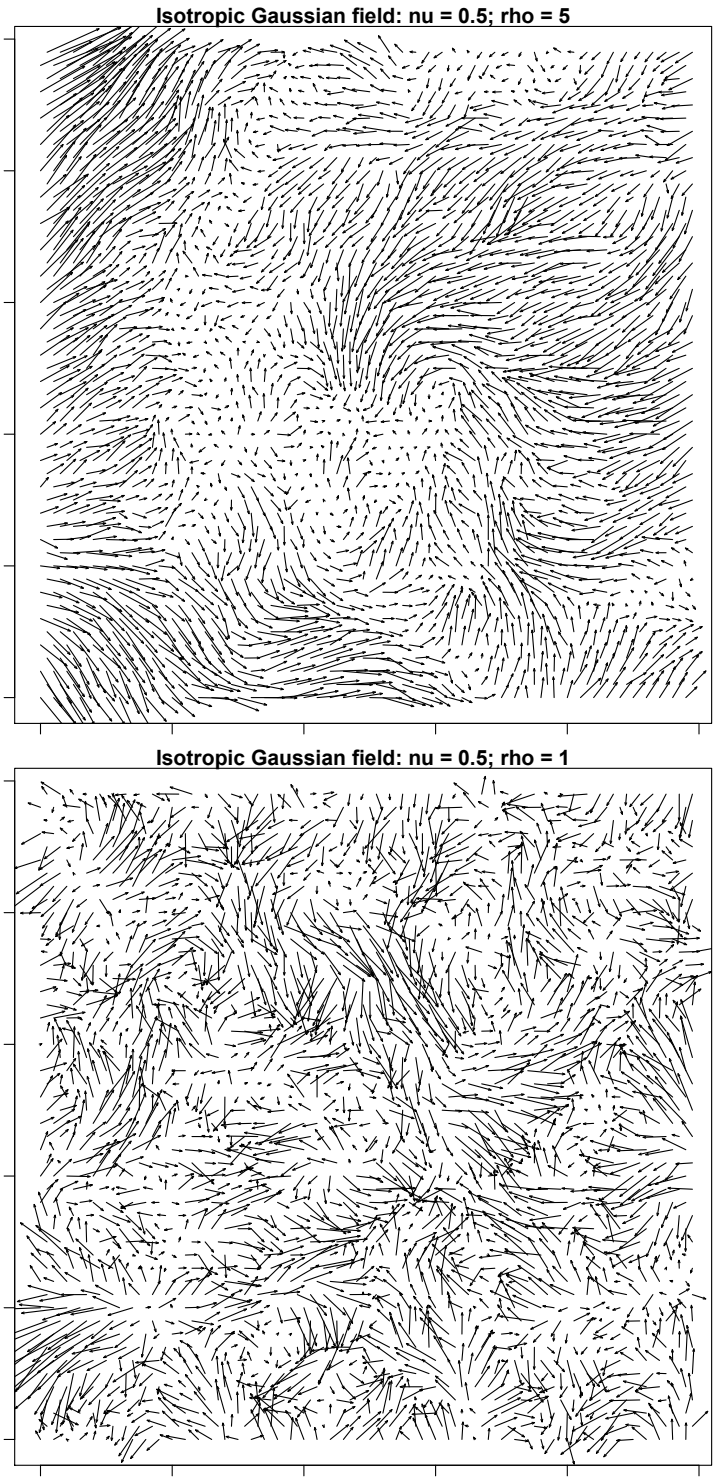

Figure 14.2: Two simulations of an isotropic complex Gaussian field

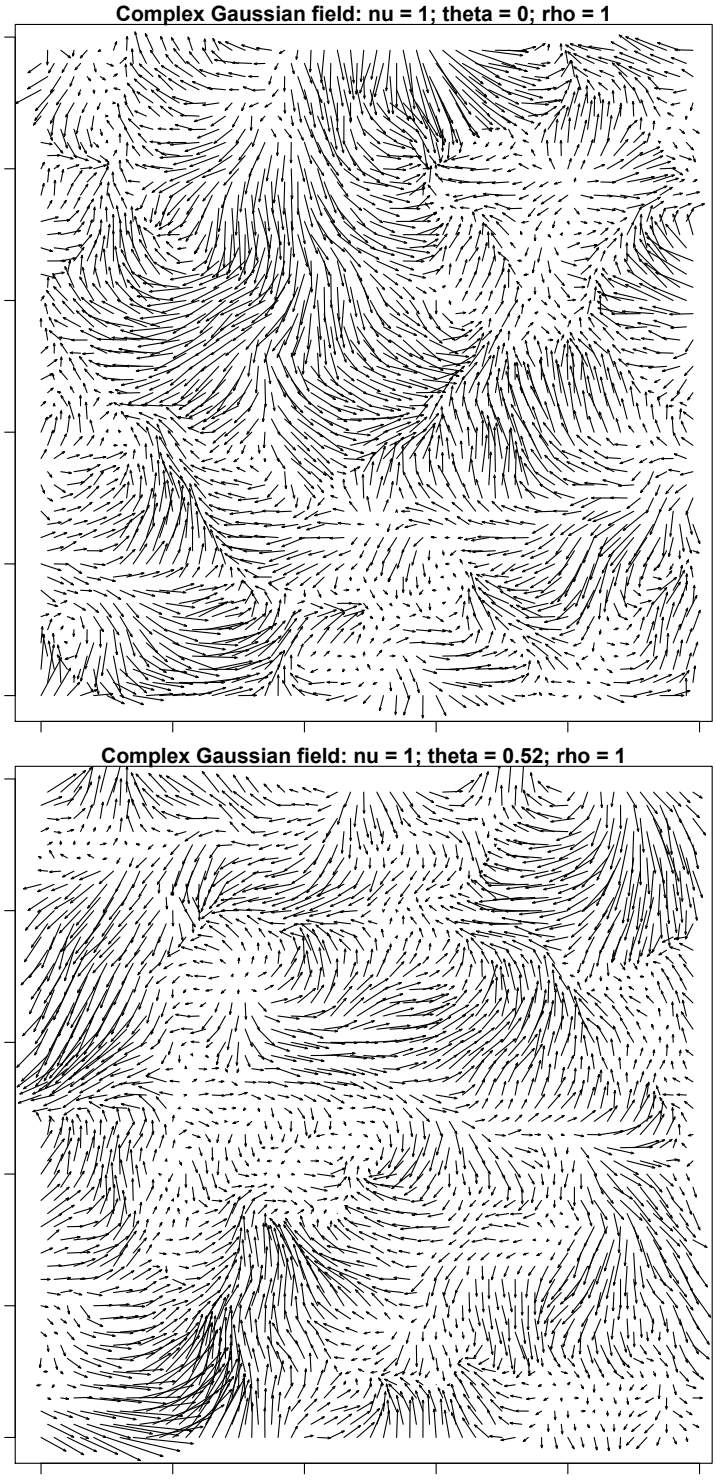

Figure 14.3: Two anisotropic Gaussian fields, one with $\theta = 0$ and one with $\theta = \pi/6$

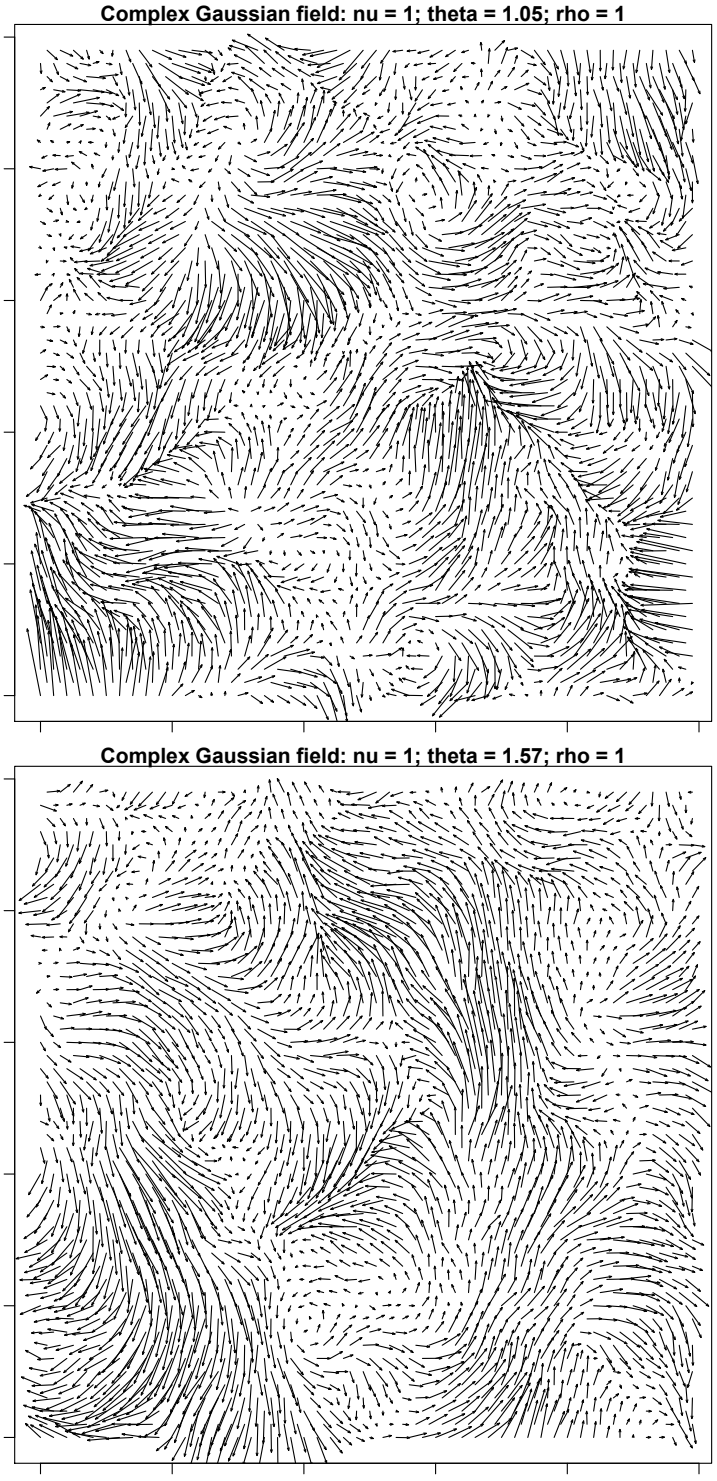

Figure 14.4: Two anisotropic Gaussian fields, one with $\theta = \pi/3$ and one with $\theta = \pi/2$

As an example, suppose that C and K are skew-Matern covariances (14.5), with polar vectors a and b respectively. The product is

$$M_\nu^2(\|x-x'\|)\Big(1+\frac{ia(x-x')}{\nu+d/2}\Big)\Big(1+\frac{ib(x-x')}{\nu+d/2}\Big),$$

which is positive-definite Hermitian on $\mathbb{R}^d$. The real part of the product is symmetric and positive definite:

$$M_\nu^2(\|x-x'\|)\bigg(1-\frac{Q(x-x')}{(\nu+d/2)^2}\bigg), \tag{14.6}$$

where $x \mapsto Q(x)$ is a rank-one quadratic form in x whose singular value $\|a\|\times\|b\|$ is less than one. The singular vectors are the unit polar vectors $a/\|a\|$ and $b/\|b\|$. Since a non-negative linear combination of positive-definite functions is positive definite, it follows that (14.6) is positive definite for every quadratic form Q whose nuclear norm satisfies $\|Q\|_* \le 1$. The nuclear norm is the sum of the singular values.

In the rank-one case, the trace of Q is $\sum a_i b_i$, which is the scalar product of the polar vectors. Thus, if the polar vectors come in mutually orthogonal pairs, Q is trace-free, which means that $x \mapsto Q(x)$ is a spherical harmonic of degree two.

14.5.2 Separable products and tensor products

Let $K(u,u')$ be a positive definite function on $\mathcal{U}$, and $C(v,v')$ a positive definite function on $\mathcal{V}$. Let the eigenvalues and eigenfunctions of K be $\{\lambda_i, \xi_i(u)\}$, so that $\int_{\mathcal{U}} K(u,u')\xi_i(u')\,du = \lambda_i\xi_i(u)$. Likewise, on $\mathcal{V}$, let the eigenvalues and vectors of C be $\{\rho_j, \zeta_j(v)\}$.

The product space $\mathcal{U}\times\mathcal{V}$ consists of ordered pairs (u,v), and the covariance product

$$K_2((u,v),(u',v')) = K(u,u')\,C(v,v') \tag{14.7}$$

is a natural candidate for a covariance function on the product space. An elementary calculation shows that $\xi_i(u)\zeta_j(v)$ is an eigenfunction of the product:

$$\begin{aligned}\int K(u,u')C(v,v')\xi_i(u')\zeta_j(v')\,du'\,dv' &= \int_{\mathcal{U}} K(u,u')\xi_i(u')\,du' \int_{\mathcal{V}} C(v,v')\zeta_j(v')\,dv' \\ &= \lambda_i\rho_j\,\xi_i(u)\zeta_j(v).\end{aligned}$$

Thus, the eigenvalues of the product are the products of the eigenvalues. Hence the covariance product is positive definite on the product space, and the rank of the product is the product of the ranks.

One important special case occurs when the spaces $\mathcal{U}$ and $\mathcal{V}$ are equal. The covariance product (14.7) restricted to the diagonal of $\mathcal{U}\times\mathcal{U}$ is nothing more than the Hadamard product of two covariance functions on $\mathcal{U}$. Positive definiteness follows trivially from the definition.

A covariance function on the product space is said to be *separable* if it is expressible as a single product, as in (14.7). A statistical covariance model is said to be separable if each covariance function in the model is separable. For example, the space-time covariance model consisting of the infinite set of covariance functions

$$\{\sigma^2 e^{-|t-t'|/\lambda}\, K_\mu(x,x') \colon \sigma^2, \lambda, \nu > 0; \|a\| \le 1\}$$

for K_μ in (14.5), is space-time separable.

A tensor product is a linear combination of pairwise products taken from two basis sets, say K_0, K_1 on $\mathcal{U}$ and C_0, C_1, C_2 on $\mathcal{V}$, and the tensor product space is the set of all such combinations. Most statistical applications employ tensor products in this form as models for covariances, with constraints on the coefficients to ensure positive definiteness. A tensor product is typically not separable.

The matrix formed by restriction of the separable product to a finite product grid is the Kronecker product of the marginal restrictions, and the inverse of a Kronecker product is the Kronecker product of the inverses. This fact makes for enormous simplification of statistical calculations related to Gaussian estimation and prediction, and that computational simplicity is the chief attraction of separable covariance models.

Separable covariances have a deservedly poor reputation in applied work for several reasons including the following. Suppose that a spatio-temporal Gaussian process Z is observed at a collection of sites $x_1, \ldots, x_n$ at times $t_1 < \cdots < t_k$, and it is required to predict the values at the same sites at a later time t_{k+1}. Suppose that the covariance function is given and separable. Then the conditional expected value of $Z(x_i, t_{k+1})$ given the data is a linear function of previous values for this site only. See exercises 14.?–? for an outline of a proof. In other words, separability implies that values observed at other nearby sites are irrelevant for prediction in this regular grid-like sampling scheme. This consequence of separability is unacceptable for almost any naturally-occurring process.

14.6 Real spatio-temporal process

14.6.1 Covariance products

The simplest way to construct a positive-definite covariance on a product space is to begin with a covariance function or set of covariance functions on each space, and to use tensor products. Usually a single product is not sufficient for applied work. Elementary examples can be found in (5.4).

For purposes of illustration, we use the temporal family (14.3) and the spatial family (14.5) with the same index. Writing x and t in place of $x - x'$ and $t - t'$, the outer product is

$$M_\nu(\|x\|)\big(1 + iax/(\nu + d/2)\big) \times M_\nu(t)\big(1 + ibt/(\nu + 1/2)\big), \tag{14.8}$$

where $-1 \leq b \leq 1$ is a scalar. This product splits into four sub-products, two real and two imaginary:

$$\begin{aligned}
&M_\nu(\|x\|)\, M_\nu(t);\\
&M_\nu(\|x\|)\, M_\nu(t) \times iax/(\nu + d/2);\\
&M_\nu(\|x\|)\, M_\nu(t) \times ibt/(\nu + 1/2);\\
&M_\nu(\|x\|)\, M_\nu(t) \times -ax\, bt/((\nu + 1/2)(\nu + d/2)).
\end{aligned} \tag{14.9}$$

For a real spatio-temporal process, the two imaginary terms can be discarded. We are left with a linear combination of the two real products,

$$M_\nu(\|x\|)\, M_\nu(t)\left(1 - \frac{ax\, bt}{(\nu + 1/2)(\nu + d/2)}\right). \tag{14.10}$$

Only the first of these is positive definite. Nevertheless, the linear combination is positive definite for all $\|a\| \leq 1$.

The complex temporal process associated with (14.3) is stationary but not reversible. The complex spatial process associated with (14.5) is stationary, but the polar parameter implies a specific directional asymmetry. By contrast, the real space-time process with covariance (14.10) is spatially isotropic at every time, and it is temporally reversible at every site. However, the space-time process is neither isotropic nor time-reversible. The fourth product in (14.9) is the interaction of one temporal asymmetry with one spatial asymmetry. It implies that the real Gaussian process with covariance (14.10) has the same distribution as the time-reversed process with polar vector $-a$.

Since b is a real number, it can be ignored or absorbed into other factors. The product (14.9) is linear in $x_1, \ldots, x_k$ with coefficients proportional to the polar coefficients $a_1, \ldots, a_k$. It can therefore be expressed as a linear combination of k similar terms, with coefficients to be estimated from the data.

14.6.2 Travelling wave

Any continuous group acting on the spatial domain $g\colon x \mapsto gx$ can be made to act on the space-time product space, either in a trivial way $(x,t) \mapsto (gx,t)$ or in a non-trivial way. The simplest non-trivial action is that of a wave travelling with constant velocity $g \in \mathbb{R}^d$, so that the group action $(x,t) \mapsto (x - gt, t)$ is a spatial shift proportional to time. Such a transformation on the domain induces a transformation on the process, sending $Z(x,t)$ to $W(x,t) = Z(x - gt, t)$ by composition. Positive-definiteness of the composite covariance function follows automatically from the definition.

If the original process Z has a separable covariance function $K(x,x')C(t,t')$, the transformed covariance function

$$\begin{aligned}
\operatorname{cov}\big(W(x,t), \bar{W}(x',t')\big) &= \operatorname{cov}\big(Z(x - gt, t), \bar{Z}(x' - gt', t')\big)\\
&= K(x - gt, x' - gt')\, C(t,t')
\end{aligned}$$

is a $K \times C$ product, but it is not a space-time separable product. At time t, the process has a random spatial profile whose covariance function is

$$\operatorname{cov}\big(W(x,t), \bar{W}(x',t)\big) = K(x - gt, x' - gt)\, C(t,t).$$

Assuming that K and C are both stationary, the spatial profile is a Gaussian process with covariance $K(x-x')\,C(0)$, which is stationary in space and constant in time. Although the spatial distribution is constant in time, the profile itself is not static unless the spatial factor is constant $C(t) = C(0)$.

As a specific example consider a complex-valued Matérn process in $\mathbb{R}^d$ with polar vector a and covariance function (14.5). The covariance function for the associated wave travelling at velocity v is

$$M_\nu(\|x - x' - v(t-t')\|)\left(1 + \frac{ia(x - x' - v(t-t')}{\nu + d/2}\right). \tag{14.11}$$

Using the temporal covariance (14.3) with $b = 1$, and writing x, t in place of $x - x'$ and $t - t'$, the four components of the product are obtained by replacing x with $x - vt$ in (14.10):

$$M_\nu(\|x - vt\|)\, M_\nu(t); \tag{14.12}$$
$$M_\nu(\|x - vt\|)\, M_\nu(t) \times ia(x - vt)/(\nu + d/2);$$
$$M_\nu(\|x - vt\|)\, M_\nu(t) \times it/(\nu + 1/2);$$
$$M_\nu(\|x - vt\|)\, M_\nu(t) \times \frac{-axt + avt^2}{(\nu + 1/2)(\nu + d/2)}. \tag{14.13}$$

Since we are interested in real-valued spatio-temporal processes, we focus on the two real parts whose sum is automatically positive definite. Note that ax is the scalar product of the polar vector with the spatial displacement x, and av is the scalar product with the velocity vector.

For application to fluid dynamics, other groups may be relevant, in particular the group of rigid Euclidean motions in $\mathbb{R}^d$, which allows the wave to rotate as it travels. It is convenient to illustrate the idea for $d = 2$ by regarding $\mathbb{R}^2 \cong \mathbb{C}$ as the spatial domain, so that the group element $g = (\theta, v)$ acts as a rigid Euclidean motion on the space-time domain by

$$g\colon (x,t) \mapsto (e^{i\theta t}x - vt,\ t).$$

The group element has two components, $\theta \in [0, 2\pi)$ or $\mathbb{R} \pmod{2\pi}$, and $v \in \mathbb{C}$. The covariance function for the transformed process is obtained by substituting $e^{i\theta t}x - vt$ for $x - vt$ in (14.11) and (14.9). If the polar vector a is also taken as a complex number, ax is the real part of the complex product $a\bar{x}$, not the product of complex numbers.

The concept of a flowing compressible fluid is of central importance in partial differential-equation models governing atmospheric physics and fluid mechanics more generally. Matter, heat and electrical particles are transferred by fluid flow, a phenomenon known as advection. It is only natural to think of v as an

advection vector in (14.11). This mechanical portrayal conveys a vivid image but the picture may be seriously misleading. Although some waves do travel in a fluid, there is no essential connection between the fluid velocity and the wave velocity. As the following example demonstrates, the wave velocity may be substantially greater than the fluid velocity—and not necessarily in the same direction.

14.6.3 Perturbation theory

One way to understand the skew-symmetric contribution in section 14.4 is to view the parameter a (or $a/(\nu+d/2)$) as a small perturbation of the Matérn spectral measure μ_{sym}. In that case, we can approximate (14.5) by a perturbation generated by the group element e^{iax} acting on the process. As always, $x \mapsto ax$ is the scalar product of vectors in $\mathbb{R}^d$. In other words, if Z is an isotropic process with covariance $M_\nu(\|x - x'\|)$, the covariance of the non-isotropic perturbation $W(x) = e^{iax}Z(x)$ is

$$\begin{aligned}\operatorname{cov}\big(W(x), \bar{W}(x')\big) &= M_\nu(\|x - x'\|)e^{ia(x-x')} \\ &= M_\nu(\|x - x'\|)\big(1 + ia(x - x') + o(\|a\|)\big).\end{aligned}$$

By construction, this function is positive definite for all vectors a; it is an approximation to (14.5) only for small a.

If we also regard the temporal covariance (14.3) as a multiplicative perturbation with parameter b, we arrive at a multiplicative perturbation of the covariance product in the form

$$e^{ia(x-x')+ib(t-t')} = e^{ia(x-vt-(x'-vt'))},$$

with $v \in \mathbb{R}^d$ as velocity vector and $b = -av$ as the scalar product. The perturbation-theory covariance function

$$M_\nu(\|x - x'\|)\, M_\nu(t - t')e^{ia(x-vt-(x'-vt'))}$$

has some of the characteristics of a travelling wave, but it is not the same as (14.11).

14.7 Summer cloud cover in Illinois

Figure 14.5 illustrates the fractional cloud cover on a 15×15 grid of points in central Illinois with 0.2 degrees separation in latitude and longitude, which implies 17.0×22.2 km cells. Successive panels show the cloud cover at 30-minute intervals from 6.00am to noon on June 9, 1998.

Solar irradiance is measured by a geostationary satellite, and fractional cloud cover is the complement of the ratio of solar irradiance relative to the clear-sky maximum at that time and location. The value lies between zero and one. On this day, the average fractional cloud cover was 35%. Cloud cover is

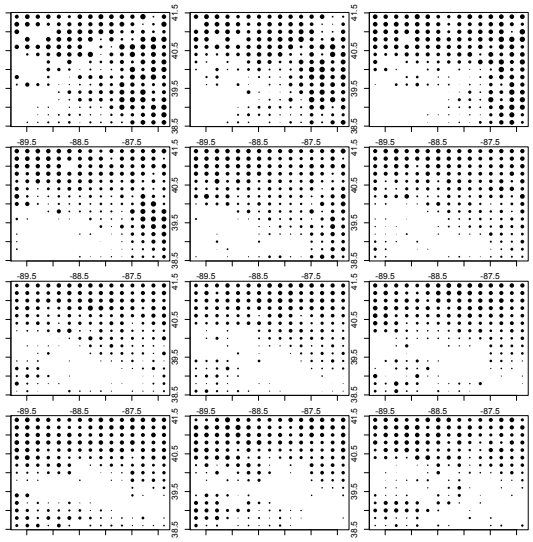

Figure 14.5: Fractional cloud cover on a 15×15 spatial grid in central Illinois in half-hour intervals from 6.00am to 11.30am on June 9, 1998

the primary variable that limits the production of solar energy. Its evolution throughout the day is of commercial interest for short-term prediction of solar electrical generating capacity, so that alternative sources may be brought online if needed.

For this illustration, only the first 2.5 hours of data from 6.00am to 8.30am are used for parameter estimation and model fitting. This corresponds to the top six panels in Fig. 14.5. The process appears to be relatively smooth in space and in time, so we use the Matérn model (14.10) with $\nu = 1$ for both space and time. Two range parameters, ρ_0 for time in minutes, and ρ_1 for distance in km. are also needed, so t is replaced by t/ρ_0 and x by x/ρ_1. The isotropic submodel has $a = 0$. The maximally anisotropic model takes $b = 1$ and advection $a = (\cos\theta, \sin\theta)$ as a unit vector in the east and north directions respectively.

For both the isotropic and the anisotropic covariance models, the mean fractional cloud cover is taken to be linear in both space and time. This may be adequate for short-term prediction, but it is not recommended for long-term prediction or extrapolation beyond the spatial domain. As always, a nugget term is included in the covariance model. The fitted range parameters for the isotropic model are 32.5 minutes and 27.0 km, while the variance components are 0.0219 for the identity matrix and 0.0283 for the Matérn product covariance. For the anisotropic model (14.10) using the unit vector $a = (\cos\theta, \sin\theta)$ with $\hat{\theta} = 3.01$, the fitted range parameters are 32.0 minutes and 26.0 km, while the variance components are 0.0215 and 0.0276.

The REML log likelihood for the fitted anisotropic model is 11.23 units higher than that for the isotropic model. Since the anisotropic model has two

Table 14.1. Summary of fitted parameters for four space-time models

Parameter	Isotropic	(14.10)	(14.12)	+(14.13)
Spatial range (km)	27.0	26.0	35.0	35.5
Temporal range (min)	32.5	32.0	61.0	62.0
Nugget variance	0.0219	0.0215	0.0240	0.0238
Matérn variance	0.0283	0.0276	0.028	0.0298
Wave speed (km/min)	0.0	0.0	0.70	0.70
Wave direction θ	—	—	0.033	0.00
Polar norm $\|a\|$	0.0	1.0	0.0	0.67
Polar direction ϕ	—	3.01	—	0.63
Log likelihood	4.724	15.954	23.522	24.920
RMSE 9.00am	0.140	0.131		

Table 14.1: Summary of fitted parameters for four space-time models

additional parameters, both $\|a\|$ and θ, the likelihood-ratio test statistic is nominally on two degrees of freedom. In fact, $\|\hat{a}\| = 1$ on the boundary, which slightly complicates the null distribution theory. Nevertheless, the observed likelihood-ratio test statistic of 22.46 leaves no doubt about the existence of space-time anisotropy for the cloud-cover process. Whether the formulation (14.10) captures adequately the full extent of anisotropy is another matter. Most likely, the polar vector could not be expected to remain constant from one day to the next.

Table 14.1 shows the fitted parameters for four spatial models, all including a nugget effect. Each of the anisotropic models is a substantial improvement over the isotropic Matérn product. The simplest travelling wave model (14.12) is the most effective; the additional polar anisotropy in (14.9) does not substantially improve the fit.

In both travelling-wave models, the estimated wave velocity is approximately 0.7 km/min, or 42 km/hr, or 26 mph from the east. However, that particular June morning was calm and humid, with light and variable winds averaging four mph. In such circumstances, a wave travelling at 42 km/hr in any direction might be attributed to changes in temperature or pressure, but it cannot be attributed to atmospheric advection.

The large value of the nugget variance relative to the spatio-temporal variance implies that even the best predictor has a substantial variance. The nugget standard error is a lower bound for the root mean square prediction error, and the fitted values are 0.148 for the isotropic model, and 0.147 for the anisotropic model (14.10). The empirical one-step-ahead root mean square prediction error averaged over 225 sites for 9.00am are 0.140 for the isotropic model and 0.131 for the anisotropic model (14.10).

14.8 More on Gaussian processes

14.8.1 White noise

An alert reader may have noticed that the definition of a Gaussian process in section 14.1, and the definitions of stationarity and isotropy in sections 14.2–3, are not sufficiently broad to include the simplest non-trivial Gaussian processes on the real line or on the plane. On an arbitrary domain with measure Λ, white noise is a zero-mean Gaussian process indexed by subsets such that, $\mathrm{cov}(W(A), W(B)) = \Lambda(A \cap B)$. The process takes independent values on disjoint sets, and variances are determined by the intensity measure. If Λ is Lebesgue measure on the real line or $\mathbb{R}^d$, which is the standard choice for those domains, the process is both stationary and isotropic. The notation here and subsequently in this section presumes that the process is real-valued.

The earlier definition is inadequate because it assumes that the domain and the index set are one and the same set. This is sufficient for processes defined pointwise, but it is not sufficient to cover many of the generalized or intrinsic processes that occur in applied work. For planar white noise, the domain is $\mathcal{D} = \mathbb{R}^2$ or $\mathbb{C}$, but the index set $\mathcal{U}$ is the set of Borel subsets in the domain. More correctly, $\mathcal{U}$ is the proper subset consisting of Borel sets of finite Λ-measure. The definition of stationarity offered in section 14.2.1 is not applicable to white noise because it presumes that $W(x)$ exists.

Stationarity and isotropy refer to a group acting on the domain $x \mapsto gx$ either by translation or rotation. There is a natural induced action on the index set $A \mapsto gA$, which is a rigid Euclidean motion of subsets. With an appropriate modification to distinguish between the domain and the index set, white-noise with intensity Λ is stationary or isotropic if the measure is invariant under this action.

Every process Z that is defined pointwise and is continuous on the domain can be extended by integration to an additive process W on domain subsets

$$W(A) = \int_A Z(x)\, d\Lambda(x).$$

Additivity for disjoint subsets means $W(A \cup B) = W(A) + W(B)$. The covariance function $K(x, x')$ of Z is the covariance density of W

$$\mathrm{cov}\big(W(A), W(B)\big) = \int_{A\times B} K(x, x')\, d\Lambda(x)\, d\Lambda(x').$$

White noise is not a continuous process and does not have a covariance density. However, $\mathrm{cov}(W(A), W(B)) = \Lambda(A \cap B)$ means that there is a covariance measure, which is the Dirac-type singular measure $\Lambda(\cdot)$ concentrated on the diagonal in $\mathcal{D}^2$.

The extension to subsets is a half-way house that suffices for a few purposes, but it is not adequate for mathematical work, which requires all variances to be

finite, and it is not entirely adequate even for applied work. Consider, for example, the additive planar process defined for regular planar subsets as follows:

$$\operatorname{cov}\bigl(W(A), W(B)\bigr) = \begin{cases} \Lambda_1(\partial A \cap \partial B), & \operatorname{Int}(A) \subset \operatorname{Int}(B) \text{ or } \operatorname{Int}(B) \subset \operatorname{Int}(A); \\ -\Lambda_1(\partial A \cap \partial B), & \operatorname{Int}(A) \cap \operatorname{Int}(B) = \emptyset. \end{cases}$$

Regular means that each planar subset A has a well-defined interior $\operatorname{Int}(A)$, and a one-dimensional boundary ∂A of finite length $\Lambda_1(\partial A)$. Additivity implies that the variance for more general regions is the boundary length, and the covariance for two regions is the total signed length of the common boundary.

It is not clear from the preceding description how we are meant to deal with a subset having an irregular boundary or an empty interior, so this specification is not entirely satisfactory. It is also unclear whether a covariance measure exists. The more satisfactory way to study such processes is to abandon subsets and to use a suitable Hilbert space as the index set. Planar white noise is associated with the space of square-integrable functions $f\colon \mathbb{R}^2 \to \mathbb{R}$; a subset is nothing more than its indicator function. The process described above is associated with functions such that the norm of the derivative vector is square-integrable $\int \|f'(x)\|^2 \, d\Lambda(x) < \infty$.

14.8.2 Limit processes

This section deals with questions of two types, the first related to limits of Gaussian processes, the second related to prediction and limits of conditional distributions.

Two families of processes are used to illustrate the development. Both are indexed by a single parameter $\theta > 0$, and the limit refers to $\theta \to \infty$. The first is an exchangeable Gaussian process in which the covariance function is

$$\operatorname{cov}(Z_i, Z_j) = \delta_{ij} + \theta, \tag{14.14}$$

which means that the covariance matrx of $Z[n]$ is $I_n + \theta J_n$. The second is a Matérn-1/2 process defined pointwise on $\mathbb{R}^d$ with covariance function

$$\operatorname{cov}\bigl(Z(x), Z(x')\bigr) = \theta e^{-\|x - x'\|/\theta}. \tag{14.15}$$

Here, θ is the range parameter, and the limit $\theta \to \infty$ addresses long-range dependence. Inessential scalar multiples, which would almost always occur in applied work, are disregarded.

Each question gives rise to a number of subsidiary questions along the following lines:

1. Does the limit process exist? (N). If not, does any limit process exist? (Y). If a non-trivial limit exists, exhibit the index set, the covariance function, and so on.

2. For fixed n, is there an invertible normalization that produces a limit distribution? (Y).

3. What is the conditional distribution of $Z(x_0)$ given $Z(x_1), \ldots, Z(x_n)$? Does the conditional distribution have a limit as $\theta \to \infty$? (Y).

4. Can the limiting conditional distribution be obtained from the limit process? (N). Can the limiting conditional distribution be obtained from the limit in part 2? (Y).

5. Can the best linear predictor (BLP) of Z_{n+1} be obtained from the limit process? (Y). Can the conditional expected value of Z_{n+1} given $Z[n]$ be computed from the limit process? (N).

The two examples are generic, so the answers indicated in parentheses apply equally to both.

Existence of a limit process

In the first example $Z_1 \sim N(0, 1+\theta)$, and in the second $Z(x) \sim N(0, \theta)$. Neither sequence of distributions has a limit as $\theta \to \infty$, so the answer to the first question is negative. On the other hand, $Z_i - Z_j \sim N(0, 2)$ for $i \neq j$, while $Z(x) - Z(x')$ has variance

$$2\theta - 2\theta e^{-\|x-x'\|/\theta} = 2\|x - x'\| + O(\theta^{-1}).$$

Both limits exist and the distributions are Gaussian. More generally, for $n \geq 1$ and any coefficient vector $\alpha = (\alpha_1, \ldots, \alpha_n)$ whose components add to zero, the linear combination $\sum \alpha_j Z_j$ is Gaussian with variance $\|\alpha\|^2 = \sum \alpha_j^2$, independent of θ. In the case of the Matérn model, $\sum \alpha_j Z(x_j)$ is Gaussian with variance $\alpha' D\alpha + O(\theta^{-1})$, where $D_{ij} = -\|x_i - x_j\|$ is the $n \times n$ matrix of negative Euclidean distances. In both cases the limit exists for arbitrary contrasts.

The easiest way to characterize the preceding limit is to use the vector space of contrasts as the index set. For ease of exposition, suppose that the set of points under consideration is fixed and finite, so that a contrast $\alpha \in \mathbf{1}_n^0$ is a vector whose components add to zero. The value at contrast α is $W(\alpha) = \alpha_1 Z_1 + \cdots + \alpha_n Z_n$ for (14.14), or $\sum \alpha_r Z(x_r)$ in the case of the Matérn model. The limiting covariance of two contrasts is the Hilbert-space inner product

$$\operatorname{cov}\big(W(\alpha), W(\beta)\big) = \langle \alpha, \beta \rangle = \begin{cases} \sum_r \alpha_r \beta_r & (14.14); \\ -\sum_{r,s} \alpha_r \beta_s \|x_r - x_s\| & (14.15). \end{cases}$$

The Hilbert space $\mathcal{H}_n^*$ has dimension $n-1$, but it is exhibited here as the subspace $\mathbf{1}_n^0$ of contrasts in a vector space of dimension n. Consequently, the inner-product matrix is of order n, and is not unique. For the first example, we could use either the identity matrix of order n or $I_n - J_n/n$.

Since every n-contrast is also a $(n+1)$-contrast whose last component is zero, the Hilbert-space $\mathcal{H}_n^*$ of n-contrasts is a subspace of H_{n+1}^*. Kolmogorov consistency is automatic, but is equivalent to the statement that the restriction or insertion $\mathcal{H}_n^* \hookrightarrow H_{n+1}^*$ is an isometry. In effect, $\mathcal{H}_\infty^*$ includes all of the finite-dimensional spaces as subspaces. In fact, the restriction to contrasts determines a consistent process, not only in the limit, but for every θ.

An equivalent way of saying the same thing is that the process for finite θ is defined conventionally for finite samples as a probability distribution $P_{n,\theta}$ on the space $\mathcal{B}(\mathbb{R}^n)$ of Borel subsets in $\mathbb{R}^n$. For any event $A \subset \mathbb{R}^n$ such that $A + \mathbf{1}_n = A$, the value assigned by $P_{n,\theta}$ to A has a limit $P_{n,\infty}(A)$, which a Gaussian probability. These translation-invariant events $A \in \mathcal{B}(\mathbb{R}^n/\mathbf{1}_n)$ are the only events to which the limit process assigns a probability. Thus, the non-trivial limit is obtained by restriction of the σ-field.

Existence of a limit distribution

In all examples of the type under consideration, the covariance matrix of $Z[n]$ is

$$\operatorname{cov}(Z[n]) = \theta J_n + \Sigma + O(\theta^{-1}) \tag{14.16}$$

where Σ is independent of θ and is also positive definite on contrasts. For the Matérn example, Σ is the matrix with components $-\|x_i - x_j\|$.

Let $P = J_n/n$ and $Q = I_n - P$ be complementary projections, so that

$$\operatorname{cov}\begin{pmatrix} \theta^{-1/2}PZ \\ QZ \end{pmatrix} = \begin{pmatrix} J_n + P\Sigma P/\theta & P\Sigma Q/\theta^{1/2} \\ Q\Sigma P/\theta^{1/2} & Q\Sigma Q \end{pmatrix} \to \begin{pmatrix} J_n & 0 \\ 0 & Q\Sigma Q \end{pmatrix}.$$

Consequently, the vector $W = \theta^{-1/2}PZ + QZ$ with components

$$W_i = \theta^{-1/2}\bar{Z}_n + (Z_i - \bar{Z}_n)$$

has a Gaussian limit distribution with covariance matrix $J_n + Q\Sigma Q'$. For each $n \geq 1$, the transformation $Z[n] \mapsto W$ is invertible, so the answer to part 2 is affirmative.

Note that $W \in \mathbb{R}^n$ is not the restriction of the corresponding transformation in $\mathbb{R}^{n+1}$, so there is no W-process associated with these transformations. The existence of a limit distribution for every n does not imply the existence of a limit process.

Limit of conditional distributions

The conditional distribution of $Z(x_0)$ given $Z(x_1, \ldots, Z(x_n)$ is Gaussian, so it is necessary only to compute the conditional mean and variance, and to observe the behaviour as $\theta \to \infty$. For the exchangeable model, the conditional mean and variance are

$$E\big(Z_{n+1} \mid Z[n]\big) = \frac{n\theta\bar{Z}_n}{1+n\theta}, \qquad \operatorname{var}\big(Z_{n+1} \mid Z[n]\big) = \frac{1+(n+1)\theta}{1+n\theta},$$

so the limit of the conditional distributions is $N(\bar{Z}_n, 1)$.

For the spatial model, the calculations for finite θ are a little more complicated, so it is necessary to take limits as the calculation progresses. Let $L = \theta^{-1/2}L_0 + L_1$ be the matrix of a linear transformation in $\mathbb{R}^{n+1}$

$$L_0 = \begin{pmatrix} P_n & 0 \\ 0 & 0 \end{pmatrix}, \qquad L_1 = \begin{pmatrix} Q_n & 0 \\ -\mathbf{1}_n/n & 1 \end{pmatrix},$$

where P_n and $Q_n = I_n - P_n$ are the complementary projections denoted by P, Q in the preceding section. From the representation (14.16), the covariance matrix of $W = LZ$ has a limit, which is the sum of two mutually orthogonal matrices

$$\begin{aligned}\operatorname{cov}(LZ) &= \theta L J_{n+1} L' + L\Sigma L' + O(\theta^{-1}) \\ &= \begin{pmatrix} J_n & 0 \\ 0 & 0 \end{pmatrix} + L_1 \Sigma L_1' + O(\theta^{-1}).\end{aligned}$$

For each n, it follows that $\theta^{-1/2}\bar{Z}_n$ has a standard normal limit as $\theta \to \infty$, and is asymptotically independent of every contrast $Z_i - \bar{Z}_n$, not only for $i \le n$, but also for $i = n+1$. Thus, the limiting conditional mean satisfies

$$E(Z_{n+1} - \bar{Z}_n \mid Z[n]) = \sum_{r=1}^{n} \beta_r Z_r, \tag{14.17}$$

$$E(Z_{n+1} \mid Z[n]) = \bar{Z}_n + \sum_{r=1}^{n} \beta_r Z_r, \tag{14.18}$$

where β is the orthogonal projection $H_{n+1}^* \to \mathcal{H}_n^*$ of the coeffficient vector $(-\mathbf{1}_n/n, 1)$ associated with the contrast $Z_{n+1} - \bar{Z}_n$. The linear combination (14.18), which is not a contrast, is often called the best linear predictor (BLP). The limiting conditional variance is the reciprocal of the last diagonal component of $(L_1 \Sigma L_1')^{-1}$.

Conditional distributions for the limit process

The situation regarding conditional distributions for the limit process is different in a fundamental but subtle way. The joint distribution for any set of contrasts is determined by the Hilbert-space inner product. In particular, the conditional distribution of the contrast $Z_{n+1} - \bar{Z}_n$ given the σ-field generated by $Z_1, \ldots, Z_n$ is Gaussian with mean (14.17) and variance as described above. However, the limit process is defined on contrasts only, so the σ-field generated by $Z_1, \ldots, Z_n$ is the σ-field generated by contrasts, which means that the coefficient vector β in (14.17) is a contrast in $\mathcal{H}_n^*$. The limit process does not admit either Z_{n+1} or $\bar{Z}_n$ as a Gaussian variable, so the crucial statement that $Z_{n+1} - \bar{Z}_n$ is independent of $\bar{Z}_n$ is either meaningless or mathematically trivial. In either case, the fiducial leap from (14.17) to (14.18) requires a σ-field extension, which cannot follow from the limit process alone.

Limit process as a Markov kernel

The limit process with probabilities defined on the σ-field generated by contrasts has a certain mathematical elegance—brutal and minimalist. But the σ-field restriction is a price too steep for any applied statistician interested in probabilistic prediction. Is there a way out, a way that retains the elegance of contrasts at a more affordable price? The answer, we hope, is yes.

A Markov kernel is a function that associates with each $\mu \in \mathbb{R}$ a Gaussian process Z such that, for each contrast $\alpha \in \mathbf{1}_n^0$, the increment or linear functional $\sum \alpha_r Z_r$ has the same distribution as that in the limit process. There is no σ-field restriction.

14.9 Exercises

14.1 By making the transformation $u = 1/(1+x^2)$ and converting to a beta-type integral on $(0,1)$, show that

$$2\int_0^\infty \frac{x^{d-1}\,dx}{(1+x^2)^{\nu+d/2}} = B(\nu, d/2) = \frac{\Gamma(\nu)\,\Gamma(d/2)}{\Gamma(\nu+d/2)},$$

where $B(\cdot,\cdot)$ is the beta function for strictly positive arguments.

14.2 The Matérn spectral measure on the real line is proportional to the symmetric type IV distribution in the Pearson class, which is also equivalent to the Student t family (Pearson type VII). For $\nu > -1/2$, show that the standardized version

$$M_1(d\omega) = \frac{\Gamma(\nu+1/2)\,d\omega}{\pi^{1/2}(1+\omega^2)^{\nu+1/2}}$$

has positive density, but the total mass is finite only for $\nu > 0$.

14.3 By transforming to spherical polar coordinates in $\mathbb{R}^d$, show that

$$\int_{\mathbb{R}^d} \frac{d\omega}{(1+\|\omega\|^2)^{\nu+d/2}} = A_{d-1}\int_0^\infty \frac{x^{d-1}\,dx}{(1+x^2)^{\nu+d/2}},$$

where $A_{d-1} = 2\pi^{d/2}/\Gamma(d/2)$ is the surface area of the unit sphere in $\mathbb{R}^d$. For $\nu > -d/2$, deduce that the Matérn measure on $\mathbb{R}^d$

$$M_d(d\omega) = \frac{\Gamma(\nu+d/2)\,d\omega}{\pi^{d/2}\,(1+\|\omega\|^2)^{\nu+d/2}}.$$

has finite mass if and only if $\nu > 0$. Show that the total mass is a constant independent of the dimension of the space.

14.4 For fixed $\nu > 0$, show that the Matérn measures are mutually consistent in the sense that $M_{d+1}(A \times \mathbb{R}) = M_d(A)$ for all $d \geq 0$ and subsets $A \subset \mathbb{R}^d$. In other words, show that M_d is the marginal distribution of M_{d+1} after integrating out the last component. For $\nu > -1$, show that the Matérn measures are mutually consistent in the sense that $M_{d+1}(A \times \mathbb{R}) = M_d(A)$ for all $d \geq 2$.

14.5 Consistency and finiteness together imply that the normalized Matérn measures define a real-valued process $X_1, X_2, \ldots$ in which $M_n/\Gamma(\nu)$ is the joint distribution of the finite sequence $X[n] = (X_1, \ldots, X_n)$. This process—a special case of the Gosset process—is not only exchangeable but also orthogonally invariant for every n. Show that the conditional distribution of X_{n+1} given $X[n]$ is Student t, with a certain location parameter, scale parameter and degrees of freedom. To what extent is finiteness needed in the construction of the process?

14.6 For the Matérn process, show that the sequence of partial averages $\bar{X}_n$ has a limit $\bar{X}_\infty = \lim_{n\to\infty} \bar{X}_n$. For $n \geq 2$, what can you say about the conditional distribution of $\bar{X}_\infty$ given $X[n]$? Consider separately the cases $\nu = 0$ and $\nu > 0$.

14.7 One definition of the Bessel-K function is the integral

$$\int_0^\infty \frac{\cos(\omega t)\, d\omega}{(1+\omega^2)^{\nu+1/2}} = \frac{\sqrt{\pi}}{2^\nu \Gamma(\nu+1/2)} \times |t|^\nu \mathcal{K}_\nu(t).$$

Deduce that $\mathcal{K}_\nu(\cdot)$ is symmetric and that

$$\lim_{t\to 0} |t|^\nu \mathcal{K}_\nu(|t|) = 2^{\nu-1}\,\Gamma(\nu).$$

14.8 For any linear functional $x\colon \mathbb{R}^d \to \mathbb{R}$, show that

$$\begin{aligned}\int_{\mathbb{R}^d} \frac{\cos(\omega x)\, d\omega}{(1+\|\omega\|^2)^{\nu+d/2}} &= I(\nu+1/2, d-1) \int_{\mathbb{R}} \frac{\cos(w\|x\|)\, dw}{(1+w^2)^{\nu+1/2}} \\ &= \frac{\pi^{d/2}}{2^{\nu-1}\Gamma(\nu+d/2)} \times \|x\|^\nu \mathcal{K}_\nu(\|x\|),\end{aligned}$$

where $I(\nu, d) = \pi^{d/2}\Gamma(\nu)/\Gamma(\nu+d/2)$.

14.9 Use integration by parts to show that

$$\begin{aligned}\int_{-\infty}^\infty \frac{\omega \sin(t\omega)\, d\omega}{(1+\omega^2)^{\nu+3/2}} &= \frac{t}{2\nu+1} \int_{-\infty}^\infty \frac{\cos(t\omega)\, d\omega}{(1+\omega^2)^{\nu+1/2}}, \\ &= \frac{\sqrt{\pi}}{2^{\nu-1}\Gamma(\nu+1/2)} \times \frac{t}{2\nu+2} |t|^\nu \mathcal{K}_\nu(t).\end{aligned}$$

Hence deduce that, for any pair of linear functionals $v, x\colon \mathbb{R}^d \to \mathbb{R}$,

$$\int_{\mathbb{R}^d} \frac{\omega v\, \sin(\omega x)\, d\omega}{(1+\|\omega\|^2)^{\nu+d/2+1}} = \frac{\pi^{d/2}}{2^{\nu-1}\Gamma(\nu+d/2)} \times \frac{vx}{2\nu+1} \|x\|^\nu \mathcal{K}_\nu(\|x\|),$$

where vx denotes the scalar product.

14.10 This exercise is concerned with stereographic projection from the unit sphere in $\mathbb{R}^{d+1}$ onto the equatorial plane $\mathbb{R}^d$. Latitude on the sphere is measured by the polar angle θ, starting from zero at the north pole, through $\theta = \pi/2$ at the equator up to $\theta = \pi$ at the south pole. Every point on the sphere is a pair $z = (e\sin\theta, \cos\theta)$ where e is a unit equatorial vector. The stereographic image of z is the point

$$\omega = e\cot(\theta/2) = e\cos(\theta/2)/\sin(\theta/2),$$

so that the southern hemisphere is projected into the unit ball, and the northern hemisphere to its complement in $\mathbb{R}^d$. Deduce that the stereographic image of the uniform spherical distribution is

$$\frac{\Gamma(d)}{\pi^{d/2}\Gamma(d/2)} \frac{d\omega}{(1+\|\omega\|^2)^d}$$

on the equatorial plane.

14.11 Points near the north pole are transformed stereographically to high frequencies, and points near the south pole to low frequencies. For $\nu > d/2$, the weighted distribution with density proportional to

$$|\sin(\theta/2)|^{2\nu-d}$$

reduces the mass on northern latitudes and increases that on southern latitudes, maintaining radial symmetry. Show that the stereographic image of the weighted distribution is inversely proportional to $(1 + \|\omega\|^2)^{\nu+d/2}$. Find the normalizing constants for both distributions.

14.12 For the special case $d = 2$, we may regard $\mathbb{R}^2 \cong \mathbb{C}$, so that ω is a complex number and e is a unit complex number. Show that the weighted spherical distribution with weight proportional to the degree k harmonic perturbation

$$1 + \Re(a\bar{e}^k)\sin^k\theta$$

is transformed to

$$\frac{\Gamma(d)}{\pi^{d/2}\Gamma(d/2)}\frac{d\omega}{(1+\|\omega\|^2)^d} \times \left(1 + \frac{2\Re(a\bar{\omega}^k)}{(1+|\omega|^2)^k}\right)$$

and is positive for $|a| \leq 1$.

14.13 Consider a fixed tessellation of the plane into a countable set of polygonal cells $A_1, \ldots$, and let $0 \leq \ell_{ij} < \infty$ be the length of the common boundary $\partial A_i \cap \partial A_j$. Associate with each ordered pair of regions (i, j) a Gaussian random variable

$$\varepsilon_{ij} = -\varepsilon_{ji} \sim N(0, \ell_{ij})$$

with iid signs independent of $|\varepsilon|$. If all boundary lengths $\ell_{i\bullet}$ are finite, the row sums $W(A_i) = \varepsilon_{i\bullet}$ define a Gaussian process indexed by cells. Find the covariances $\mathrm{cov}(W(A_i), \bar{W}(A_j))$ for $i = i$ and $i \neq j$.

14.14 In the setting of the previous exercise, let $W = L\varepsilon$, where L is a Boolean matrix. Show that W is a process defined on general planar regions and that it coincides with the process described at the end of section 14.8.1.

Chapter 15

Likelihood

15.1 Introduction

15.1.1 Non-Bayesian model

A non-Bayesian statistical model is a set of processes or a set of probability distributions $\{P_\theta\}$ on the sample space indexed by the points θ in the parameter space Θ. According to the standard paradigm, Nature chooses a point θ^* and generates the process $\{Y(x) : x \in \mathcal{D}\}$ on some domain according to the distribution P_{θ^*}. The observer chooses a fixed design or sample $\mathbf{x} = \{x_1, \ldots, x_n\}$ and observes or measures the sample values $Y[\mathbf{x}]$. The data consists of the design points $\mathbf{x}$, the values $Y[\mathbf{x}] \in \mathbb{R}^n$, and any other recorded baseline information.

Before the model can be used for inference, it is necessary to estimate the parameter from the data. The point estimator is a function $\hat{\theta}_n \colon \mathbb{R}^n \to \Theta$ from the observation space into the parameter space, defined for every adequately large design. The numerical value $\hat{\theta}_n(Y)$ determines a process $P_{\hat{\theta}}$, called the fitted process or bootstrap process, which serves as our best guess about what Nature might have been up to. If the goal is parametric inference, it is essential to quantify the estimation error, i.e., to quantify the magnitude of the difference $\hat{\theta}_n - \theta^*$ in some suitable sense. If the goal is not parametric, for example, the prediction of future values, it is necessary to compute the fitted conditional distribution

$$P_{n+m,\hat{\theta}}(A \mid Y[\mathbf{x}])$$

for events $A \subset \mathbb{R}^{n+m}$. In either case, the inferential goal requires not only a point estimate of the parameter, bu also some measure of its uncertainty and the effect of uncertainty on inferences.

The estimation step is sometimes called 'learning' in computer-science circles. But the parameter value is never learned with the certainty that that word implies; it is only estimated with error, which might be small or large. A large error in estimation does not necessarily lead to a large error in prediction. Generally speaking, parameters that are difficult to estimate have little effect

on single-point predictions that are local in a suitable sense.

In all cases considered in this book, Θ is a smooth manifold—at least locally near most points. There may be boundary points or or points of singularity. We say that the model is finite-dimensional if the dimension $\dim(\Theta) = p$ is finite. Otherwise the model is infinite-dimensional. The phrase *non-parametric model* is sometimes used in the literature as a synonym for *infinite-dimensional parametric model*. In these notes, *parametric inference* is meant literally in the sense of inferences about $\theta^* \in \Theta$ whether the dimension is finite or infinite; *nonparametric inference* refers to inferential goals that are beyond the parameter space.

Although infinite-dimensional problems occur as exercises, the focus here is on finite-dimensional models. There is an intermediate class of problems in which the space $\Theta \equiv \Theta_n$ is design-dependent or sample-size dependent, with finite dimension p_n dependent on n. Very often, dimension-dependent spaces are used as an artificial mathematical device to gauge the effect of 'many parameters' on the behaviour of the estimation procedure in extreme situations. Every model considered in this book is a family of processes. Although the parameter space may be infinite-dimensional, it is fixed and independent of the design.

The emphasis in this chapter is on normal behaviour of models and estimates, not on anomalies. Typically, we assume that the model is *identifiable*, which means that

$$\theta \neq \theta' \implies P_\theta \neq P_{\theta'}.$$

In other words, different parameter values correspond to distinct processes. Identifiability does not necessarily imply $P_{n,\theta} \neq P_{n,\theta'}$ for small samples, say $n = 1$ or $n = 2$. Nor does it imply that θ is estimable from the data, even for large n. Identifiability is not a strong condition, nor is it an essential condition: see the mixture problem in Exercise 15.1.

15.1.2 Bayesian resolution

A Bayesian model has all of the ingredients listed in the preceding section—plus one other. The additional feature is a probability distribution $\pi(\cdot)$ on the parameter space. The non-Bayesian model is generally portrayed as a stochastic formulation whose appropriateness in a given application is widely agreed, whereas no broad consensus is expected regarding the choice of $\pi(\cdot)$. One is said to be objective and the other subjective. These adjectives are not only provocative and unhelpful, but also devoid of mathematical content.

The net effect of the prior is that a Bayesian model is either (i) a single distribution $\pi(d\theta)P_{\theta,n}(\cdot)$ on the product space $\Theta \times \mathbb{R}^n$; or (ii) a single mixture process $P_\pi(\cdot) = \int P_\theta(\cdot)\,\pi(d\theta)$. In principle, the reduction to a mixture is a huge simplification because the estimation step is by-passed, the model comprises a single process, and the ambiguity about the choice of process for prediction is eliminated. Even for problems of parametric inference, it is usually possible in principle to by-pass the parameter space entirely by re-phrasing the target as a

tail event associated with a limit statistic, for example by computing $P_\pi(\bar{Y}_\infty \in A \mid Y[\mathbf{x}])$ or $P_\pi(\hat{\theta}_\infty \in A \mid Y[\mathbf{x}])$.

In practice, two difficulties must be overcome before we can confidently take advantage of the Bayesian solution. The first is to select a suitable prior distribution and, more importantly, to convince the reader that this prior is appropriate for the problem. The Bayes resolution calls for a single prior distribution selected to represent the information available a priori. In practice, it is often better to depart from the paradigm by considering a sequence of distributions π_ν for $\nu > 0$ such that the available information corresponds either to the limit $\nu \to 0$ or to the asymptote in which ν is small but strictly positive.

Absence of information may be represented by a sequence of distributions such that $\pi_\nu(A) \to 0$ at rate $\rho_\nu > 0$ on bounded subsets in such a way that $\rho_\nu^{-1}\pi_\nu(dx)$ has a finite non-zero limit. The limit is a measure, sometimes termed 'improper' because it is not a probability distribution. However, for sufficiently large samples, the conditional distribution given the data may have a limit that is satisfactory for inference. Usually it depends on the limit measure, but is otherwise independent of the sequence. At the other end of the spectrum, strong information such as sparsity corresponds to a sequence that tends to the Dirac measure in a suitably regular way that permits limits for a certain class of integrals; (McCullagh and Polson, 2018).

These limit recipes are reasonably satisfactory for stylized problems in low-dimensional parameter spaces. For high-dimensional spaces, assumptions of independence for selected components are not to be taken lightly because their effect on conclusions may be substantial.

The second problem, perhaps less of an obstacle today than in the recent past, is to manage the computations. Posterior distributions can often be approximated by simulation in various ways, for example, using Markov-chain Monte Carlo. Bayesian computation is not a focus of this book, so we do not make sweeping recommendations regarding the choice of prior or now to manage the computation.

15.2 Likelihood function

15.2.1 Definition

In the simplest setting where the response distribution has a density $P_\theta(dy) = p_\theta(y)\,dy$, the density $p_\theta(y)$ as a function of θ for fixed y is the likelihood function. The function $p_\theta(y)$ is the density of the probability relative to Lebesgue measure. For present purposes, there is nothing special about Lebesgue measure, so the likelihood function is the density ratio relative to an arbitrary fixed density. For example, if zero is a point in the parameter space, we could adopt $L(\theta) = p_\theta(y)/p_0(y)$ as the likelihood function provided that $p_0(y)$ is strictly positive throughout the space. The important point to remember is that only likelihood ratios $p_\theta(y)/p_{\theta'}(y)$ are well defined. Even in that case, it is necessary to handle points of zero density and points of infinite density with care.

On the technical side, we assume that there is a dominating measure that covers all distributions in the model. Usually this is Lebesgue measure or counting measure. But in some settings such as survival models, the measure has a discrete part associated with censored values, and a continuous part associated with failure times.

The likelihood function is a fundamental object for statistical estimation and inference. For parametric Bayes tasks, the likelihood function is the ratio of the posterior distribution to the prior on Θ:

$$P_\pi(d\theta \mid Y) \propto L(\theta; y)\pi(d\theta).$$

Its non-Bayesian role is a little more complicated, but it is equally fundamental. Mostly it is used for point estimation and interval estimation.

15.2.2 Bartlett identities

The likelihood is a function $L(\theta; y)$ of the parameter θ and the data y, and the same applies to the log likelihood $l(\theta; y) = \log L(\theta; y)$. Since the likelihood is defined up to an arbitrary multiplicative factor that is constant in θ, the log likelihood is defined up to an arbitrary additive term that is constant in θ.

The Bayesian goal is to compute the conditional probability of some specified inferential event given the data, and in that calculation y is regarded as a fixed constant. However, frequentist properties of estimators are connected with the statistical behaviour of log likelihood derivatives and related procedures for fixed θ as a function of the random variable whose distribution is P_θ. The Bartlett identities are fundamental for deriving large-sample asymptotic distributions in regular problems. To keep notation digestible, we pretend that θ is a scalar, so that each log likelihood derivative is also a scalar. Results for vector-valued parameters are obtained by replacing scalars with vectors or matrices as appropriate.

The first two log likelihood derivatives are

$$U_1(\theta; y) = dl(\theta; y)/d\theta; \quad U_2(\theta; y) = d^2 l(\theta; y)/d\theta^2.$$

The Bartlett identities are connected with the moments of these and higher-order derivatives. The first identity follows from the constancy of the integral $\int p_\theta(y)\, dy$ as a function of θ.

$$\begin{aligned} 0 &= \frac{\partial}{\partial\theta} \int p_\theta(y)\, dy \\ &= \int \frac{\partial p_\theta(y)}{\partial\theta}\, dy \\ &= \int \frac{\partial \log p_\theta(y)}{\partial\theta} p_\theta(y)\, dy \\ &= E\big(U_1(\theta; Y); \theta\big). \end{aligned}$$

The first step in this derivation is to switch the order of differentiation with respect to θ and integration over the observation space. This step requires a

regularity condition, which fails if the support of P_θ is parameter-dependent. Regularity conditions must be taught by faculty and learned by students, if only to demonstrate mastery of Fubini's theorem, but they almost never fail in practical work. In the last expression θ occurs twice, the first to indicate the differentiation point, the second to indicate that the parameter of the distribution $Y \sim P_\theta$ is the same as the point at which the derivative is computed. The random variable $U_1(\theta; Y)$ does not have zero mean under the distribution $Y \sim P_{\theta^*}$.

The second identity, which follows from the second derivative of the probability integral, establishes the role of the Fisher information matrix

$$\begin{aligned} 0 &= \frac{\partial^2}{\partial\theta^2} \int p_\theta(y)\, dy, \\ &= \frac{\partial}{\partial\theta} \int \frac{\partial \log p_\theta(y)}{\partial\theta} p_\theta(y)\, dy, \\ &= \int \left(\frac{\partial^2 \log p_\theta(y)}{\partial\theta^2} + \Big(\frac{\partial \log p_\theta(y)}{\partial\theta} \Big)^2 \right) p_\theta(y)\, dy, \\ &= E\big(U_2(\theta; Y); \theta\big) + E\big(U_1^2(\theta; Y); \theta\big). \end{aligned}$$

The final expression implies that the variance of the first derivative is the negative expected value of the second derivative, which is called the Fisher information matrix:

$$I(\theta) = -E\big(U_2(\theta; Y); \theta\big) = \operatorname{cov}\big(U_1(\theta; Y); \theta\big).$$

It follows that $I(\theta) > 0$, and, for vector parameters, that $I(\theta)$ is positive definite.

15.2.3 Implications for estimation

Regularity conditions for statistical work are of two types, those that can be checked or verified and those that cannot. Fubini-type conditions permitting the interchange of sample-space integration with parameter-space differentiation are verifiable. Conditions regarding the smoothness of functions or topological adequacy of the parameter space are also verifiable. Statistical models that occur in applied work are usually field-tested and are seldom in violation of verifiable conditions.

Asymptotic conditions holding in the large-sample limit are a different matter. Much of the theory of statistical estimation uses asymptotic theory as a device for distributional approximation. For simple processes having independent and identically distributed components, the only route to infinity is 'more independent copies of the same'. For more general spatial or temporal processes, or processes involving covariates, the routes to infinity are more numerous. By their nature, asymptotic conditions are not verifiable in any finite sample because any finite design can be embedded into a sequence of larger designs in countless ways. The question to be asked is not whether the given design is part of a particular sequence but whether one conceptual design sequence provides a better distributional approximation than another.

The motivation for large-sample theory is most straightforward for independent and identically distributed sequences. Such sequences seldom occur naked in applied work, so the iid theory is not directly relevant. However, the crucial parts of the theory carry over with relatively minor modification to models having independent observations. Additional conditions are needed for asymptotic regularity in specialized models for genetics, time series and spatial processes. The count of individual numbers or observations or rows in a data file may be impressive, but that does not necessarily translate into an impressive quantity of information.

In a setting where the components of the response are independent, or conditionally independent given treatment, the log likelihood is a sum of n independent contributions, and the same applies to the log likelihood derivatives. In particular, the total Fisher information $I_\bullet(\theta) = \sum I_i(\theta)$ is the sum of positive contributions coming from individual components. The first derivative at θ^* is the sum of independent random variables, $U_1, \ldots, U_n$, having zero mean and finite variances $I_i(\theta) < \infty$. Provided that n is large and that no small subset of components dominates the contribution to the total Fisher information, the central limit theorem implies that the first derivative at θ^* is approximately normally distributed

$$U_\bullet(\theta^*) \sim N\big(0, I_\bullet(\theta^*)\big)$$

under the distribution P_{θ^*}. Assuming that the maximum is a stationary point, Taylor approximation in a neighbourhood of θ^* gives

$$0 = U_\bullet(\hat\theta) = U_\bullet(\theta^*) - I_\bullet(\theta^*)(\hat\theta - \theta^*) + O_p(1).$$

To first order in the sample size, this implies

$$\hat\theta - \theta^* \simeq I_\bullet^{-1}(\theta^*) U_\bullet(\theta^*) \sim N\big(0, I_\bullet^{-1}(\theta^*)\big) \tag{15.1}$$

in which the uncomputable $I_\bullet(\theta^*)$ may be replaced with the computable $I_\bullet(\hat\theta)$. Probability calculations using this asymptotic approximation have error of order $O(n^{-1/2})$ in the sample size. However, the error can often be reduced to acceptable levels by parameter transformation. More accurate approximations using bias corrections and Edgeworth series are available in the literature.

The linear approximation (15.1) is often is re-packaged as a computational algorithm, which generates from a starting point $\hat\theta_0$ a parameter sequence satisfying

$$\hat\theta_{r+1} = \hat\theta_r + I_\bullet^{-1}(\hat\theta_r) U_\bullet(\hat\theta_r). \tag{15.2}$$

If this sequence converges, it converges to a stationary point of the log likelihood, which is a local maximum and usually the global maximum. Technically, the sequence (15.2) is not Newton-Raphson because it uses the Fisher information or expected second derivative at $\hat\theta_r$, which is not usually the same as the observed second derivative at that point.

15.2.4 Likelihood-ratio statistic I

Taylor expansion of the log likelihood function about θ^* including terms up to degree two in $\hat{\theta} - \theta^*$ gives

$$l(\hat{\theta}; y) - l(\theta^*; y) = U_\bullet(\theta^*)(\theta^* - \hat{\theta}) - \tfrac{1}{2} I_\bullet(\theta^*)(\theta^* - \hat{\theta})^2 + \cdots .$$

For models having independent ad identically distributed components, the first derivative is $O_p(n^{1/2})$, while second and higher-order derivatives are $O_p(n)$. As a result, both terms shown are formally $O_p(1)$ while the error term is $O_p(n^{-1/2})$. Under suitable asymptotic conditions, these asymptotic orders also hold more broadly for generalized linear models and many models having temporal or spatial correlation.

Using the one-step approximation (15.1) for the parameter estimate, the likelihood-ratio statistic satisfies

$$\begin{aligned} 2l(\hat{\theta}; y) - 2l(\theta^*; y) &= U_\bullet(\theta^*) I_\bullet^{-1}(\theta^*) U_\bullet(\theta^*) + O_p(n^{-1/2}), \\ &= U_\bullet(\theta^*) I_\bullet^{-1}(\hat{\theta}) U_\bullet(\theta^*) + O_p(n^{-1/2}), \\ &= (\hat{\theta} - \theta^*) I_\bullet(\hat{\theta})(\hat{\theta} - \theta^*) + O_p(n^{-1/2}). \end{aligned}$$

The first and second versions are positive definite quadratic forms in the vector of first derivatives at the true or hypothesized parameter point; the third is a quadratic form in the parameter space. The likelihood ratio statistic is invariant under smooth reparameterization, and that property is inherited by the first quadratic form shown, which is called the Rao statistic, or Fisher-Rao statistic. The third version, called the Wilks statistic, is not invariant under reparameterization. Invariance is desirable in applied work, but perhaps not absolutely essential.

The central limit approximation for the distribution of log likelihood derivatives implies that all three versions of the likelihood-ratio statistic are first-order equivalent, and that the limit distribution is χ^2_p in all cases. They are not second-order equivalent, either in power or in distribution. A more refined analysis taking account of higher-order terms shows that the expected value of the likelihood-ratio statistic is $p(1 + b(\theta)/n)$, and that the asymptotic distribution is $(1 + b(\theta)/n)\chi^2_p$ with error $O(n^{-2})$. The use of the Bartlett correction factor greatly improves the accuracy of the χ^2_p approximation. This adjustment holds for regular problems with continuous distributions.

15.2.5 Profile likelihood

Most parametric models that occur in applied work make a distinction between parameters of interest and other parameters, loosely called nuisance parameters. Despite the nomenclature, nuisance parameters are essential for satisfactory inferences.

The parameter of interest is defined by a differentiable function $T\colon \Theta \to \Theta'$ from the parameter space of dimension p into a manifold of dimension $q \leq p$.

We suppose without loss of generality that this mapping is onto, i.e., $T\Theta = \Theta'$. To each $\tau \in \Theta'$ there corresponds a sub-manifold of dimension $p - q$

$$\Theta_\tau = \{\theta : T(\theta) = \tau\} \subset \Theta.$$

All points in Θ_τ are similar in the sense that they have the same value of the parameter of interest; differences are associated with nuisance parameters. By construction, the sub-manifolds are disjoint and exhaustive in Θ; they form a partition or a foliation of the parameter space.

The profile likelihood for τ is the maximum achieved on Θ_τ:

$$l_p(\tau; y) = \max_{\theta \in \Theta_\tau} l(\theta; y) = l(\hat{\theta}_\tau; y).$$

To first order in the sample size, the profile likelihood behaves like an ordinary likelihood function. For example, the first derivative has mean of order $O(n^{-1})$, which is not zero but is small enough to permit the standard asymptotic argument to proceed. Likewise, the expected value of the second derivative is not exactly the variance of the first, but the difference is small enough that it does not affect first-order asymptotic approximations under standard regularity conditions. Consequently, the subset consisting of parameter values achieving near-maximum likelihood

$$\{\tau \in \Theta' : 2l(\hat{\theta}; y) - 2l(\hat{\theta}_\tau; y) \leq \chi^2_{q,1-\alpha}\}$$

is an approximate $1 - \alpha$-confidence subset for the parameter of interest.

15.2.6 Two worked examples

Example 1: Treatment effect estimation

The standard Gaussian model for a completely randomized design has three parameters $\theta = (\mu_0, \mu_1, \sigma^2)$, two means and one variance $\sigma^2 > 0$, so Θ has dimension three. The log likelihood function is

$$l(\theta; y) = -\frac{n_0(\bar{y}_0 - \mu_0)^2}{2\sigma^2} - \frac{n_1(\bar{y}_1 - \mu_1)^2}{2\sigma^2} - \frac{(n-2)s^2}{2\sigma^2} - n \log \sigma$$

in standard notation for sample sizes n_0, n_1, sample means $\bar{y}_0, \bar{y}_1$, and pooled sample variance s^2. The treatment effect is the difference $T(\theta) = \mu_1 - \mu_0$, and the focus of the analysis is primarily on that parameter. The profile log likelihood for the treatment effect is the maximum value achieved on the subset $\Theta_\tau \subset \Theta$

$$\Theta_\tau = \{\theta : \mu_1 - \mu_0 = \tau\}; \qquad l(\hat{\theta}_\tau; y) = \max_{\theta \in \Theta_\tau} l(\theta; y).$$

Usually the maximum for fixed τ must be computed numerically, but it can be evaluated explicitly in this instance:

$$\begin{aligned}
\hat{\mu}_0 &= \big(n_0\bar{y}_0 + n_1\bar{y}_1 - n_1\tau\big)/n;\\
\hat{\mu}_1 &= \big(n_0\bar{y}_0 + n_0\tau + n_1\bar{y}_1\big)/n = \hat{\mu}_0 + \tau;\\
n\hat{\sigma}^2 &= (n-2)s^2 + n_0(\bar{y}_0 - \hat{\mu}_0)^2 + n_1(\bar{y}_1 - \hat{\mu}_1)^2;\\
l(\hat{\theta}_\tau; y) &= \tfrac{n}{2}\log(\hat{\sigma}^2) + \text{const}\\
&= \tfrac{n}{2}\log\Big((n-2)s^2 + n_0(\bar{y}_0 - \hat{\mu}_0)^2 + n_1(\bar{y}_1 - \hat{\mu}_1)^2\Big) + \text{const}'\\
&= \tfrac{n}{2}\log\Big((n-2)s^2 + n_0n_1(\bar{y}_0 - \bar{y}_1 + \tau)^2/n\Big) + \text{const}'.
\end{aligned}$$

The partially maximized likelihood function is called the profile likelihood for the parameter of interest. By construction, the overall maximum occurs at the ordinary maximum of the likelihood, $\hat{\tau} = \bar{y}_1 - \bar{y}_0$ in this example.

Asymptotically, the profile log likelihood has all of the essential properties of a log likelihood function. For example, an approximate level-α likelihood-based confidence interval can be obtained in the standard manner

$$\{\tau : 2l(\hat{\theta}; y) - 2l(\hat{\theta}_\tau; y) \le \chi^2_{1,1-\alpha}\}. \tag{15.3}$$

In this example, it is possible to construct the standard exact confidence interval for τ using the ratio

$$t_\tau = \sqrt{\frac{n_0n_1}{n}}\,\frac{\bar{Y}_0 - \bar{Y}_1 + \tau}{s},$$

which has the Student t distribution on $n-2$ degrees of freedom. The exact coverage of the likelihood-based interval can be inferred from the fact that the likelihood-ratio statistic $n\log\big(1 + t^2_\tau/(n-2)\big)$ is monotone in t^2_τ.

Example 2: Inference for the LD_{90}

Suppose that the response of unit i to dose x is a Bernoulli variable with parameter $\pi(x)$ satisfying the linear logistic model

$$\operatorname{logit}\pi(x) = \theta_0 + \theta_1 x, \tag{15.4}$$

with independent responses for distinct units. The goal is to estimate the dose τ for which $\pi(\tau) = 0.9$, the so-called lethal dose 90%. The LD_{90} is a non-linear function of the parameters

$$\begin{aligned}
\operatorname{logit}(0.9) = \log(9) &= \theta_0 + \theta_1\tau;\\
\tau &= \big(\log(9) - \theta_0\big)/\theta_1,
\end{aligned}$$

so we take $T(\theta) = \big(\log(9) - \theta_0\big)/\theta_1$. To compute the profile likelihood for τ, it is necessary to fit the logistic model (15.4) by maximizing over the parameter subset

$$\Theta_\tau = \{(\theta_0, \theta_1) : T(\theta) = \tau\} = \{(\log(9) - \tau\theta_1,\, \theta_1) : \theta_1 \in \mathbb{R}\}.$$

In other words, we aim to fit the one-parameter sub-model

$$\text{logit}\,\pi(x) = \log(9) - \tau\theta_1 + \theta_1 x = \log(9) + \theta_1(x - \tau)$$

for arbitrary but fixed τ. This is not a linear logistic model in the strict technical sense, but most computer packages have the option to cater for an offset, which is the constant $\log(9)$ in this setting. The likelihood-based confidence region for τ is the set of values for which the likelihood is sufficiently large in the sense of (15.3).

If we replace the linear logistic model (15.4) with a linear Gaussian model and ask for the x-value that makes the mean response zero, the goal is the abscissa or parameter ratio $\tau = -\theta_0/\theta_1$. Fieller's method is tailored for problems of this sort. However, the likelihood-ratio statistic is a function of the standardized ratio on which Fieller's method is based, so the two approaches are essentially identical.

One point to note is that a likelihood-based confidence set is not necessarily an interval. Equivariance under reparameterization makes this unavoidable. For instance, if the likelihood-based confidence set for $\tau = \theta_0/\theta_1$ is a bounded interval containing zero, the confidence set for $1/\tau$ is necessarily an 'interval' containing $\pm\infty$ but not zero. In both the linear logistic and Gaussian cases, the likelihood-based confidence set is either an interval or the complement of an interval, or possibly the whole space.

15.3 Generalized linear models

15.4 Linear Gaussian models

15.5 Mixture models

15.5.1 Two-component mixtures

Let ψ_0 and ψ_1 be the density functions of two distributions on the real line. Both densities are assumed to be strictly positive, so the density ratio $\zeta(y) = \psi_1(y)/\psi_0(y)$ is finite, as is the inverse ratio. The mixture model refers to the family of distributions

$$\psi_\theta(y) = (1 - \theta)\psi_0(y) + \theta\psi_1(y),$$

which is a convex set indexed by the mixture parameter $0 \leq \theta \leq 1$.

According to the standard statistical paradigm, the observations $Y_1, \ldots, Y_n$ are independent and identically distributed as ψ_θ for some unknown parameter value. Statistically speaking, the estimation and testing problems are regular if $0 < \theta < 1$; in such circumstances, the standard asymptotic approximations hold for the distribution of $\hat{\theta}$ and for the likelihood-ratio statistic. Otherwise, if $\theta = 0$ or $\theta = 1$ on the boundary, the problem is non-regular; standard asymptotic

approximations cannot be relied upon for either the distribution of $\hat{\theta}$ or of the likelihood-ratio statistic.

Given the observation $y = (y_1, \ldots, y_n)$, the log likelihood function for θ is

$$l(\theta; y) = \sum \log(\psi_\theta(y_i)) = \sum \log(1 - \theta + \theta\zeta(y_i)) + \text{const}(y).$$

To understand the behaviour as a function of θ, we examine the derivatives

$$l'(\theta; y) = \sum_i \frac{\zeta(y_i) - 1}{1 - \theta + \theta\zeta(y_i)};$$
$$l''(\theta; y) = -\sum_i \left(\frac{\zeta(y_i) - 1}{1 - \theta + \theta\zeta(y_i)} \right)^2 < 0.$$

If all of the observation points satisfy $\psi_0(y_i) = \psi_1(y_i)$, then $\zeta(y_i) = 1$ for each i, and the log likelihood is constant in θ. Otherwise, the second derivative is everywhere strictly negative, implying concavity. Every stationary point is a global maximum, and there is at most one such point in $(0, 1)$.

At the left end-point $l'(0; y) = \sum \zeta(y_i) - n$; if the derivative at zero is negative, i.e., if $\sum \zeta(y_i) \leq n$, the maximum occurs at $\hat{\theta} = 0$. At the right end-point $l'(1; y) = n - \sum 1/\zeta(y_i)$. If the derivative is positive, i.e., if $\sum 1/\zeta(y_i) \leq n$, the maximum occurs at $\hat{\theta} = 1$. The likelihood function has a maximum in the interior of the interval if and only if $\sum \zeta(y_i) > n$ and $\sum 1/\zeta(y_i) > n$. In that case, the maximum can be computed by a straightforward Newton-Raphson iteration.

For $0 < \hat{\theta} < 1$, the condition $l'(\hat{\theta}; y) = 0$ implies

$$\sum \frac{\zeta(y_i)}{1 - \hat{\theta} + \hat{\theta}\zeta(y_i)} = \sum \frac{1}{1 - \hat{\theta} + \hat{\theta}\zeta(y_i)} = n,$$

which can be viewed as a self-consistency condition. If we associate with each i the class-I assignment probability

$$\hat{\theta}(y_i) = \frac{\hat{\theta}\zeta(y_i)}{1 - \hat{\theta} + \hat{\theta}\zeta(y_i)} = \text{pr}(i \mapsto \text{class I} \mid Y),$$

then $\hat{\theta} = n^{-1} \sum \hat{\theta}(y_i)$ is the sample mean of the assignment probabilities.

15.5.2 Likelihood-ratio statistic

For likelihood-ratio statistics it is convenient to take ψ_0 as the reference point. Relative to that point, the maximized likelihood ratio statistic is

$$l(\hat{\theta}; y) - l(0; y) = \sum \log\bigl(1 - \hat{\theta} + \hat{\theta}\zeta(y_i)\bigr).$$

In particular, the likelihood-ratio statistic is zero if $\hat{\theta} = 0$, i.e., if $\sum \zeta(y_i) \leq n$.

If we regard ψ_0 as the null hypothesis, it must be understood that $\theta = 0$ is a boundary point, and that the standard asymptotic theory may fail—and indeed it does fail spectacularly. For the null distribution theory, the observations are independent with distribution ψ_0. An elementary computation shows that if $Y \sim \psi_0$, the random variable $\zeta(Y) = \psi_1(Y)/\psi_0(Y)$ is non-negative with mean one. Thus, by the law of large numbers, $n^{-1}\sum \zeta(Y_i) \to 1$. In addition, if $\zeta(Y)$ has finite variance, the central limit theorem implies asymptotic normality, so that the event $\sum \zeta(Y_i) \leq n$ occurs with limiting probability one half. In those cases, the null distribution of $\hat{\theta}$ has an atom of $1/2$ at the origin, and the same goes for the likelihood-ratio statistic. This sort of behaviour is non-standard, but it is classical for boundary-point problems.

For the more usual sorts of mixtures that occur in practical applications, $\zeta(Y)$ does not have finite variance. In those cases, the convergence of the average $n^{-1}\sum \zeta(Y_i) \to 1$ does not imply that the event $n^{-1}\sum \zeta(Y_i) \leq 1$ has a limiting probability or that the limit is one half. As an example, if ψ_0 is standard normal, and ψ_1 is Cauchy, the random variable $\zeta(Y)$ has a density whose tail behaviour is $O(z^{-2}\log(z)^{-3/2})$. The mean is one, but there are no finite moments beyond the first. The limit distribution appears from simulation to be such that

$$n^{-1}\sum_{i=1}^{n} \zeta(Y_i) = 1 - \frac{\text{const}}{\log\log n} + \frac{\epsilon}{\log n \, \log\log n},$$

where ϵ is a random variable in the Landau class (stable with $\alpha = \beta = 1$). The event $\sum \zeta(Y_i) > n$ is equivalent in the limit to $\epsilon > \text{const} \times \log n$. Since the Landau density has an inverse-square right tail, the probability is $O(1/\log n)$.

Every mixture model in which ψ_1 is symmetric with inverse-square tails gives the same limit. For other distributions having sub-Gaussian tails such as $e^{-|y|}$, the same limit is approached at a possibly different rate. In all such cases, the limiting null distribution for $\hat{\theta}$ and the likelihood-ratio statistic are degenerate at zero. This is not standard asymptotic behaviour for boundary-point problems.

15.5.3 Sparse signal detection

Given a random signal $X \sim P$ and an observation $Y = X + \varepsilon$ contaminated by additive independent Gaussian noise, how do we estimate the signal? The non-sparse signal estimation problem was first posed by F. Dyson in 1926. Eddington's solution, which is described in section 13.4.4, depends only on the marginal density $m(y)$ of the observation. The sparse version of the problem is discussed by Johnstone and Silverman (200?). For simplicity, we assume here that ε is standard normal.

A signal $X \sim P$ is said to be sparse if its distribution is symmetric and most of the mass is concentrated at or near the origin. In that case, the sparsity rate ρ is defined by the integral

$$1 - \rho = \int e^{-x^2/2} p(x)\, dx.$$

The statement that ρ is small is to be interpreted a mathematical code or convention, which implies a formal limit $\rho \to 0$ even if that is not explicitly stated. The reason for focusing on the sparsity rate as opposed to the null atom $P(X = 0)$ is that the null atom may be zero; more crucially, ρ is a mixture fraction that is identifiable from observations whereas the null atom is not. Subsequent conclusions depend only on the sparsity rate, which is strictly smaller than the probability of a non-null signal.

Under regularity conditions given in McCullagh and Polson (2018), the marginal density is a Gaussian mixture

$$m(y) = \phi(y)\big(1 - \rho + \rho\zeta(y)\big) + o(\rho),$$

where ζ is a symmetric non-negative convex function satisfying $\zeta(0) = 0$. In practice, ρ must first be estimated from the data.

The essence of the matter is that all sparse scale families having similar tail behaviour give rise to the same zeta function. The horseshoe family with density $\log(1 + y^{-2})/(2\pi)$ has the same inverse-square tail behaviour as the Cauchy family, and the zeta function for both satisfies $\zeta''(y) = \exp(y^2/2)$. This implies $\zeta(y) = \sum_{r\geq 1} \mu_{2r-2} y^{2r}/(2r)!$, where $\mu_{2r} = 1 \cdot 3 \cdots (2r-1)$ is the $2r$th standard Gaussian moment. It is possible to make an elaborate argument for one over the other, but such arguments are futile because the marginal distributions are indistinguishable to first order.

Eddington's signal-estimation formula reduces to

$$E(X_i \mid Y) = \frac{\rho\zeta'(y_i)}{1 - \rho + \rho\zeta(y_i)} + o(\rho)$$
$$E(X_i^2 \mid Y) = \frac{\rho\zeta''(y_i)}{1 - \rho + \rho\zeta(y_i)} + o(\rho).$$

In addition, the signal identification or conditional exceedance probability for threshold $\epsilon > 0$ is formally the same as $E(|X_i|^0 \mid Y)$:

$$P(|X_i| > \epsilon \mid Y) = \frac{\rho\zeta(y_i)}{1 - \rho + \rho\zeta(y_i)} + o(1).$$

For a given suitably low but strictly positive threshold, the exceedance probability is approximately independent of the threshold (McCullagh and Polson, 2018), and $\zeta(y)$ is interpretable as the posterior-to-prior odds ratio, also called the Bayes factor. For example, if $\rho = 0.05$ and $y = 3.5$, the inverse-square zeta value is $\zeta(y) = 55.3$ and the exceedance probability is 0.74. Provided that below-threshold signals are counted as null or false, there is a close formal connection with the concept of false discovery rate, and particularly with the local false discovery rate (Benjamini and Hochberg, 199?; Efron, 200?).

In the great majority of sparse signal identification and detection formulations the second component of the mixture $\psi_2(y) = \phi(y)\zeta(y)$ has heavy tails. The tails are governed by the signal distribution, which may be either Laplace-type $e^{-|y|}$ or Cauchy-like with regularly varying tails. In all such models, the

asymptotic null distribution of $\hat{\rho}$ and of the likelihood-ratio statistic is is degenerate at zero. However, a very small signal little larger than $\rho \simeq \log(n)/n$ is enough to change the calculus for signal detection, at least in the Cauchy case.

15.6 Inferential compromises

15.6.1 The dictatorial compromise

The fundamental difficulty with the non-Bayesian model $\{P_\theta : \theta \in \Theta\}$ for inferential purposes is that it contains more than one stochastic process. Which one, if any, are we to use for prediction? The Bayesian paradigm resolves the difficulty by compromise, which—however reasonable it may be and however little its effect on conclusions may be—is ultimately dictatorial. That compromise consists of a prior distribution or mixture $\pi(d\theta)$ so that the set $\{P_\theta\}$ is replaced with the single mixture $P_\pi = \int P_\theta \, \pi(d\theta)$. Prediction is straightforward.

For parametric inference, the event $\theta \in A$ is identified with the set of sequences $y \in \mathbb{R}^\infty$ such that $\hat{\theta}(y) = \lim_{n\to\infty} \hat{\theta}_n(y[n])$ exists and belongs to A. In that way, the conditional probability of the event $\theta \in A$ given $Y[n]$ is computable as a tail event

$$P_\pi(\theta \in A \mid Y[n]) = P_\pi(\hat{\theta}(y) \in A \mid Y[n]).$$

Any consistent estimator can be used in place of $\hat{\theta}_n$, so this description is not tied in any way to maximum likelihood.

15.6.2 The one-time democrat

Consider a standard non-Bayesian model consisting of processes P_θ, with finite-dimensional distributions $P_{n,\theta}$ on $\mathbb{R}^n$. Maximum-likelihood estimation offers two ways to generate a new process that is related to the family. The first of these is the standard parametric bootstrap, or parametric simulation.

Bootstrap process: Given a observation point $y \in \mathbb{R}^m$, and the corresponding point $\hat{\theta} = \hat{\theta}_m(y)$ in Θ, the entire family $\{P_\theta\}$ is replaced with the maximum-likelihood representative $\hat{P} = P_{\hat{\theta}}$. The finite-dimensional distributions are $\hat{P}_n = P_{n,\hat{\theta}}$ on $\mathbb{R}^n$. In particular, if each P_θ defines a process with independent components, the bootstrap representative is also a process whose components are conditionally independent given $\hat{\theta}_n$.

15.6.3 The sequential democrat

The maximum-likelihood process (MLP) operates in a different manner and exhibits fundamentally different behaviour, which is analogous to the behaviour of a Polya urn. Every process P_θ determines a Markov kernel, which associates with each $n \geq 0$ and each point $y \in \mathbb{R}^n$, a conditional distribution

$$Q_{n1,\theta}(dy_{n+1}; y) = P_{n+1,\theta}(dy_{n+1} \mid Y[n] = y)$$

on $\mathbb{R}$. The finite-dimensional joint density is the product of such kernels

$$P_{n,\theta}(dy) = Q_{1,\theta}(dy_1)Q_{2,\theta}(dy_2; y[1]) \cdots Q_{n,\theta}(dy_n; y[n-1]).$$

The conditional distribution given $Y[m]$ is a truncated kernel product

$$Q_{m+1,\theta}(dy_{m+1}; y[m])Q_{m+2,\theta}(dy_{m+2}; y[m+1]) \cdots Q_{m+n}(dy_{m+n,\theta}; y[m+n-1]).$$

In the maximum-likelihood process, each transition kernel $Q_{n+1,\theta}(dy_{n+1}; y)$ is replaced with the maximum-likelihood estimate

$$\hat{Q}_{n+1}(dy_{n+1}; y) = Q_{n+1,\hat{\theta}_n(y)}(dy_{n+1}; y).$$

This makes sense only for n sufficiently large that $\hat{\theta}_n = \hat{\theta}_n(y[n])$ exists. Given an initial sequence $y[m]$, the distribution of successive values in the MLP is defined by the kernel product

$$Q_{m+1,\hat{\theta}_m}(dy_{m+1}; y[m]) \times Q_{m+2,\hat{\theta}_{m+1}}(dy_{m+2}; y[m+1]) \times \cdots$$

in which the maximum-likelihood point is updated at each stage.

15.7 Acknowledgements

I am grateful to Nick Polson for bringing the Eddington/Dyson paper to my attention.

15.8 Exercises

15.1 Maximum-likelihood for mixtures: Let $\psi_0(\cdot), \ldots, \psi_k(\cdot)$ be given probability density functions on $\mathbb{R}$, and let

$$m_\theta(y) = \theta_0\psi_0(y) + \cdots + \theta_k\psi_k(y)$$

be a $k+1$-component mixture with non-negative weights adding to one. Suppose that $Y_1, \ldots, Y_n$ are independent and identically distributed with density m_θ, assumed to be strictly positive for θ strictly positive. Under what conditions is the mixture model with iid observations identifiable? Show that the maximum-likelihood estimator satisfies the condition

$$\sum_{i=1}^{n} \frac{\psi_r(y_i)}{\hat{m}(y_i)} \leq n,$$

with equality for every r such that $\hat{\theta}_r > 0$. Discuss the 'almost-true' claim that $\hat{m}$ exists and is unique for every $n \geq 1$ and every $y \in \mathbb{R}^n$, even if the model is not identifiable.

15.2 Let $\psi_0(y) = e^{-y^2/2}/\sqrt{2\pi}$ be the standard normal density. Assume that $Y_1, \ldots, Y_n$ are independent standard normal. Show that the random variables $X_i = \psi_1(Y_i)/\psi_0(Y_i)$ have unit mean, and hence, by the law of large numbers, that the sample average tends to one as $n \to \infty$.

15.3 If the claim made in the last paragraph of section 15.5.2 is to be believed, the re-scaled limit distribution of $\bar{X}_n$ does not have a mean. Discuss this apparent contradiction.

15.4 Consider the two-component mixture with ψ_0 standard normal, and ψ_1 standard Cauchy. The null hypothesis is all Gaussian, i.e., $\theta = (1, 0)$. Show that $\hat{\theta}_1 > 0$ if and only if $\bar{X}_n > 1$. By simulation or otherwise, show that $P_0(\bar{X}_n > 1) \to 0$ as $n \to \infty$. What is the effect of changing the Cauchy scale parameter?

15.5 Show that the random variables $X_i = \psi_1(Y_i)/\psi_0(Y_i)$ in the preceding exercise have a density whose tail behaviour is $1/f(x) \sim x^2 \log(x)^{3/2}$ as $x \to \infty$.

15.6 Explain why the observation $P_0(\bar{X}_n > 1) \to 0$ as $n \to \infty$ deduced from simulations does not conflict with the law of large numbers $\bar{X}_n \to 1$.

15.7 Consider the two-component mixture with ψ_0 standard normal and ψ_1 standard Laplace, or double exponential. Investigate the behaviour of $P_0(\bar{X}_n > 1)$ as a function of n for large n. What is the effect of changing the scale parameter? What do these calculations this imply about the null distribution of the likelihood-ratio statistic?

15.8 Sparse signal detection. Suppose that the observation $Y = X + \varepsilon$ is the sum of a signal X plus independent Gaussian noise $\varepsilon \sim N(0, 1)$. For any signal distribution $X \sim P_\nu$, the sparsity rate is defined by the integral

$$\rho = \int (1 - e^{-x^2/2})\, P_\nu(dx).$$

Suppose that the signal is distributed according to the Dirac-Cauchy mixture $P_\nu(dx) = (1-\nu)\delta_0(dx) + \nu C(dx)$ in which the null atom $1 - \nu$ is the null-signal rate. Find the sparsity rate corresponding to 5% non-zero signals.

15.9 For the setting of the previous exercise, show that Y is distributed according to the mixture with density

$$m(y) = (1-\rho)\phi(y) + \rho\psi(y) + o(\rho) = \phi(y)\big(1 - \rho + \rho\zeta(y)\big) + o(\rho)$$

where $\psi(\cdot)$ is a probability density, $\zeta(y) = \psi(y)/\phi(y)$ is the density ratio, and $\zeta(0) = 0$. Fill in the details needed to express $\zeta(\cdot)$ or $\psi(\cdot)$ as a function of the family P_ν.

15.10 Suppose that $Y_1, \ldots, Y_n$ are independent and identically distributed with density $m(y)$. Ignoring the error term, show that the maximum-likelihood

estimate of the mixture fraction is zero if $\sum \zeta(y_i) \le n$, one if $\sum 1/\zeta(y_i) \le n$, and otherwise is a point $0 < \hat\rho < 1$ satisfying

$$\sum \frac{\zeta(y_i) - 1}{1 - \hat\rho + \hat\rho\zeta(y_i)} = 0.$$

Hence or otherwise deduce that the maximum-likelihood estimate of the mixture satisfies the self-consistency condition

$$\sum \frac{\psi(y_i)}{\hat m(y_i)} = \sum \frac{\phi(y_i)}{\hat m(y_i)} = n.$$

In what sense does this equation imply self-consistency?

15.11 A sequence $\epsilon_\nu \to 0$ such that $P_\nu(|X| < \epsilon_\nu) \to 1$ as $\nu \to 0$ is called a signal negligibility threshold. Show that the conditional probability of a non-negligible signal is

$$P_\nu(|X| > \epsilon_\nu \mid Y) = \frac{\rho\zeta(y)}{1 - \rho + \rho\zeta(y)} + o(1),$$

which implies that the 'true discovery rate' is essentially independent of the threshold.

15.12 What does the preceding equation imply about the fraction of non-negligible signals among sites in the sample such that $|Y_i| \ge 3$?

15.13 Let $\kappa_0 = \rho\zeta(y)/(1 - \rho + \rho\zeta(y))$ be the exceedance probability, and let κ_r be the rth derivative of $\log(1 - \rho + \rho\zeta(y))$. For $\zeta(y) \simeq e^{y^2/2}/y^2$ for large y, show that Eddington's formulae give

$$E(X \mid Y) \simeq \kappa_0\big(y^2 - 2\big)/|y|, \quad \operatorname{var}(X \mid Y) \simeq \kappa_0(1 - \kappa_0)(y^2 - 3) + \kappa_0^2$$

for large y. Discuss the implications for mean shrinkage and variance inflation.

15.14 For $1 \le i \le k$, suppose $Y_i = \alpha_i + \epsilon_i$, where $\epsilon_1, \dots, \epsilon_k$ are independent standard normal variables, and $\alpha_1, \dots, \alpha_k$ are exchangeable and independent of ϵ. Let F be the joint distribution of α. The goal of this exercise is to find an estimator of F as a function of the observation $y \in \mathbb{R}^k$. Ideally the estimator should be the maximum-likelihood estimator or an approximation thereof within a set of distributions having some natural symmetry. In the candidate estimators listed below, λ and s are unspecified scalars, $\delta_x(\cdot)$ is the Dirac measure at x, $\mathcal{M}_k$ is the set of functions $[k] \to [k]$, $\mathcal{S}_k \subset \mathcal{M}_k$ is the set of permutations, and τy is the composition $(\tau y)_i = y_{\tau(i)}$.

$$\hat F_0(\cdot) = \frac{1}{k!} \sum_{\tau \in \mathcal{S}_k} \delta_{\lambda\tau y}(\cdot);$$

$$\hat F_1(\cdot) = \frac{1}{k^k} \sum_{\tau \in \mathcal{M}_k} \delta_{\lambda\tau y}(\cdot);$$

$$\hat F_2(\cdot) = N_k(\mathbf{1}\bar y,\ s^2 I_k).$$

Show that $\hat{F}_0$ and $\hat{F}_1$ are both exchangeable with the same marginal distribution, and that $\hat{F}_1$ also has independent components. For $\lambda = 1$, these are called the permutation estimator and the bootstrap estimator respectively.

Chapter 16

Parametric models

Chapter 17

Residual likelihood

17.1 Background

Residual maximum likelihood (REML) is a technique proposed by Patterson and Thompson (1971) for estimating variances and variance components in a linear Gaussian model in which the observation $y \in \mathbb{R}^n$ is regarded as a realization of the Gaussian random vector $Y \sim N_n(X\beta, \Sigma)$. The model matrix, or design matrix, X is of order $n \times p$ and known, with image subspace $\mathcal{X} = \text{span}(X)$. In the simplest version of the covariance model, the matrix is expressed as a linear combination of given symmetric non-negative definite matrices

$$\Sigma = \sigma_1^2 V_1 + \cdots + \sigma_k^2 V_k \tag{17.1}$$

with non-negative coefficients $\sigma_1^2, \ldots, \sigma_k^2$ to be estimated. These coefficients are called *variance components*; the space of matrices determined by (17.1) is the convex cone spanned by the given matrices. Usually V_1 is the identity matrix of order n; the remaining matrices are typically block factors or other known relationships among the observational units.

In other settings, the model for Σ may not be linear in all parameters, but partial linearity is fairly common, as is linearity after transformation. In a spatial or time-series setting, the variance model may be a combination such as

$$\text{cov}(Y_s, Y_t) = \sigma_0^2 \delta_{s-t} + \sigma_1^2 e^{-\lambda|s-t|}$$

which is linear for fixed λ. On the other hand, a Gaussian graphical model is an additive specification for the inverse covariance matrix. The simplest version is

$$\Sigma^{-1} = \tau_0 I_n + \tau_1 G,$$

where $G \subset [n]^2$ is the graph incidence matrix, and the coefficients are subject to positive-definiteness conditions. Usually, this means $\tau_0 > 0$ and $\tau_1 \leq 0$.

Residual likelihood differs from ordinary likelihood in that it uses only the residuals $R = LY$, where L is any linear transformation such that

$$\ker(L) = \mathcal{X} = \{X\beta : \beta \in \mathbb{R}^p\}.$$

By focusing on the residuals, the regression parameters are eliminated from the distribution

$$R \sim N_n(LX\beta,\, L\Sigma L') = N_n(0,\, L\Sigma L').$$

It is crucial that R be observable, which means that L is a fixed linear transformation independent of all parameters. Under (17.1), the residual covariance matrix $L\Sigma L'$ is a linear combination of the matrices LV_rL', so the residual likelihood is a function of the variance components only. Ordinarily, the matrices LV_rL' are non-zero and linearly independent, which implies that the variance-components are identifiable from the residuals. After estimating the variance components by maximizing the residual likelihood, the second step is to compute

$$\hat{\Sigma} = \sum_{r=1}^{k} \hat{\sigma}_r^2 V_r,$$

its inverse $\hat{W} = \hat{\Sigma}^{-1}$, and the weighted least squares estimate of β

$$\hat{\beta} = (X'\hat{W}X)^{-1}X'\hat{W}Y. \tag{17.2}$$

The covariance matrix of $\hat{\beta}$ is then reported as $(X'\hat{W}X)^{-1}$.

17.2 Simple linear regression

In a simple linear regression model with a single variance component, the covariance matrix is $\Sigma = \sigma^2 V$, where V is known and strictly positive definite. It is convenient in this setting to take $W = V^{-1}$ as the inner-product matrix, so that $P_X = X(X'WX)^{-1}X'W$ and $Q = I - P$ are complementary W-orthogonal projections. Then WQ and QV are both known and symmetric. The model for the residual $QY \sim N(0, \sigma^2 QV)$ has only a single parameter. For this full exponential-family setting, the quadratic form $\|QY\|^2 = Y'WQY$ is minimal sufficient, and the REML estimate is obtained by equating the observed value $\|Qy\|^2$ to its expected value:

$$\|Qy\|^2 = E(Y'WQY; \hat{\sigma}^2) = \hat{\sigma}^2 \operatorname{tr}(VWQ) = \hat{\sigma}^2 \operatorname{tr}(Q).$$

Since Q is a projection with rank $\operatorname{tr}(Q) = n - p$, the REML estimate reduces to the standard unbiased estimator that is universally recommended and used in all computer packages

$$\hat{\sigma}^2 = y'WQy/(n-p).$$

Note that the REML estimate is strictly larger than the ordinary maximum-likelihood estimator which is $y'WQy/n$. The lesson here is that REML, not ML, is the norm for variance estimation.

17.3 The REML likelihood

17.3.1 Projections

For the development in this section, K is a matrix of order $n \times k$, whose columns span a subspace $\mathcal{K}$ of dimension k, and T is a complementary matrix of order $n \times (n-k)$ such that $T'K = 0$. In other words, the linear transformation $T' \colon \mathbb{R}^n \to \mathbb{R}^{n-k}$ satisfies $\ker(T') = \mathcal{K}$. For the moment, the relation between $\mathcal{K}$ and $\mathcal{X}$ is left unspecified, but $\mathcal{K} = 0$, $\mathcal{K} = \mathcal{X}$ and $\mathcal{K} \subset \mathcal{X}$ are the most important special cases.

The observation space $\mathbb{R}^n$ is regarded as a real inner-product space with inner product matrix $W = \Sigma^{-1}$. For most purposes, W can be replaced with any proportional matrix, as was done in section 1.2. Consider the three $n \times n$ matrices:

$$P = K(K'WK)^{-1}K'W; \qquad Q = I - P; \qquad A = \Sigma T(T'\Sigma T)^{-1}T'. \tag{17.3}$$

It is readily checked that $P^2 = P$, $Q^2 = Q$ and $A^2 = A$, so all three are idempotent, and thus linear projections $\mathbb{R}^n \to \mathbb{R}^n$. They are also self-adjoint, meaning that WP, WQ and WA are symmetric, which implies they are orthogonal projections. In addition, $x \in \mathcal{K}$ implies $T'x = 0$, which implies $Ax = 0$, and hence $\mathcal{K} \subset \ker(A)$. Finally, $Ax = 0$ implies $T'Ax = 0$, which implies $T'x = 0$, which implies $x \in \mathcal{K}$, and hence $\ker(A) \subset \mathcal{K}$. Thus, A is the orthogonal projection with kernel $\mathcal{K}$, and Q is also the orthogonal projection with kernel $\mathcal{K}$. Uniqueness implies $A = Q$ and $I - A = P$.

17.3.2 Determinants

Now consider the partitioned matrix $H = [T, K]$, which is invertible of order n, and the related matrix

$$H'\Sigma H = \begin{pmatrix} T'\Sigma T & T'\Sigma K \\ K'\Sigma T & K'\Sigma K \end{pmatrix}.$$

The condition $T'K = 0$ implies that $H'H$ is block-diagonal with determinant $\det(H'H) = \det(T'T)\det(K'K)$, and hence that

$$\det(H'\Sigma H) = \det(H'H)\det(\Sigma) = \det(T'T)\det(K'K)\det(\Sigma).$$

Using the standard formula for the determinant of a partitioned matrix, we find

$$\begin{aligned}
\det(H'\Sigma H) &= \det(T'\Sigma T)\det(K'\Sigma K - K'\Sigma T(T'\Sigma T)^{-1}T'\Sigma K) \\
&= \det(T'\Sigma T)\det\big(K'[\Sigma - T(T'\Sigma T)^{-1}T'\Sigma]K\big) \\
&= \det(T'\Sigma T)\det(K'(I-A)\Sigma K) \qquad \text{from (17.3)} \\
&= \det(T'\Sigma T)\det(K'K(K'WK)^{-1}K'W\Sigma K) \\
\det(T'T)\det(K'K)\det(\Sigma) &= \det(T'\Sigma T)\det{}^2(K'K)/\det(K'WK) \\
\frac{\det(T'T)}{\det(T'\Sigma T)} &= \frac{\det(K'K)}{\det(K'WK)\det(\Sigma)}.
\end{aligned}$$

For REML applications where the kernel in specified by K, the determinantal term in the marginal likelihood is the expression on the right.

17.3.3 Marginal likelihood with arbitrary kernel

For any linear transformation such as T' having kernel $\mathcal{K}$, the linear transformation $Y \mapsto T'Y$ is called a residual modulo $\mathcal{K}$. All transformations having the given kernel determine the same likelihood function. The marginal log likelihood based on the linear transformation $T'Y \sim N(T'\mu, T'\Sigma T)$ is

$$l = -\tfrac{1}{2}(y-\mu)'T(T'\Sigma T)^{-1}T'(y-\mu) - \tfrac{1}{2}\log\det(T'\Sigma T) + \text{const.}$$

In this setting, l is a function on the parameter space, and the additive constant may be any function that is constant on the parameter space. It is convenient here to take a particular constant, namely $\frac{1}{2}\log\det(T'T)$ plus any function of y. This choice ensures that, for every invertible matrix L of order $n-k$, the linear transformations T' and LT' produce identical versions of the log likelihood. With this choice, the marginal log likelihood based on the residuals modulo $\mathcal{K}$ is one half of

$$\begin{aligned} 2l &= -(y-\mu)'WA(y-\mu) + \log\det(T'T) - \log\det(T'\Sigma T) \\ &= -(y-\mu)'WQ(y-\mu) - \log\det(\Sigma) \\ &\qquad - \log\det(K'WK) + \log\det(K'K), \end{aligned} \tag{17.4}$$

where Q is the orthogonal projection with kernel $\mathcal{K}$, and K is any matrix whose columns span $\mathcal{K}$.

In applications where $\mathcal{X}$ is the model subspace, the most common choice is $\mathcal{K} = \mathcal{X}$, but expression (17.4) is valid for all subspaces, and $\mathcal{K} \subset \mathcal{X}$ arises in the computation of likelihood-ratio statistics. The ordinary log likelihood with kernel $\mathcal{K} = 0$ is obtained by setting $K = 0$. The standard REML likelihood has $K = X$ and $\mathcal{K} = \mathcal{X}$ so that $\mu \in \mathcal{X}$ implies $Q\mu = 0$:

$$2l = -y'WQy - \log\det(\Sigma) - \log\det(X'WX) + \log\det(X'X). \tag{17.5}$$

Formulae (17.4) and (17.5) may be used directly in computer software. The constant term $\log\det(K'K)$ is included to ensure that the log likelihood depends on the kernel subspace, not on the particular choice of basis vectors.

For general-purpose software, these formulae are not recommended because the marginal likelihood requires only that $T'\Sigma T$ be positive definite, which is a weaker condition than positive-definiteness for Σ. Marginal likelihood modulo a suitable kernel may be used for fitting generalized Gaussian processes, sometimes called intrinsic processes, that are defined by a generalized covariance function, which is not positive definite in the normal sense, but for which $T'KT$ is positive definite. For example, if $i \mapsto z_i$ is a quantitative covariate taking values in $\mathbb{R}^k$, the matrix $\Sigma_{ij} = -\|z_i - z_j\|$ is positive definite in the Euclidean space $\mathbb{R}^n/\mathbf{1}$, which is the space of residuals modulo the one-dimensional subspace of constant functions.

17.3.4 Likelihood ratios

A likelihood ratio at E is the ratio of probabilities assigned to the event E by two probability measures:

$$LR_{\theta',\theta}(E) = \frac{P_{\theta'}(E)}{P_\theta(E)}.$$

A maximized likelihood ratio is a similar expression

$$\frac{\sup_{\theta\in\Theta_1} P_\theta(E)}{\sup_{\theta\in\Theta_0} P_\theta(E)}$$

in which the numerator and denominator are maximized over the respective parameter spaces. It is crucial that all probability measures be defined on the same σ-field and that event in the numerator be the same as the event in the denominator; otherwise the ratio is not a fair comparison. In fact, the event is always an observation or singleton event, which is best regarded as an infinitesimal event, and commonly denoted by $E = dy$. Operationally speaking, dy is the limiting ϵ-ball $B(y,\epsilon)$ centered at the observation point $y \in \mathbb{R}^n$, and the likelihood ratio is the density ratio at y.

In the case of marginal likelihood, however, the event $E \subset \mathbb{R}^n$ is necessarily an event in the σ-field generated by the linear transformation T' into the Borel space $\mathbb{R}^{n-k}$. The induced σ-field in $\mathbb{R}^n$ is the class of residual events, which are the Borel subsets of $\mathbb{R}^n$ such that $E + \mathcal{K} = E$. In other words, a residual is a point in the quotient space $\mathbb{R}^n/\mathcal{K}$, and each residual event E is a union of translates of $\mathcal{K}$, i.e., a union of $\mathcal{K}$-cosets. The residual event, $E = B(y,\epsilon)+\mathcal{K}$, is the union of $\mathcal{K}$-cosets that intersect the ball. This is, of course a Borel subset in the space of residuals modulo $\mathcal{K}$. A residual likelihood ratio statistic modulo $\mathcal{K}$ is thus a ratio of the form

$$\frac{\sup_{\theta\in\Theta_1} P_\theta(dy+\mathcal{K})}{\sup_{\theta\in\Theta_0} P_\theta(dy+\mathcal{K})}$$

in which the limiting event $B(y,\epsilon)+\mathcal{K}$ is the observed residual. A ratio such as

$$\frac{\sup_{\theta\in\Theta_1} P_\theta(dy+\mathcal{K}_1)}{\sup_{\theta\in\Theta_0} P_\theta(dy+\mathcal{K}_0)}$$

is not a likelihood ratio unless $\mathcal{K}_0 = \mathcal{K}_1$.

17.4 Computation

17.4.1 Software options

By default, the function `lmer(...)` estimates the variance components by maximizing the residual log likelihood (17.5). As a follow-up, it reports the weighted least squares estimate (17.2) of the regression coefficients. The square

roots of the diagonal components of the inverse Fisher information $(X'\hat{W}X)^{-1}$ serve as standard errors. The optional argument `REML=FALSE` is a cop-out, which overrides the default, and reverts to ordinary maximum likelihood instead. This option produces a valid likelihood ratio statistic, which is not the one recommended by Welham and Thompson (1997) or by the writer. Maximum likelihood with $\mathcal{K} = 0$ is not recommended because the variance estimates have a multiplicative bias of order $O(p/n)$, whose effect is sometimes not negligible.

The function `regress(y~X, ~block+V, kernel=K)` has a three-part syntax, permitting greater flexibility, in which the setting for `kernel` determines the method of estimation. The first part is a standard model-formula for the mean-value subspace $\mathcal{X}$; the second part, which may be empty or missing, is a simple model formula for the covariances. Each term in the second part is either a symmetric matrix or a factor; each factor is converted internally into a block matrix by `outer(fac, fac, "==")`, and $\hat{\Sigma}$ is a linear combination of these matrices. The identity matrix is included by default as the first element in the list. The set of matrices must be linearly independent as vectors in $\mathbb{R}^{n^2}$. For the third part, the default is $\mathcal{K} = \mathcal{X}$, i.e., REML, not $\mathcal{K} = 0$. The log likelihood value reported by `regress(...)$llik` is the maximized log likelihood (17.4), using whatever kernel is specified or implied. The zero-dimensional and one-dimensional options `kernel=0` and `kernel=1` are permitted but not recommended.

17.4.2 Likelihood-ratios

To compute a likelihood ratio, we need a null model and an alternative model, preferably nested. It is essential that both models be fitted based on the same information or data, i.e., that the same kernel be used in both.

For the comparison of mean-values $H_0 : \mu \in \mathcal{X}_0$ versus $H_1 : \mu \in \mathcal{X}_1 \supset \mathcal{X}_0$ as alternative, residual likelihood may be used in the following manner:

```
X0 <- model.matrix(~mf0);   X1 <- model.matrix(~mf1)
fit0 <- regress(y~mf0, ~block+V, kernel=X0);   # default kernel
fit1 <- regress(y~mf1, ~block+V, kernel=X0);   # default over-ridden
2*(fit1$llik - fit0$llik);
```

Here, `mf0` and `mf1` denote the model formulae for $\mathcal{X}_0$ and $\mathcal{X}_1$ respectively. The space of covariance matrices is fixed but arbitrary, and `block+V` is used solely for illustration.

Welham and Thompson (1997) recommend setting the kernel equal to the null subspace, i.e., $\mathcal{K} = \mathcal{X}_0 \subset \mathcal{X}_1$, which is the computation illustrated above. Provided that $\mathcal{K}$ is fixed, $\mathcal{X}_0 \subset \mathcal{X}_1$ and $\mu \in \mathcal{X}_0$, the log likelihood ratio is distributed approximately as χ^2 on $\dim(\mathcal{X}_1 + \mathcal{K}) - \dim(\mathcal{X}_0 + \mathcal{K})$ degrees of freedom, which simplifies for $\mathcal{K} \subseteq \mathcal{X}_0$ to $q = \dim(\mathcal{X}_1) - \dim(\mathcal{X}_0)$ independent of the kernel. The numerical value of the likelihood-ratio statistic for $\mathcal{K} \subseteq \mathcal{X}_0$ depends on the kernel, but the first-order asymptotic approximation to the null distribution is χ^2_q, which is independent of $\mathcal{K}$. The choice $\mathcal{K} = \mathcal{X}_0$ is thought

to be optimal in the sense of power, and in the sense of accuracy of the χ^2 distributional approximation.

The option `kernel=0` is permitted but not encouraged; it implies ordinary maximum likelihood, and is equivalent to the `REML=FALSE` option in `lmer()`. The option `kernel = X1` is also allowed; this option produces a valid likelihood ratio statistic that is exactly zero. Why? Because the hypothesis concerns $\mu \in \mathcal{X}_1$, and the residuals modulo $\mathcal{X}_1$ contain no information about the parameter.

For the comparison of two nested models having the same mean-value subspace, the REML default option is recommended:

```
fit0 <- regress(y~mf0, ~block);
fit1 <- regress(y~mf0, ~block+site);
2*(fit1$llik - fit0$llik);
summary(fit1)
```

Ordinarily, the one-dimensional subspace of constant functions is a subspace of $\mathcal{X}$, while `block` and `site` are factors having at least two levels. Every factor that occurs in a covariance model is converted internally into a block factor or equivalence matrix

```
Vb <- outer(block, block, "==")    Vs <- outer(site, site, "==")
```

so the first model specifies a linear combination of $V_1 = I_n$ and $V_2 =$ `block` as a block factor, while the second specifies a linear combination of three matrices. Positivity of coefficients is not automatically enforced. Provided that the third coefficient is not restricted to be positive, the asymptotic null distribution is χ^2_1. To force positivity for the site variance component, the code may be modified as follows:

```
fit0 <- regress(y~mf0, ~block);
fit1 <- regress(y~mf0, ~block+site, pos=c(0,0,1), start=c(fit0$sigma, 1));
2*(fit1$llik - fit0$llik);
summary(fit1)
```

The asymptotic null distribution is a 50% mixture $0.5\delta_0 + 0.5\chi^2_1$.

17.4.3 Testing for interaction

Consider an experimental design consisting of three physical sites, one northern one southern and one western, separated by a considerable distance that is sufficient to affect the local climate. Each site consists of four blocks of six plots, all of which are outdoors. Each plot is assigned by randomization to one of two treatment levels, which are constant in time. On 127 days over a two-year period, measurements are made on one plant in certain designated plots. By construction, there is one treatment factor; `site` is regarded as a classification factor. `block` is a factor with 12 levels, while `plot` has 72 levels; both are naturally regarded as block factors. The levels of the remaining factor `day` have a temporal component, which may be important for certain purposes,

but the temporal aspect is ignored in the initial discussion, which is concerned with treatment by site interaction. Is the treatment effect constant over sites?

The simplest null model includes additive independent and identically distributed random effects for observations in the same block, additional additive random effects for observations in the same plot, and independent additive effects for observations on the same day. The simplest models with and without interaction are specified implicitly as follows:

```
X0 <- model.matrix(~site+treat);
fit0 <- regress(y~site+treat, ~block+plot+day);
fit1a <- regress(y~site*treat, ~block+plot+day, kernel=X0);
fit1b <- regress(y~site*treat, ~block+plot+day, start=fit1a$sigma);
2*(fit1a$llik - fit0$llik);   summary(fit1b)
```

The *treat*×*site* interaction space has dimension two, so the null distribution of the likelihood ratio statistic is χ^2_2. The parameter estimates reported by `fit1a` and `fit1b` should be very similar, but not identical. If the number of observations is large, it is helpful to supply an initial value for the iteration, as illustrated for `fit1b` above. Note that `fit1b` uses the default kernel `site*treat`, so `fit1b$llik` is not comparable with `fit0$llik`.

It may happen that the response has a temporal component that is continuous in time, as opposed to the process implied by the inclusion of `day` as a block factor above, in which the daily contributions are independent and identically distributed. One simple option is to assume that the temporal component behaves like free Brownian motion, with generalized covariance function proportional to $-|t - t'|$. Free Brownian motion has independent increments on non-overlapping intervals; it is a stationary process in the sense that the distribution of increments are constant in time.

```
dayv <- as.numeric(paste(day));  BM <- -abs(outer(dayv, dayv, "-"))
X0 <- model.matrix(~site+treat);
fit0 <- regress(y~site+treat, ~block+plot+BM);
fit1a <- regress(y~site*treat, ~block+plot+ BM, kernel=X0);
fit1b <- regress(y~site*treat, ~block+plot+BM, start=fit1a$sigma);
2*(fit1a$llik - fit0$llik);   summary(fit1b)
```

It is essential in this script that the kernel subspace include the constant functions. The option `kernel=1` is acceptable; `kernel=0` is not acceptable, and may produce an error message.

17.4.4 Singular models

There are various ways in which singularities may arise in a variance-components model. Consider, for example, a design in which each subject is one experimental unit, and several response measurements of the same type are made on each individual. See project 1 in which up to five measurements are made at different sites on each rat. Then `subject` is a partition of the experimental units,

which is a sub-partition of `treatment`. Ordinarily, the covariance model contains `subject` as a block factor with independent and identically distributed effects, and the model for the mean contains the treatment factor. This model is non-singular, and the computation should be straightforward. However, if the exchangeability assumption for subject effects is dropped, and the effects are included additively in the mean model, the subspace spanned by `subject` includes the subspace spanned by `treatment`. As a consequence, treatment effects are not identifiable, i.e., they are confounded with subject effects..

A different sort of singularity arises when a factor such as `block` is included both in the model for the mean and in the model for covariances. While the software may complain about singularities or lack of identifiability, it is feasible to examine this situation analytically. The model is technically identifiable in the sense that distinct parameter values give rise to distinct probability distributions. However, the likelihood function achieves its maximum on the boundary of the space at which `block` has a zero coefficient in the covariance model. In other words, the factor in the mean model trumps the block factor in the covariance model. The model can thus be fitted by dropping the block factor from the covariances.

17.5 Numerical example

17.6 Exercises

17.1 Consider a balanced block design having m blocks each consisting of b observational units, and let B be the associated block factor as a Boolean matrix of order $n = mb$. The three-parameter Gaussian model with moments

$$\mu \in \mathbf{1}_n, \quad \Sigma = \sigma^2(I_n + \theta B),$$

is parameterized by two scalars μ, $\sigma > 0$ and one additional parameter. For the purposes of this exercise $\theta > -1/b$ is not necessarily positive, but Σ is positive definite. In addition, the residual refers to any linear transformation, such as $Y_{ij} - \bar{Y}_{\bullet\bullet}$, whose kernel is $\mathbf{1} \subset \mathbb{R}^n$.

Let Y_{ij} be the observation for unit j in block i. Show that the within- and between- quadratic forms

$$\mathrm{SS}_W = \sum_{ij} (Y_{ij} - \bar{Y}_{i\bullet})^2, \quad \mathrm{SS}_B = b \sum_{i} (\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet})^2,$$

are independent with distributions $\sigma^2\chi^2_{n-m}$ and $\sigma^2(1 + b\theta)\chi^2_{m-1}$ respectively. Hence deduce that, if $\theta = 0$, the mean-square ratio

$$F = \frac{\mathrm{SS}_B \,/(m-1)}{\mathrm{SS}_W \,/(n-m)}$$

is distributed according to Fisher's $F_{m-1,n-m}$ distribution.

17.2 For the balanced block design in the preceding exercise, show that the implied distribution for residuals is a two-parameter full exponential-family model with canonical sufficient statistic $\mathrm{SS}_W, \mathrm{SS}_B$. Hence deduce that the [residual] maximum-likelihood estimate satisfies

$$\hat{\sigma}^2 = \mathrm{SS}_W/(n-m), \quad 1 + b\hat{\theta} = F.$$

Show also that the sub-model with $\theta = 0$ is a one-parameter exponential family model with sufficient statistic $\mathrm{SS}_B + \mathrm{SS}_W$.

17.3 For the balanced block design, show that the log determinant is

$$\log\det(\Sigma) = n\log(\sigma^2) + m\log(1 + b\theta).$$

Show that the ML estimate satisfies $1 + b\hat{\theta} = (m-1)F/m$. Hence deduce that the ordinary log likelihood ratio statistic for testing $\theta = 0$ is

$$\log\det\hat{\Sigma}_0 - \log\det\hat{\Sigma}_1 = n\log\left(\frac{n-m+(m-1)F}{n}\right) - m\log\left(\frac{(m-1)F}{m}\right),$$

while the REML statistic is

$$(n-1)\log\left(1 + \frac{(m-1)(F-1)}{n-1}\right) - (m-1)\log F.$$

What does this expression tell you about the null distribution of the REML likelihood-ratio statistic?

17.4 Show that the REML estimate with positivity constraint satisfies $1+b\hat{\theta} = \max(F, 1)$. What is the REML estimate for the second component? Express the constrained REML likelihood-ratio statistic as a function of F, and compute the atom at the origin.

17.5 The following exercise is concerned with the distribution of the likelihood-ratio statistic in a 'fixed-effects' model for a balanced design, where $\Sigma = \sigma^2 I_n$, and either $\mu \in \mathbf{1}_n$ under the null hypothesis or $\mu \in \mathrm{span}(B)$ under the alternative. The meaning of the term 'residual' is unchanged, and the F-statistic in exercise 17.2 is also unchanged.

Show that the residual log likelihood-ratio statistic for testing $\mu \in \mathbf{1}$ versus $\mu \in \mathrm{span}(B)$ is

$$(n-1)\log\left(1 + \frac{(m-1)F}{n-m}\right).$$

By simulation or otherwise, show also that the null expected value exceeds that of χ^2_{m-1} by the approximate multiplicative factor

$$1 + \tfrac{1}{2}(m+1)/(n-m).$$

This is a particular instance of the Bartlett correction factor.

17.6 The null hypothesis being tested in exercise 17.5 is the same as that in exercise 17.3, but the alternatives are different: one implies exchangeability of block effects, the other does not. Discuss the implications of the fact that one statistic is strictly increasing as a function of F, whereas the other is strictly decreasing for $F < 1$ and strictly increasing for $F > 1$.

17.7 Positive definiteness of a function $\mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ means that,F for integer $n \geq 1$ and each finite collection of points $\mathbf{x} = \{x_1, \ldots, x_n\}$, the $n \times n$ matrix $K[\mathbf{x}, \mathbf{x}]$ with components

$$K_{ij} = K(x_i, x_j)$$

is positive definite.

Let x be a point in real Euclidean space $\mathbb{R}^d$. For each $\lambda > 0$, the function

$$K(x, x') = e^{-\lambda \|x - x'\|}$$

is the covariance function for an autoregressive process of order 1 if $d = 1$, also called the Ornstein-Uhlenbeck process for general d. Show that K is positive definite.

Chapter 18

Response transformation

18.1 Likelihood for Gaussian models

In applied work, it is frequently advantageous to transform the observations prior to fitting a linear Gaussian model. Invariably, this means that the state space for each observation $Y_i \in S$ is an open real interval such as $S = (0, \infty)$ or $S = (0, 1)$ or $S = \mathbb{R}$. A transformation $g\colon S \to \mathbb{R}$ is identified and applied component-wise to the vector $Y \in \mathbb{R}^n$ in the hope that the transformed variable gY might be approximately normally distributed with mean $\mu \in \mathcal{X}$, and covariance Σ belonging to some family of covariance matrices such as those described in chapter ?. According to this scenario, the joint density of the observation Y at the point $y \in S^n$ is

$$(2\pi)^{-n/2}|\Sigma|^{-1/2}e^{-(gy-\mu)'\Sigma^{-1}(gy-\mu)/2}\prod_{i=1}^{n}|g'(y_i)|$$

provided that g is 1–1 differentiable with a differentiable inverse, i.e., a diffeomorphism $S \to \mathbb{R}$. To specify the likelihood function, it is necessary to identify the set of transformations $g \in \mathcal{G}$ under consideration, plus the mean-value space $\mathcal{X} = \text{span}(X)$ and the space Θ of covariance matrices. To be clear, these moment spaces are moment spaces for the transformed variable gY, not for Y. As a function on $\mathcal{G} \times \mathcal{X} \times \Theta$, this density is the likelihood function.

It is helpful at this stage to insert two additional technical conditions. First, the space $\mathbf{1}$ of constant n-vectors is a subspace of $\mathcal{X}$; this is not required in the theory of linear models, but it is universal in applied work and it is needed at certain points in the argument that follows. Second, the space of covariance matrices is a cone, i.e., $\Sigma \in \Theta$ implies $\tau\Sigma \in \Theta$ for every scalar multiple $\tau > 0$. Both conditions are mathematically essential but relatively benign; the cone need not be convex. The cone condition extends to Σ^{-1} and ensures that the maximum-likelihood estimate $\hat{\mu}_g$, $\hat{\Sigma}_g$ for fixed g satisfies

$$(gy - \hat{\mu}_g)'\hat{\Sigma}_g^{-1}(gy - \hat{\mu}_g) = n.$$

As a consequence, the profile log likelihood for the transformation $g \in \mathcal{G}$ is

$$l_p(g) = -\tfrac{1}{2} \log \det(\hat{\Sigma}_g) + \sum_{i=1}^{n} \log |g'(y_i)|. \tag{18.1}$$

Finally, for all scalars $a, b \neq 0$, the cone condition and $\mathbf{1} \subset \mathcal{X}$ imply $l_p(a + bg; y) = l_p(g; y)$, so that the profile likelihood is invariant with respect to affine composition. In other words, the transformations $y \mapsto g(y)$ and $y \mapsto a+bg(y)$ are equivalent for this comparison: $gY \sim N(\mu, \Sigma)$ implies $a+bgY \sim N(a+b\mu, b^2\Sigma)$, and vice-versa.

The preceding analysis assumes that the maximum-likelihood estimate $\hat{\mu}_g, \hat{\Sigma}_g$ exists. Existence and uniqueness cannot be guaranteed in general, but failure is rare in practice provided that $p < n$ and the residual space is adequate to estimate all variance components.

18.2 Box-Cox transformation

18.2.1 Power transformation

One very natural option is to choose a simple parametric family such as the family of power transformations (Box and Cox, 1964). Provided that $\mathbf{1} \subset \mathcal{X}$ and all observations are strictly positive, the transformation $(0, \infty) \to \mathbb{R}$ may be taken in the form $y \mapsto (y^\lambda - 1)/\lambda$ for some scalar λ, with the limit $\lambda \to 0$ corresponding to the log function. The derivative $y^{\lambda-1}$ is strictly positive, so the profile log likelihood for λ is

$$l_p(\lambda; y) = -\tfrac{1}{2} \log \det(\hat{\Sigma}_\lambda) + (\lambda - 1) \sum_{i=1}^{n} \log y_i \tag{18.2}$$

provided that the maximum-likelihood estimate $\hat{\Sigma}_\lambda$ exists. It is a straightforward exercise to plot $l_p(\lambda)$ against λ to check whether there is a clear maximum in the range of interest, which is typically $-1 \leq \lambda \leq 1$. A large value of the likelihood-ratio statistic $2l_p(\hat{\lambda}) - 2l_p(1)$ indicates a need for transformation.

The profile log likelihood is meant to be used only as a rough guide. In practice, the only transformations that are ordinarily considered for linear statistical analysis are (i) the logarithm if the response scale is strictly positive with a well-defined origin, and effects are expected to be multiplicative; (ii) the identity if treatment effects are expected to be additive on the given scale; (iii) occasionally the reciprocal, square root or cube root if there is a reasonable justification based on the physical units of measurement. For example, if the observation is a volume, an argument might be made for the cube-root; if the observation is a time or duration, conversion by reciprocals to the rate scale or frequency scale might make sense. But additivity of effects on such scales is usually dubious, so the log transformation is the preferred choice for most physical variables such as mass, volume, length, time, or ratios such as speed, density, miles per gallon, and so on. Under no circumstances should the reported analysis be done on the scale $Y^{\hat{\lambda}}$, where $\hat{\lambda}$ is the maximum-likelihood estimate from (18.2).

18.2.2 Re-scaled power transformation

Let $\tau > 0$, and let $y \mapsto \tau(y^\lambda - 1)/\lambda$ be the re-scaled power transformation applied component-wise on the transformed scale. The Jacobian is $\tau^n \prod y_i^{\lambda-1}$, so the log Jacobian is

$$\log J = n(\lambda - 1) \log \dot{y} + n \log \tau,$$

where $\dot{y}$ is the geometric mean of the observations. As a numerical device, it is sometimes helpful to set the scale parameter to $\tau = \dot{y}^{1-\lambda}$ so that the Jacobian is one. With this choice, the profile likelihood reduces to the determinantal term in (18.2).

The preceding discussion refers to re-scaling $g \mapsto \tau g$ by composition on the left, i.e., by multiplication after power transformation. Composition on the right $g \mapsto g\tau$ refers to the effect of re-scaling $y \mapsto \tau y$ before power transformation:

$$\begin{aligned} \text{left}: &\quad y \overset{g}{\longmapsto} g(y) \overset{\tau}{\longmapsto} \tau g(y) \\ \text{right}: &\quad y \overset{\tau}{\longmapsto} g(y) \overset{g}{\longmapsto} g(\tau y). \end{aligned}$$

Composition on the right sends $g(\cdot)$ to $g(\tau\cdot)$, which is an affine transformation of $g(\cdot)$:

$$g(\tau y) = \tau^\lambda g(y) + \frac{\tau^\lambda - 1}{\lambda} = \tau^\lambda g(y) + \text{const.}$$

The assumption $\mathbf{1} \subset \mathcal{X}$, and the cone condition on covariance matrices, are sufficient to ensure that likelihood-based conclusions are unaffected by scalar composition on the right. Invariance is absolutely essential in applied work, where the choice of physical units—inches versus centimetres or minutes versus seconds—is quite arbitrary.

The purpose of re-scaling on the left is not to modify the power transformation in a substantive way, but to simplify the computation. Nonetheless, the argument in the first paragraph could easily be misconstrued as a statement that the modified power transformation

$$y_i \mapsto \frac{y_i^\lambda}{\lambda \dot{y}^{\lambda-1}} \quad \text{or} \quad y_i \mapsto \frac{y_i^\lambda - 1}{\lambda \dot{y}^{\lambda-1}}$$

has Jacobian equal to one. As they are written above, these transformations do not act component-wise. The first transformation satisfies $g(\tau y) = \tau g(y)$, but the Jacobian $J = |\lambda|^{-1}$ is discontinuous at $\lambda = 0$. For the second transformation, the Jacobian

$$J = \frac{1}{\lambda} + \frac{\lambda - 1}{n\lambda} \sum y_i^{-\lambda}$$

is continuous at $\lambda = 0$, but there do not exist constants a, b such that $g(\tau y) = a + bg(y)$. A modified power transformation satisfying both conditions—$g(\tau y) = \tau g(y)$ and continuity in λ—is described in Exercise 18.7. None of these modifications has a parameter-independent Jacobian, so the Jacobian cannot be ignored in likelihood calculations.

		Team					
Species	Bark	I	II	III	IV	V	VI
spruce	no	6.4 F	10.9 E	9.8 D	7.5 B	4.6 A	4.1 C
pine	no	6.8 B	6.2 C	7.9 E	6.0 A	4.0 D	4.2 F
larch	no	12.7 E	13.4 A	12.5 B	7.3 C	6.1 F	7.4 D
spruce	yes	8.8 C	10.2 D	12.5 A	8.6 F	6.1 E	5.6 B
pine	yes	7.4 D	10.0 B	8.3 F	6.4 E	4.3 C	5.6 A
larch	yes	13.1 A	12.0 F	12.0 C	11.3 D	6.1 B	9.7 E

Table 18.1: Time in minutes taken by six teams to complete a woodcutting task using one of six available saws A–F.

18.2.3 Worked example

This example is taken from Bliss (1970, p. 440–441). The woodcutting efficiencies of three brands of saw were compared in a fractional factorial design using six cutting teams, three species of softwood (spruce, pine and larch) both with bark and without bark. The response variable is the time in minutes taken to complete the designated cutting task. The fractional factorial is embedded in a 6×6 Latin square whose columns correspond to six teams of workmen covering the range from experienced woodcutters to seasonal labourers. The letters correspond to six distinct saws, where A,D are duplicates of brand 1, B,E are duplicates of brand 2, and C,F are duplicates of brand 3.

Bearing in mind that the chief purpose of transformation is not so much to induce normality, but to achieve additivity of effects, two Gaussian models were selected as targets. In the first version, the mean of the transformed variable is additive in the four factors *species+bark+team+saw.id*, while the variances are constant, and the covariances are zero. This is a rank-14 sub-model of the standard Latin-square model, which has rank 16. The transformation model has two additional parameters, σ^2 and λ, making 16 total. In the second version, the mean is additive in the three factors *species+bark+saw.brand*, which is a subspace of dimension 6, while there are two additional variance components *team+saw.id*, making a total of ten parameters. Both profile log likelihoods for λ in Fig 18.1 have their maxima near $\hat{\lambda} = -0.34$; both 95% confidence intervals include $\lambda = 0$, but the identity is excluded. The conclusion is that the effects on the time scale are approximately multiplicative, so taking logs is the natural remedy, as indicated by Bliss. Most experienced statisticians would transform instinctively to the log scale on the grounds that additive effects on the time scale are less plausible than multiplicative effects.

As it happens, the variation between duplicate saws is small, but brand three is about 15% more efficient than the others. Mean cutting times are in the ratios 1.28:1.00:0.80 for larch:spruce:pine, with an additional factor of 1.14 for bark. There is substantial variation among the teams.

A crucial point in the computation of log likelihoods for Gaussian transformation models is that REML, or residual log likelihood, must not be used

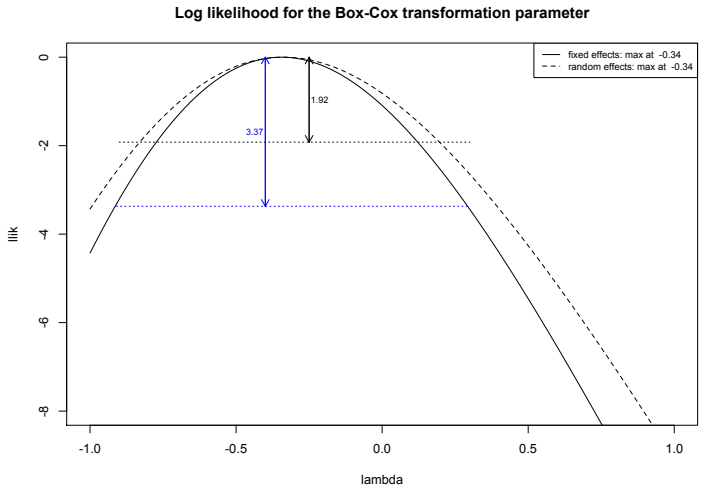

Figure 18.1: Log likelihood for the transformation parameter λ for two linear models.

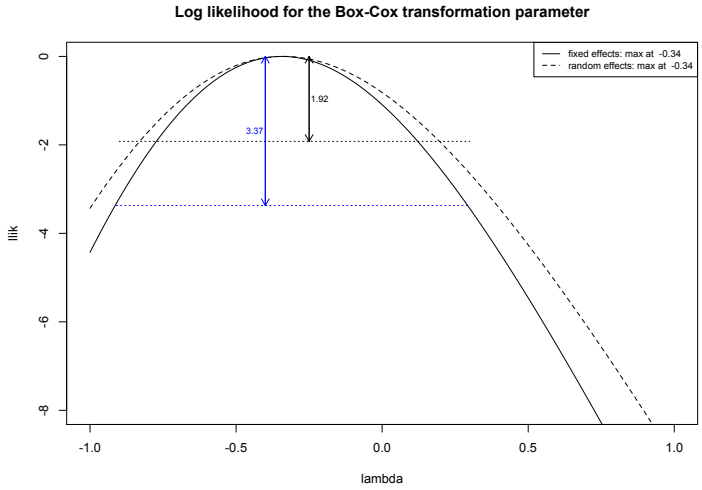

Figure 18.2: Log likelihood for the transformation parameter λ for two linear models.

under any circumstances. REML calculations are based on the residuals in $\mathbb{R}^{n-p}$, whereas the Jacobian is the determinant of a transformation $\mathbb{R}^n \to \mathbb{R}^n$. These are not compatible; the power transformation does not act on residuals. As a function of λ the residual likelihood criterion is not a likelihood in the conventional Radon-Nikodym sense. If the REML criterion were used with the Jacobian as in (18.2), the plots shown in Fig. 18.1 would look substantially different. Details are discussed in the next section.

Figure 18.1 indicates that a $1-\alpha$ confidence interval for the transformation parameter may be obtained by the likelihood-ratio formula

$$\{\lambda : l(\hat{\lambda}; y) - l(\lambda; y) \geq \tfrac{1}{2}\chi^2_{1,\alpha}\},$$

which is based on large-sample distribution theory: For $\alpha = 0.05$, the cutoff allowance $1.92 = 1.96^2/2$ is indicated on the log likelihood plot. In the present setting, the effective sample size is the residual degrees of freedom, which is $36 - 14 = 22$. Given that we are interested only in whether the interval includes zero, the asymptotic approximation is reasonably adequate. However, the coverage level can be improved appreciably by replacing the $\frac{1}{2}\chi^2_1$ threshold with the threshold based on Fisher's F-ratio,

$$\frac{n}{2}\log\left(1 + \frac{F_{1,n-p-1,\alpha}}{n-p-1}\right),$$

which is the exact threshold for $1-\alpha$-coverage in the setting of nested linear models. The 95% F-threshold for $n = 36$ and $p = 14$ is 3.37, which is also shown in Fig. 18.1 for comparison. The greater allowance produces a wider interval, but does not materially alter the conclusion or subsequent analysis.

An analysis-of-variance decomposition on the log scale shows that the interactions *species.bark* and *bark.brand* are negligible. Bark removal reduces the *mean* log cutting time by an estimated 0.152 ± 0.031 units for each species and each brand, so the cutting-time distribution is reduced multiplicatively by about 14%.

Notice that the effect of treatment is not to modify the *response* as a random variable $Y \mapsto Y + \tau$, but to modify the *distribution*. Mathematically speaking, the treatment parameter is a real number τ whose effect for linear Gaussian models is ordinarily the transformation

$$N(\mu, \sigma^2) \overset{\tau}{\longmapsto} N(\mu + \tau, \sigma^2)$$

from the set of Gaussian distributions into itself. The distinction between a sample-space transformation and a distributional transformation is more clear-cut for a Poisson or Bernoulli or beta model in which

$$\begin{aligned}
\mathrm{Po}(\mu) &\overset{\tau}{\longmapsto} \mathrm{Po}(\mu e^{\tau}),\\
\mathrm{Ber}\Big(\frac{e^{\eta}}{1+e^{\eta}}\Big) &\overset{\tau}{\longmapsto} \mathrm{Ber}\Big(\frac{e^{\eta+\tau}}{1+e^{\eta+\tau}}\Big),\\
\mathrm{Beta}(\alpha, \beta) &\overset{\tau}{\longmapsto} \mathrm{Beta}\big(\alpha e^{\tau/2}, \beta e^{-\tau/2}\big),
\end{aligned}$$

are homomorphisms from $\mathbb{R}$ (as a group) into a group of transformations on Poisson or Bernoulli or beta distributions. In the Gaussian case, the action on probability distributions is induced by the additive transformation $Y \mapsto Y+\tau$ on the observation space. But there is no comparable sample-space transformation $\{0,1\} \to \{0,1\}$ for the Bernoulli model, or $\mathbb{R}_+ \to \mathbb{R}_+$ for the Poisson model, or $(0,1) \to (0,1)$ for the beta model.

To express the rationale in more concrete terms, suppose that the response distribution for one control unit is F belonging to a given class $\mathcal{F}$, and that the treatment effect is a real number τ. Then the response distribution for a comparable treated unit is τF, also belonging to the same set $\mathcal{F}$. In general, the response distribution for control unit u depends on the covariate x_u, so not all controls have the same distribution. Whatever the control distribution $F \in \mathcal{F}$ may be, the distribution for a comparable treated unit u' having the same covariate $x_u = x_{u'}$ is τF. Thus $(F, \tau F)$ is not simply a fixed pair of distributions in $\mathcal{F}$, but τ is a function $\mathcal{F} \to \mathcal{F}$. Group action implies that $\tau\colon \mathcal{F} \to \mathcal{F}$ is invertible; the inverse is $-\tau$.

In addition to the logit model, probit and complementary log-log models also determine group actions on Bernoulli distributions; the identity-link model does not. Crowder's (19??) beta regression model is a group action on beta distributions,

$$(\alpha, \beta) \overset{\tau}{\longmapsto} \left(\frac{\alpha(\alpha+\beta)e^\tau}{\alpha e^\tau + \beta}, \frac{\beta(\alpha+\beta)}{\alpha e^\tau + \beta} \right),$$

in which the treatment effect is multiplicative on the ratio: $\alpha/\beta \mapsto e^\tau \alpha/\beta$. In the first version of the beta model, the product $\alpha\beta$ is invariant, and in the second, the sum is invariant. Note that $Y \sim \text{Beta}(\alpha, \beta)$ implies $\bar{Y} = 1 - Y$ is distributed as $\text{Beta}(\beta, \alpha)$, with transposition of components. Both beta examples are consistent in the sense that, the distribution obtained from (α, β) with transposition followed by τ is the same as the that obtained by $-\tau$ followed by transposition.

If we fit the standard additive model *team+species+bark+brand*, the estimate reported for the brand 3 versus brand 1 contrast is -0.1484 with standard error 0.040 on 25 degrees of freedom, suggesting fairly strongly that brand 3 is more efficient than brand 1. This analysis is potentially misleading because, with only six distinct saws, there cannot be more than five degrees of freedom for the estimation of between-saw variability. The situation for the bark contrast is markedly different because there are 36 logs. The natural analysis based on the design associates with each saw an independent additive random effect, so that the six saw averages

saw	*A*	*B*	*C*	*D*	*E*	*F*
mean	2.122	2.060	1.975	2.070	2.156	1.920

are independent with equal variance to be estimated from the three replicate pairs. This reduction to saw averages does not affect the point estimate for any brand contrast, but it reduces the degrees of freedom to three, and it increases the standard error of each pairwise brand contrast to 0.050. The net effect is

to decrease the F-ratio for brand effects from $F_{2,25} = 9.98$ to $F_{2,3} = 6.38$. The preferred Gaussian model is one that incorporates independent and identically distributed additive effects for replicate saws. Admittedly, the variance component for replicate saws is not very large, but the fact that only six saws were used in the design is decisive.

18.2.4 Transformation and residual likelihood

If the transformation under consideration can be regarded as an invertible transformation on residuals, the residual likelihood modulo the subpace $\mathcal{X}$ may be used in place of (18.2). A transformation $g\colon \mathbb{R}^n \to \mathbb{R}^n$ may be regarded as a transformation on residuals if and only if each coset $y + \mathcal{X}$ has an image that is either a coset or a subset of a coset. In that case, g induces a transformation $\mathbb{R}^n/\mathcal{X} \to \mathbb{R}^n/\mathcal{X}$ on residuals, which is assumed to be measurable with respect to the Borel subsets of $\mathbb{R}^n/\mathcal{X}$. For example, a linear transformation $g\colon \mathbb{R}^n \to \mathbb{R}^n$ induces a transformation on residuals if and only if $g\mathcal{X} \subset \mathcal{X}$. Unfortunately, a component-wise non-linear transformation does not induce a measurable transformation on residuals except for trivial cases such as $\mathcal{X} = 0$ and $\mathcal{X} = \mathbb{R}^n$. Even $\mathcal{X} = \mathbf{1}$ fails. Thus, residual likelihood is not available as an option for the comparison of response transformations.

To see why and how a naive version of REML fails, we compare the standard log likelihood with a REML-style criterion first proposed by Shi and Tsai (2002) and subsequently by Gurka, Edwards, Muller and Kupper (2006). For the power-transformation model $gY \sim N(X\beta, \Sigma)$, the two criteria are

$$\begin{aligned} l &= -\tfrac{1}{2}(gy - \mu)'\Sigma^{-1}(gy - \mu) - \tfrac{1}{2}\log\det\Sigma + (\lambda - 1)\sum\log(y_i), \\ l^\dagger &= l(\mu, \Sigma, \lambda) - \tfrac{1}{2}\log\det(X'WX) + \tfrac{1}{2}\log\det(X'X), \end{aligned}$$

where $W = \Sigma^{-1}$. Here, $gy = y^\lambda/\lambda$ for $y > 0$ is the component-wise power transformation, so that the log Jacobian is $(\lambda - 1)\sum\log(y_i)$. In either case, maximization over the space of mean-vectors $\mu \in \mathcal{X}$ gives $\hat{\mu} = P(gy)$, the W-orthogonal projection of the transformed vector. The profile criteria are

$$\begin{aligned} l(\Sigma, \lambda; y) &= -\tfrac{1}{2}\, gy'WQgy - \tfrac{1}{2}\log\det\Sigma + (\lambda - 1)\sum\log(y_i), \\ l^\dagger(\Sigma, \lambda; y) &= l(\Sigma, \lambda; y) - \tfrac{1}{2}\log\det(X'WX) + \log\det(X'X). \end{aligned}$$

Suppose that two statisticians are asked to examine the same data, which is concerned with vehicle fuel economy. Statistician I analyzes the consumption rates in miles per gallon, and statistician II in kilometres per litre, so the pairs of numbers differ by a constant multiple: $y_i^{(1)} = \tau y_i^{(2)}$ with $\tau \simeq 2.8$. For each λ, the transformed values differ by a parameter-dependent factor τ^λ, the associated variance matrices satisfy $\Sigma^{(1)} = \tau^{2\lambda}\Sigma^{(2)}$, and the inverse matrices satisfy $W^{(1)} = \tau^{-2\lambda}W^{(2)}$. As we should expect, the log likelihood function is scale-invariant in the sense that, the two versions differ by an additive constant

$$l(\tau^{2\lambda}\Sigma, \lambda; \tau y) = l(\Sigma, \lambda; y) - n\log(\tau).$$

Invariance with respect to scalar multiplication means that two statisticians analyzing the same data on different scales must arrive at the same conclusion regarding the transformation parameter. By contrast, the two versions of the modified criterion differ linearly in λ:

$$l^\dagger(\tau^{2\lambda}\Sigma, \lambda; \tau y) = l^\dagger(\Sigma, \lambda; y) - n\log(\tau) + \lambda p \log(\tau),$$

where $p = \dim(\mathcal{X})$. Lack of invariance means that two statisticians using $l^\dagger$ as the selection criterion are liable to arrive at contradictory conclusions for λ. For the continuity-modified transformation $gy = (y^\lambda - 1)/\lambda$, the analysis is slightly more complicated, but the conclusions are essentially the same provided that $\mathbf{1} \subset \mathcal{X}$.

The extreme example $X = I_n$ and $\mathcal{X} = \mathbb{R}^n$ is not of practical interest because $Q = 0$ implies that the residual is identically zero. But it suffices to show that $l^\dagger$ is a non-trivial function of the observations, and thus not a function of residuals. With $X = I_n$ the three determinantal terms vanish, leaving

$$l^\dagger(\Sigma, \lambda; y) = (\lambda - 1)\sum \log(y_i).$$

In the absence of information, constancy in Σ is correct, but linearity in λ is misleading. The slope is positive if the geometric mean observation is greater than one, and negative otherwise, so scale conversion can change the slope from positive to negative or vice-versa.

18.3 Quantile-matching transformation

18.3.1 Likelihood function

In certain 'big-data' settings such as the analysis of micro-array gene-expression data, transformation to a marginal reference distribution is sometimes recommended as a way to reduce the impact of unwanted structural effects. Quantile matching is a strictly monotone transformation $h = G^{-1} \circ F$, which is defined by a domain distribution F and a target distribution G. Both distribution functions are assumed to be strictly monotone and differentiable. Quantile matching is applied component-wise to the data, and transforms $Y \sim F$ into $hY \sim G$. The empirical version, denoted by $\tilde{h}$, transforms the finite set $\{y_1, \ldots, y_n\}$ into specific quantiles of G:

$$\begin{aligned} h &: y \longmapsto F(y) \longmapsto G^{-1}\big(F(y)\big) \\ \tilde{h} &: y \longmapsto \tilde{F}_n(y) \longmapsto G^{-1}\big(\tilde{F}_n(y)\big). \end{aligned}$$

For the specific requirements of this section, $\tilde{F}$ is a strictly monotone continuously differentiable function satisfying $0 < \tilde{F}(t) < 1$. At each observation point $y \in \{y_1, \ldots, y_n\}$, the value is the average of the left and right limits of the empirical distribution function

$$\tilde{F}_n(y) = \tfrac{1}{2}\hat{F}_n(y^-) + \tfrac{1}{2}\hat{F}_n(y^+).$$

Elsewhere in the domain, $\tilde{F}_n(t)$ is subject to differentiability and strict monotonicity, but, apart from the sample points, the values are otherwise unspecified. The numbers $\tilde{F}(y_1), \dots, \tilde{F}(y_n)$ are the uniform sample quantiles in $(0, 1)$, and the target G-quantiles are the transformed values

$$q_{i:n} = \tilde{h}(y_i) = G^{-1}(\tilde{F}(y_i)),$$

taken with multiplicity in ascending order. If there are no ties, the uniform quantiles are the numbers $(2i - 1)/2n$ for $1 \le i \le n$.

The Jacobian of the transformation $h \colon \mathbb{R}^n \to \mathbb{R}^n$ is the product of the derivatives at the domain points

$$\prod_{i=1}^{n} h'(y_i) = \prod_{i=1}^{n} \frac{F'(y_i)}{g(h(y_i))},$$

so the log Jacobian and its empirical version are

$$\sum \log F'(y_i) - \sum \log g(h(y_i)) \quad \text{and} \quad \sum \log \tilde{F}'(y_i) - \sum \log g(q_{i:n}),$$

where $g = G'$ is the target density. The last term is a quadrature sum, which is an approximation to the entropy integral

$$\tilde{J}(G) = n^{-1} \sum \log g(q_{i:n}) = \int \log g(x)\, dG(x) + O(n^{-1}).$$

From (18.1), the profile log likelihood function for the quantile-matching transformation $\tilde{h}$ is

$$-\tfrac{1}{2} \log \det \hat{\Sigma}_h + \sum_i \log \tilde{F}'(y_i) - \sum \log g(q_{i:n}), \tag{18.3}$$

where $\hat{\Sigma}_h$ is the maximum-likelihood estimate after transformation. However, the derivatives $\tilde{F}'(y)$ are not readily available, so the log likelihood (18.3) is not computable.

Now consider two quantile-matching transformations, which are defined by their target distributions G_0, G_1. From (18.3), the profile log likelihood ratio of G_1 to G_0, is

$$-\tfrac{1}{2} \log \det(\hat{\Sigma}_1 \hat{\Sigma}_0^{-1}) - n\, \tilde{J}(G_1) + n\, \tilde{J}(G_0), \tag{18.4}$$

If n is sufficiently large, the quadrature sums $\tilde{J}(G_0)$ and $\tilde{J}(G_1)$ can be replaced with the corresponding integrals. The quadrature errors are typically $O(n^{-1})$ for both $J(G_0)$ and $J(G_1)$, but if the distributions are contiguous or similar, the quadrature error for the difference is $o(n^{-1})$, and thus negligible for present purposes.

Figure 18.3: Log likelihood for α in the quantile function (18.5)

18.3.2 Simulated example

As an illustration, we simulate data from a row-column design with independent and identically distributed additive row and column effects as follows:

```
nrows <- 50;  ncols <- 30;  n <- nrows*ncols
row <- gl(nrows, 1, n);  col <- gl(ncols, nrows, n)
set.seed(3142)
mu <- rcauchy(nrows)[as.numeric(row)] + rcauchy(ncols)[as.numeric(col)]
y <- 5 + mu + rnorm(n);  hist(y, nclass=50)
```

The histogram in Figure 18.2 is reasonably symmetric but markedly non-Gaussian.

The percentile-matching model is determined by the subspace $\mathcal{X} = row{+}col$, the covariance specification $\Sigma \propto I_n$, plus the quantile function $q\colon (0,1) \to \mathbb{R}$, which we take to be of the form

$$q(p) = p^\alpha/\alpha - (1-p)^\alpha/\alpha \tag{18.5}$$

for some real number α. The limit $\alpha \to 0$ is the logistic model. In practice, it suffices to focus on the range $-1 < \alpha < 1$, or some subset thereof. We now compute the profile log likelihood for a range of parameter values as follows:

```
pc <- (rank(y) - 1/2)/n
alpha <- seq(-0.25, 0.5, 0.05);  llik <- rep(0, along=alpha)
for(i in 1:length(alpha)){
   a <- alpha[i]
   if(a==0) gy <- log(pc/(1-pc)) else gy <- (pc^a - (1-pc)^a)/a
   fit <- lm(gy~row+col)
   s2 <- sum(fit$resid^2)/n
   llik[i] <- -n*log(s2)/2 + sum(log(pc^(a-1) + (1-pc)^(a-1)))
   }
```

The next step is to plot the log likelihood as a function of α. The set of quantile functions does not include the Gaussian or probit function, so the log likelihood is computed separately and indicated on the plot. Ordinarily, Gaussian quantile matching is quite effective, but Fig. 18.2 shows that, for these data, the logistic quantile-matching function with $\alpha = 0$ works appreciably better.

Since the response values were generated additively according to the Gaussian model, the optimal transformation in this setting is linear or affine. But the identity and other linear transformations are not among the options accessible by quantile matching, which is a function of the rank vector only. Nonetheless, among the transformations considered, the correlation matrix shows that

the quantile transformation for $\hat{\alpha} = -0.05$ is maximally correlated with the optimum:

Correlations with quantile-transformed variables				
	$\hat{\alpha}$	logistic	Gaussian	t_7
Identity	0.927	0.917	0.888	0.919

The quantiles of the Student t family provide a viable alternative to (18.5). The optimum, t_7 in this instance, is better than Gaussian, and approximately as effective as logistic matching.

In the preceding analysis, the likelihood is determined by the mean model $\mathcal{X} = \textit{row+col}$ and the covariance model $\Sigma \propto I_n$. One reasonable variation in the present setting is to use an additive Gaussian random-effects model for gY with $\mathcal{X} = \mathbf{1}$, and Σ a linear combination of the block matrices I_n, `row` and `col`. For a balanced design such as this, maximum-likelihood estimates of all four parameters are available in closed form, so the computations are not onerous. The profile log likelihood plot for α looks much the same as Fig. 18.2 except that all log likelihood values are reduced by approximately 250 units. The reduction is not constant in α, but the variation is insufficient to change the conclusion in a material way.

18.4 Exercises

18.1 Let Y be a non-negative random variable with cumulants κ_r such that $\kappa_r/\kappa_1^r = O(\rho^{r-1})$ as $\rho \to 0$. In other word, the scale-free variable $Z = Y/\kappa_1$ has variance $\rho = \kappa_2/\kappa_1^2$, which is the squared coefficient of variation of Y, and the higher-order scale-free cumulants are $O(\rho^{r-1})$. Show that the cumulants of the power-transformed variable are

$$\begin{aligned} E(Z^\lambda) &= 1 + \frac{(\lambda-1)\kappa_2}{2\kappa_1^2} + o(\rho); \\ \operatorname{var}(Z^\lambda) &= \frac{\kappa_2}{\kappa_1^2} + o(\rho); \\ \operatorname{cum}_3(Z^\lambda) &= \frac{\kappa_3}{\kappa_1^3} + 3(\lambda-1)\frac{\kappa_2^2}{\kappa_1^2} + o(\rho^2). \end{aligned}$$

Hence deduce that the approximate symmetry-inducing power transformation is $\hat{\lambda} = 1 - \kappa_1\kappa_3/(3\kappa_2^2)$.

18.2 Wilson-Hilferty transformation: Show that the rth cumulant of the exponential distribution is $\kappa_r = \kappa_1(r-1)!$, and hence that $Y^{1/3}$ is approximately symmetrically distributed.

18.3 Show that the rth cumulant of the Poisson distribution is $\kappa_r = \kappa_1$, and hence that $Y^{2/3}$ is approximately symmetrically distributed.

18.4 Show that the transformation $\mathbb{R}^n \to \mathbb{R}^n$ defined by

$$\bar{u} \mapsto \bar{u} + \text{const}, \quad u_i - \bar{u} \mapsto \lambda(u_i - \bar{u})$$

is linear and invertible with Jacobian $J = |\lambda|^{n-1}$. Here, $\bar{u}$ is the mean of the components of the vector $u \in \mathbb{R}^n$, and λ is a non-zero constant.

18.5 Consider the non-component-wise transformation

$$y_i \mapsto \frac{y_i^\lambda}{\lambda \dot{y}^{\lambda-1}}$$

where $\dot{y}$ is the geometric mean of the components of $y \in \mathbb{R}^n_+$. Using the result of the previous exercise, show that the modified transformation is invertible $\mathbb{R}^n_+ \to \mathbb{R}^n_+$ with Jacobian $J = |\lambda|^{-1}$.

18.6 As a function of λ, show that the the transformation $\mathbb{R}^n_+ \to \mathbb{R}^n$

$$(gy)_i = \frac{y_i^\lambda - 1}{\lambda \dot{y}^{\lambda-1}}$$

is continuous at $\lambda = 0$, and that the Jacobian is the absolute value of

$$\frac{1}{\lambda} + \frac{\lambda - 1}{n\lambda} \sum y_i^{-\lambda}.$$

Find the limits for $\lambda = 0, \pm 1$. Discuss the implications regarding invertibility? For $\tau > 0$, show that $g(\tau y)$ is not expressible in the form $a + bg(y)$ for any constants a, b depending on τ, λ.

18.7 For $\tau > 0$, show that the modified transformation $\mathbb{R}^n_+ \to \mathbb{R}^n$

$$(gy)_i = \dot{y} + \frac{y_i^\lambda - \dot{y}^\lambda}{\lambda \dot{y}^{\lambda-1}}$$

is continuous at $\lambda = 0$ and satisfies $g(\tau y) = \tau g(y)$. What are the implications for statistical applications? Show that the Jacobian is

$$\frac{1}{\lambda} + \frac{\lambda - 1}{n\lambda} \dot{y}^\lambda \sum y_i^{-\lambda},$$

which is positive, and that the limits for $\lambda \to 0$ and $\lambda \to 1$ are equal.

18.8 For which values of α is the transformation

$$g_\alpha(x) = \frac{(-\log(1-x))^\alpha}{\alpha} - \frac{(-\log(x))^\alpha}{\alpha}$$

differentiable and strictly monotone $(0, 1) \to \mathbb{R}$?

18.9 For the simulated data in the section 18.?, compute and plot the profile log likelihood for the preceding family of transformations as a function of α in a suitable range that includes the maximum. Comment on any unusual aspects of the likelihood function.

18.10 Repeat the preceding exercise taking $\mathcal{X} = \mathbf{1}$, and Σ a linear combination of the block matrices I_n, `row` and `col`.

Chapter 19

Missing values

19.1 Pattern of missing components

19.1.1 Complementary subsets and subspaces

In a typical units-by-variables setting, the 'complete' observation is a function $Y\colon NJ \to \mathbb{R}$, where $NJ = [N] \times [J]$ is the product set. The pattern of recorded components is a subset $r \subset NJ$, and the pattern of missing components is the complementary subset $\bar{r}$. These subsets are identified with their indicator functions $r\colon NJ \to \{0,1\}$ and $\bar{r} = 1-r$, so the observation consists of the mask r together with the component-wise product $Y_{\text{obs}} = (r,\, r\cdot Y)$. The complementary unobserved part is $Y_{\text{mis}} = (\bar{r},\, \bar{r}\cdot Y)$. Either Y_{obs} or Y_{mis} determines the mask r, while the sum $Y_{\text{obs}} + Y_{\text{mis}} = (NJ, Y)$ determines only Y, not the mask.

In general, $Y_{\text{obs}}(i,j) = 0$ implies either $Y_{i,j} = 0$ or $r_{i,j} = 0$. If $Y_{i,j} = 0$ is not physically possible, or if it is an event of probability zero, then $Y_{\text{obs}}(i,j) = 0$ implies $r_{i,j} = 0$, in which case there is no ambiguity in writing $Y_{\text{obs}} = r\cdot Y$ and $Y_{\text{mis}} = \bar{r}\cdot Y$.

No distinction is made in the notation between a subset $r \subset NJ$ and its indicator function or indicator vector. But, in general, a set or subset consists of *elements*, while a vector or matrix has *components*, which are the coefficients with respect to the indicator basis, one indicator function for each element in NJ.

To each subset $r \subset NJ$ there corresponds a vector subspace $\mathbb{R}^r \subset \mathbb{R}^{NJ}$:

$$\mathbb{R}^r = \{r\cdot v : v \in \mathbb{R}^{NJ}\} = \{x \in \mathbb{R}^{NJ} : \bar{r}\cdot x = 0\}$$

where 0 is the zero vector in $\mathbb{R}^{NJ}$. The masking function $Y \mapsto r\cdot Y$ is a linear transformation $\mathbb{R}^{NJ} \to \mathbb{R}^{NJ}$, in fact the unique linear projection with image $\mathbb{R}^r$ and kernel $\mathbb{R}^{\bar{r}}$. The condition $x \in \mathbb{R}^r$ implies that each component $x(i,j)$ is a real number and that $x(i,j) = 0$ for $(i,j) \notin r$. Evidently, $\dim(\mathbb{R}^r) = \#r$ is equal to the number of elements in the set r, or the number of non-zero components in the indicator matrix r.

Let W be a vector space, and let U, V be complementary subspaces. Complementary in W means (i) $U \cap V = 0$, i.e., the intersection is the zero subspace, and (ii) $\text{span}(U, V) = U + V = W$. Condition (ii) implies that each vector $x \in W$ can be decomposed as a sum $x = u + v$ with $u \in U$ and $v \in V$, and condition (i) implies that this decomposition is unique. Equivalently, to each $x \in W$ there corresponds an ordered pair $x \mapsto (u, v)$ with $u \in U$ and $v \in V$ such that $x = u+v$. If U, V are complementary in W, we write $W = U \oplus V$, meaning that W is the direct sum of subspaces. Thus $W = U \oplus V$ implies $W = U + V$ and $U \cap V = 0$.

For the setting under discussion here, where $r \subset NJ$, and the subspaces are defined component-wise by $U = \mathbb{R}^r$ and $V = \mathbb{R}^{\bar{r}}$, the projections onto U and V are $u = r \cdot x$ and $v = \bar{r} \cdot x$ respectively. These projections are complementary but not necessarily orthogonal because no inner product has been specified.

19.1.2 Probability distributions

Let $r \subset NJ$ be a fixed subset, let $U = \mathbb{R}^r$, $V = \mathbb{R}^{\bar{r}}$ be complementary subspaces, and let P be a probability distribution with density $p(x)$ at $x \in \mathbb{R}^{NJ}$. The linear projection $x \mapsto r \cdot x$ has kernel V, and $A \times V = A + V$ is the inverse image of $A \subset U$, so the marginal distribution P_r on U is such that

$$P_r(A) = P(A \times V)$$

for each Borel subset $A \subset U$. Since $x \mapsto (u, v)$ is a linear transformation with Jacobian one (a coordinate permutation), the marginal density at u is

$$p_r(u)\, du = P(du \times V) = du \int_{v \in V} p(u, v)\, dv.$$

If Y is a random variable distributed as P, the marginal variable or masked variable $r \cdot Y$ is distributed as $P_r(\,\cdot\,)$ on the subspace U.

To make a connection with the notation typically used in the literature on missing data, $\bar{r}$ is a mask concealing part of the response $Y_{\text{mis}} = \bar{r} \cdot Y$, the values $Y_{\text{obs}} = r \cdot Y$ are recorded on r, and both $Y_{\text{obs}}, Y_{\text{mis}}$ are regarded as $N \times J$ matrices. In this setting $r \subset NJ$ is an arbitrary subset, and the transformation from ordered pairs to ordered pairs $(r, Y) \mapsto (Y_{\text{obs}}, Y_{\text{mis}})$ is one-to-one. If (u, v) are complementary vectors in $\mathbb{R}^{NJ}$ in the sense that the support subsets

$$r = I_u = \{(i, j) : u_{i,j} \neq 0\} \quad \text{and} \quad I_v = \{(i, j) : v_{i,j} \neq 0\}$$

are complementary in NJ, the inverse transformation is $(u, v) \mapsto (I_u, u + v)$ on ordered pairs. Thus, when we say that Y_{obs} is the observation, it is assumed implicitly that $Y_{\text{obs}}(i, j) = 0$ if and only if $r(i, j) = 0$, which is true with probability one if Y is continuously distributed. Otherwise, it is necessary to distinguish these possibilities by introducing a distinct symbol such as $*$ or 0^* for censored components, $r = I_{u \neq 0^*}$, and so on. In arithmetic operations, 0^* behaves as zero.

In the computational literature, a different convention is employed in which the elements of NJ are implicitly stored in a specific order, and the components

of every matrix such as Y or r are listed in parallel. Then $Y_{[r]} \equiv Y[r]$ is the restriction of Y to $r \subset NJ$, which is a list $\{Y_{i,j} \colon (i,j) \in r\}$ consisting of $\#r$ numbers presented in the same relative order without gaps. In other words, $Y_{[r]}$ is not a set or a matrix, but an ordered list of real numbers in the same relative order as the elements of $r \subset NJ$. A simple example with $N = 4$ and $J = 3$ illustrates the idea:

$$Y = \begin{pmatrix} 1.2 & 3.6 & 2.7 \\ 3.1 & 2.2 & 3.7 \\ 4.3 & 0.7 & 1.9 \\ 4.7 & 0.0 & 2.1 \end{pmatrix}, \qquad r = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}, \qquad Y_{\text{obs}} = r \cdot Y = \begin{pmatrix} 0^* & 3.6 & 2.7 \\ 3.1 & 0^* & 0^* \\ 0^* & 0^* & 1.9 \\ 0^* & 0.0 & 0^* \end{pmatrix}$$

$$Y_{[r]} = Y[r] = (3.1, 3.6, 0.0, 2.7, 1.9).$$

Note that $(r \cdot Y)(4,1) = 0^*$ implies $r(4,1) = 0$, whereas $(r \cdot Y)(4,2) = 0.0$ implies $Y(4,2) = 0.0$ and $r(4,2) = 1$.

The expression for $Y_{[r]}$ presumes that the elements of NJ are stored column-wise. If the elements of NJ are stored row-wise, we have

$$Y'_{[r]} = (3.6, 2.7, 3.1, 1.9, 0.0),$$

which is the same set but a different vector. Matrix transposition changes the implicit storage order, so that $Y'_{[r]} \equiv Y'[r']$.

For any given $r \subset NJ$, the matrix representation $Y_{\text{obs}} = r \cdot Y$ the list representation $Y_{[r]}$ and the transposed list $Y'_{[r]}$ are equivalent in the sense that any one may be computed from the other. For fixed r, the random variables $Y_{\text{obs}} \cong Y_{[r]} \cong Y'_{[r]}$ are equivalent in the sense that they determine the same σ-field. The situation for variable or random $r \subset NJ$ is entirely different, so the distinction between Y_{obs}, $Y_{[r]}$ and $Y'_{[r]}$ as random variables is not a matter of stylistic preference, but a matter of informational content. The notation implies correctly that Y_{obs} is the information recorded, and Y_{obs} determines both $Y_{[r]}$ and $Y'_{[r]}$ and also r, but $Y_{[r]}$ without r does not determine either $Y'_{[r]}$ or Y_{obs}.

19.1.3 Random masking and MAR

Let P be the joint distribution of the pair (R, Y). The *missing-value mechanism* is the conditional distribution

$$q(r, y) = P(R = r \mid Y = y),$$

which is a family of of probability distributions on subsets of NJ, one distribution for each point $y \in \mathbb{R}^{NJ}$. In this context, 'all $y \in \mathbb{R}^{NJ}$' is understood in the almost-all P sense, in essence all y at which the marginal distribution has positive density. The missing-value mechanism is said to be *missing at random* (MAR) if

$$q(r, y) = q(r,\, y + \bar{r} \cdot x) \tag{19.1}$$

for all $r \subset NJ$ and all real matrices x, y. In particular, $x = -y$ gives $q(r, y) = q(r, r \cdot y)$. This is the strong MAR condition (Rubin, 1976), also called MAAR, which can be expressed equivalently in the form

$$q(r, y) = q(r, y') \tag{19.2}$$

for all $r \subset NJ$ and all pairs y, y' such that $r \cdot y = r \cdot y'$, or, equivalently, for all pairs y, y' such that $y[r] = y'[r]$.

If (19.1) is satisfied for a particular subset, for example $r = NJ$ or $r = \emptyset$, we say that the masking distribution is MAR at r. Condition (19.1) is null for $r = NJ$, so every distribution is automatically MAR at NJ. At the other extreme, the distribution is MAR at the empty set if and only if the event $R = \emptyset$ is independent of the random variable Y.

Consider a family of distributions P_θ differing only in their Y-marginal distributions $p_\theta(y)$, so that the joint density at (r, y) is of the form $p_\theta(y)\, q(r, y)$. The MAR condition is a restriction on the conditional distribution given Y, so each distribution in the family is MAR if q satisfies (19.1).

Example 1: Optional stopping 1. Let $Y = (Y_0, Y_1, \ldots)$ be a real-valued process in discrete time, and let $B \subset \mathbb{R}$ be a subset of the state space such that the first exit time $T(Y) = \min\{t \geq 0 : Y_t \notin B\}$ is finite with probability one. The values $Y_{\rm obs} = (Y_0, \ldots, Y_T)$ are observed, so $R = [0, T]$ and $\bar{R} = [T + 1, \infty]$. Note that the observed sequence contains at least one value $Y_T \in \bar{B}$ with $T \geq 0$, so R is not empty.

It is helpful in the first instance to consider a specific case with $B = [0, \infty)$ and a specific sequence

$$y = (0.0, 0.5, 0.7, -0.4, 0.3, \ldots)$$

with $y_0 = 0.0 \in B$ and $T(y) = 3$. Then $\text{pr}(R = r \mid Y = y) = 1$ for $r = [0, 3]$ and zero for all other values. Moreover, if y' is any other sequence such that $y'[r] = y[r]$ are equal on r, then $\text{pr}(R = r \mid Y = y')$ is also equal to one for $r = [0, 3]$ and zero otherwise. Thus, the MAR condition is satisfied for this y-sequence, and all other sequences having at least one negative component.

Example 2: Optional stopping 2. Consider the same set-up as in the preceding example except that the value Y_T on exit from B is not recorded. Thus, $R = [0, T - 1]$ if $T \geq 1$ and $R = \emptyset$ otherwise. For the particular sequence y shown above, $r = \{0, 1, 2\}$ and $y' = (0, 0.5, 0.7, 0.4, -0.3, \ldots)$ we have $y[r] = y'[r]$ but $q(r, y) = 1$ whereas $q(r, y') = 0$. The MAR condition fails unless B is such that $R = \emptyset$ with probability one,

Example 3: Deterministic censoring. Let $B \subset \mathbb{R}$ be given, and let $R_{i,j} = 1$ if $Y_{i,j} \notin B$, and zero otherwise. In other words, components of Y are censored if they belong to B, so $Y_{\rm obs}(i, j) = Y_{i,j}$ if $Y_{i,j} \in B$, and $Y_{\rm obs}(i, j) = 0^*$ otherwise. Let $r \subset NJ$ be given. The MAR condition (1') is satisfied if, $q(r, y) = q(r, y')$ for each pair of matrices y, y' that are equal on r. In particular, $q(\emptyset, y) = q(\emptyset, y')$ for all pairs implies that the event $R = \emptyset$ must be independent of Y. By definition, $R = \emptyset$ if and only if $Y \in B^{NJ}$, so the MAR condition is satisfied only if the event

$Y \in B^{NJ}$ is independent of itself, i.e., $\mathrm{pr}(Y \in B^{NJ})$ is either zero or one. A similar argument shows that the marginal event $Y_{i,j} \in B$ must be independent of $Y_{i,j}$. In other words, the MAR condition is not satisfied unless the censoring set is trivial.

Example 4: Let $\alpha > 0$, let $Y = (Y_0, Y_1, Y_2)$ be a random permutation of $(0, 1, 2)$, and R a random subset of $[3] = \{0, 1, 2\}$ such that

$$q_\alpha(r, y) = \frac{\prod_{i \in r}(\alpha + y_i)}{(\alpha + 1)(\alpha + 2)(\alpha + 3)}. \tag{2}$$

The empty product is defined to be one, so that R is empty with probability $1/(\alpha + 1)$[↑3], independently of Y. Since $q_\alpha(r, y) = q_\alpha(r, r \cdot y)$ depends only on the component-wise product, the MAR condition is satisfied for every r and every y having positive probability. Equivalently, $y[r] = y'[r]$ implies $q(r, y) = q(r, y')$, satisfying $(1')$.

Example 3 suffices to illustrate the difference between Y_{obs} and $Y_{[r]}$. Suppose $Y_{[r]} = 2$, implying $R \subset [3]$ is a singleton. The pair $(Y_{[r]}, R)$ determines Y_{obs}, and the possible values are $(2, *, *)$, $(*, 2, *)$ and $(*, *, 2)$. These occur with equal probability if the marginal distribution of Y is uniform on the six permutations.

MAR is a property of a joint probability distribution, which depends only on the conditional distribution given Y. In the case of a family of probability distributions such as example 3, it is possible that some members satisfy the MAR condition and other not. If each distribution satisfies the MAR condition, we say that the family or statistical model is MAR.

19.1.4 Conditional independence

Let X, Y, Z be random variables defined on the same probability space. If the joint distribution is such that

$$\mathrm{pr}(X \in A, Y \in B \mid Z) = \mathrm{pr}(X \in A \mid Z)\ \mathrm{pr}(Y \in B \mid Z)$$

then the events $X \in A$ and $Y \in B$ are conditionally independent given Z. If this condition holds for all events A and B, then the conditional distribution given Z factors as a product of conditional distributions. The random variables X and Y are said to be conditionally independent given Z. (Here $A \subset \mathcal{X}$ is an event in the image space of X, B is an event in the image of Y, so, by definition, the inverse images $X^{-1}A$ and $Y^{-1}B$ are events in the original probability space.) Conditional independence is denoted by $X \perp\!\!\!\perp Y \mid Z$ or $Y \perp\!\!\!\perp X \mid Z$.

Density factorization: Suppose that the joint density at (x, y, z) can be factored as

$$p(x, y, z) = q_1(x, z)\, q_2(y, z)\, q_3(z).$$

Then, for any z such that $q_3(z) > 0$, the conditional density of (X, Y) given $Z = z$ is proportional to the product

$$p(x, y \mid Z = z) \propto q_1(x, z)\, q_2(y, z),$$

implying that X, Y are conditionally independent given $Z = z$. Although it is invariably taken for granted, it is worth stating explicitly that p must be the joint density with respect to some product measure on the product space. Otherwise, conditional independence does not follow.

19.1.5 Sufficiency and equality of conditional distributions

The literature on missing values is replete with statements of the general type

$$P(X \in A \mid Y) = P(X \in A \mid Z) \tag{19.3}$$

involving equality of conditional distributions given two statistics. The conditional distribution given Y associates with each point y in the image of Y, a probability distribution $Q(\cdot\,; y)$ on $\mathcal{X}$, while the conditional distribution given Z associates with each point z in the image of Z a second distribution $Q'(\cdot; z)$ on $\mathcal{X}$. The preceding statement says that these distributions are equal, i.e., $Q(A; y) = Q'(A; z)$ for all pairs (y, z) in the image of (Y, Z), and all events $A \in \sigma(X)$ generated by X.

The conditional distributions are equal if $X \perp\!\!\!\perp Y$ and $X \perp\!\!\!\perp Z$; the stronger condition $X \perp\!\!\!\perp (Y, Z)$ is not necessary. Apart from this trivial case, the conditional distributions may be equal if $Z = T(Y)$ is a function of Y or vice-versa, in which case the value $z = T(y)$ suffices to compute the conditional probability $P(X \in A \mid Y = y)$.

Suppose now that $Z = T(Y)$, i.e., the σ-field generated by Z is a subfield of that generated by Y. Suppose also that (19.3) holds for each event $A \subset \mathcal{X}$. Then,

$$\begin{aligned} \mathrm{pr}(X \in A,\, Y \in B \mid Z) &= \mathrm{pr}(X \in A \mid Y, Z)\ \mathrm{pr}(Y \in B \mid Z) \\ &= \mathrm{pr}(X \in A \mid Y)\, \mathrm{pr}(Y \in B \mid Z) \\ &= \mathrm{pr}(X \in A \mid Z)\, \mathrm{pr}(Y \in B \mid Z), \end{aligned}$$

implying $X \perp\!\!\!\perp Y \mid Z$. The second line follows from the fact that Z is a function of Y, i.e., $\sigma(Y, Z) = \sigma(Y)$, and the third line from the hypothesis (19.3). This conclusion is intuitively obvious, and is closely related to the notion of statistical sufficiency for parameter estimation. It is important that Z be a function of Y alone, not a function of (X, Y).

19.1.6 Conditional independence and mistaken identities

Let $R \subset NJ$ be a random subset whose conditional distribution satisfies (1). It follows from the definition that $Y = Y_{\mathrm{obs}} + Y_{\mathrm{mis}}$ is the sum of two complementary random matrices, so the ordered pair $(Y_{\mathrm{obs}}, Y_{\mathrm{mis}})$ determines Y. But, with probability one if Y is continuously distributed,

$$R = \{(i, j) \in NJ : Y_{\mathrm{obs}}(i, j) \neq 0\}$$

is the indicator function for $Y_{\mathrm{obs}} \neq 0$, so the ordered pair determines R. In other words, the ordered pair $(Y_{\mathrm{obs}}, Y_{\mathrm{mis}})$ is equivalent to the ordered pair (R, Y) in

the sense that they have the same information content: as random variables, they generate the same σ-field of events. Although it is commonplace to write $Y = (Y_{\text{obs}}, Y_{\text{mis}})$, rather than $Y = Y_{\text{obs}} + Y_{\text{mis}}$, such expressions demonstrate only how easy it is to compose a logically false mathematical statement.

It is also commonplace to write $Y = (Y_{[r]}, Y_{[\bar{r}]})$ as an ordered pair of random vectors or two lists of random length. This statement is false in both directions. On the one hand, Y does not determine either $Y_{[r]}$ or the number of components; on the other hand, the ordered pair determines $Y_{[r]}$ and $\#R = \#Y_{[r]}$, but Y does not. The ordered pair determines the concatenated list $\text{cat}(Y_{[r]}, Y_{[\bar{r}]})$, which is a straight list of JN real numbers that may be formatted as a matrix, `matrix(c(Y1, Y0), N, J)`. But this matrix is not to be confused with Y.

Abuse of notation is common throughout mathematics, and is essentially unavoidable in the statistics literature. Overloading is the most common benign abuse, in which a symbol such as r does double duty as a subset $r \subset NJ$ and a function $NJ \to \{0, 1\}$. In factorial models, it is well understood that a classification or treatment factor A does quadruple duty as (i) a function $A\colon [n] \to \mathcal{A}$ from the observational units into the factor levels; (ii) the associated list of levels as a spreadsheet column; (iii) the vector subspace $A \subset \mathbb{R}^n$ induced by composition $[n] \xrightarrow{A} \mathcal{A} \xrightarrow{f} \mathbb{R}$; (iv) a set of basis vectors spanning the subspace. Such overloading is difficult to avoid, and seldom causes confusion. But it may lead to confusion in unusual circumstances when a function such as r is also regarded as a matrix, which may subsequently be transposed, so that the matrices r and r' are two representations of the same subset.

The identification of Y with either $(Y_{\text{obs}}, Y_{\text{mis}})$ or $(Y_{[\bar{r}]}, Y_{[r]})$ is not so much an abuse of notation as an error of logic. One consequence is that the MAR condition is sometimes expressed incorrectly in the form

$$\text{pr}(R = r \mid Y) = \text{pr}(R = r \mid (Y_{\text{obs}}, Y_{\text{mis}})) = \text{pr}(R = r \mid Y_{\text{obs}}).$$

The first equality is a consequence of the false equivalence $Y \cong (Y_{\text{obs}}, Y_{\text{mis}})$. The second part asserting that $\text{pr}(R = r \mid Y_{\text{obs}} = y_{\text{obs}}, Y_{\text{mis}} = y_{\text{mis}})$ is equal to $\text{pr}(R = r \mid Y_{\text{obs}} = y_{\text{obs}})$, is a true statement implying that $R \perp\!\!\!\perp (Y_{\text{obs}}, Y_{\text{mis}}) \mid Y_{\text{obs}}$, which is also true. But the statement is true only because both distributions are degenerate at the point $r = I_{y_{\text{obs}} \neq 0}$. In summary, the second equality is an identity holding for all [continuous] distributions by virtue of degeneracy, while the first equality is essentially always false, even if R, Y are independent. Neither part has any connection with MAR.

Sometimes an alternative definition of MAR is offered along the following lines:

$$\text{pr}(R = r \mid Y) = \text{pr}(R = r \mid (Y_{[r]}, Y_{[\bar{r}]})) = \text{pr}(R = r \mid Y_{[r]}). \tag{19.4}$$

The first equality, which is false in most circumstances, is a consequence of the mistaken identity $Y \cong (Y_{[\bar{r}]}, Y_{[r]})$. Since $Y_{[r]}$ determines the number of components recorded $\#R = \#Y_{[r]}$, the middle distribution implies that $\#R$ is degenerate even if the distribution of $\#R$ given Y is not degenerate. The second

part of (19.4), which is not trivial by reason of degeneracy, implies

$$R \perp\!\!\!\perp Y \mid Y_{[r]} \quad \text{or} \quad R \perp\!\!\!\perp Y_{[\bar{r}]} \mid Y_{[r]}, \tag{19.5}$$

a conditional independence statement that follows from (19.3) under the mistaken identity $Y \cong (Y_{[\bar{r}]}, Y_{[r]})$. It may come as a surprise that this conditional independence statement is often put forward as the definition of MAR, even though it is entirely unrelated to MAR, it is not satisfied in most instances of MAR, and it is demonstrably false in Example 4.

Example 5: (continuation of Example 4). Let the marginal distribution of Y be uniform on permutations of $(0, 1, 2)$. Given the restriction $Y[R] = Y_{(1)} = 2$, the random subset $R \subset [3]$ is a singleton whose conditional distribution is independent of α and uniform on singletons in $[3]$. The joint distribution given $Y_{[r]} = 2$ is shown below together with the marginal distribution of R given $Y_{[r]} = 2$.

	$r \subset \{0,1,2\}$							
(y_0, y_1, y_2)	$\emptyset$	0	1	2	01	02	12	012
$(0,1,2)$	0	0	0	1/6	0	0	0	0
$(0,2,1)$	0	0	1/6	0	0	0	0	0
$(1,0,2)$	0	0	0	1/6	0	0	0	0
$(1,2,0)$	0	0	1/6	0	0	0	0	0
$(2,0,1)$	0	1/6	0	0	0	0	0	0
$(2,1,0)$	0	1/6	0	0	0	0	0	0
$p(R = r \mid Y_{[r]} = 2)$	0	1/3	1/3	1/3	0	0	0	0

Although the strong MAR condition is satisfied, the example demonstrates that

$$\text{pr}(R = r \mid Y) \quad \textit{and} \quad \text{pr}(R = r \mid Y_{[r]})$$

are different, contradicting (19.4). *It is evident also that R and Y are not conditionally independent given $Y_{[r]} = 2$, which is a counterexample to* (19.5).

19.1.7 Likelihood and MAR

Let P_θ be a family of distributions on $\mathbb{R}^{NJ} \times \{0,1\}^{NJ}$ indexed by $\theta \in \Theta$ such that the density at (y, r) is

$$q(r, y)\, p(y; \theta).$$

In other words, the marginal density of Y at y is $p(y; \theta)$, and the conditional distributon of R given Y does not depend on the parameter.

Given a fixed subset $r \subset NJ$, each point y may be decomposed as $y = (u, v)$ with $u = r \cdot y$ and $v = \bar{r} \cdot y$ belonging to complementary subspaces. By definition, the marginal density of $r \cdot Y$ at u is

$$p_r(u; \theta) = \int_V p(u + v; \theta)\, dv.$$

If, however, the sampling scheme is such that $R = r$ is the outcome of selective random sampling with distribution (8), the marginal density of $Y_{\text{obs}} = R \cdot Y$ at the point (r, u) is

$$p_r^*(u;\theta) = \int_V q(r, u+v)\, p(u+v;\theta)\, dv. \tag{19.6}$$

Here $r \subset NJ$ is an arbitrary subset, $u = r \cdot y$ is a matrix in $U = \mathbb{R}^r$ whose non-zero components determine r, and the two distributions are such that

$$\int_{\mathbb{R}^r} p_r(u;\theta) = 1$$
$$\sum_{r \subset NJ} \int_{\mathbb{R}^r} p_r^*(u;\theta)\, du = 1.$$

If the MAR condition is satisfied at r, i.e., $q(r, u+v) = q(r, u)$, the integral in (19.6) simplifies to

$$p_r^*(u;\theta) = q(r,u) \int_V p(u+v;\theta)\, dv = q(r,u)\, p_r(u;\theta),$$

which implies equality of ratios

$$\frac{p_r^*(u;\theta)}{p_r^*(u;\theta')} = \frac{p_r(u;\theta)}{p_r(u;\theta')}.$$

Consequently, the likelihood for θ based on the observation Y_{obs} with selectively sampled components is the same as the likelihood for θ computed incorrectly from the marginal distribution (19.6) as if the sampled components were fixed at r in advance.

The MAR condition is not related to independence or conditional independence, and it does not imply that the conditional distribution of Y_{obs} given $R = r$ is the same as the distribution of $r \cdot Y$.

Chapter 20

Presentations and reports

20.1 Coaching tips I

The first eight of these tips are concerned with technical aspects of statistical reports. The last four are concerned with English usage, style and semantics.

1. Length: Reports should be no longer than necessary. A short report that makes the salient points is preferable to a long rambling philosophical essay, even if the longer essay makes the same points somewhere along the way. Above all, have compassion for the reader (and grader).

2. Graphs and plots: A plot either of the raw data or of the residuals is almost always essential at some point in the analysis. Not all plots are helpful or especially interesting. Although you should indicate what plots were made, it is generally not necessary to include in the report a copy of all plots made and all analyses performed. If necessary for examination purposes, extra plots and lengthy analyses can be included in an appendix.

3. Executive summary: All major conclusions should be stated at the beginning in a summary intended for a scientifically literate reader who is not a statistician. Technical terms associated with the context of the problem are unavoidable, but technical statistical terms should so far as possible be avoided. One page is the upper limit. Remember, few readers progress beyond the summary. It is up to the author to state the conclusions early in as persuasive a manner as possible if the reader is to be convinced.

4. Statistical analyses: Following the summary, the report should describe the models fitted, the tests performed, and how these support the conclusions. The relevance of the models to the context under study is important. Technical statistical terms are acceptable here only if they are essential to support the conclusions.

5. Model specification: Statistical models are used by statisticians, computer scientists, engineers, quantitative sociologists, biostatisticans and

epidemiologists, and even by research workers in literature, law and the humanities. The term does not necessarily mean the same thing to all users. Some users think that a statistical model is an equation beginning with $y =$ and ending in $\ldots + \epsilon$. Others are of the opinion that a statistical model is a syntactical expression such as `~A+x+...` containing the symbol $\sim$, or more generally any machine-learning algorithm that is coded in R. A professional statistician knows that a statistical model is a non-empty set whose elements are probability distributions on the observation space, usually $\mathbb{R}^n$.

A *model specification* calls for a statement indicating which distributions are included in the set and which are excluded. Parameter estimation calls for an estimation method, usually maximum likelihood, but very frequently in a modified form such as REML. Maximum likelihood calls for an algorithm, preferably one that is efficient and coded in readily accessible software. Software syntax is important, but an estimation method is not an algorithm, and an algorithm is not a model specification.

A model may be specified *indirectly* by offering a description of how a random draw $Y \sim P_\theta$ may be generated from an arbitrary distribution P_θ in the model. GLMs are usually specified in this manner by a three-step procedure:

$$\eta = X\theta; \quad \pi_i = e^{\eta_i}/(1 + e^{\eta_i}); \quad Y_i \sim \mathrm{Ber}(\pi_i); \quad \text{(indep.)}$$

A great many Gaussian models may be generated by adding several independent Gaussian processes, each associated with a different factor or interaction. For example

$$Y_{it} = \alpha_i + \eta_0(t) + \eta_i(t) + \epsilon_{it}$$

as a sum of four independent zero-mean processes, can be matched up with the direct specification of the covariance function

$$\mathrm{cov}(Y_{it}, Y_{i't'}) = \delta_{ij} + K_0(t, t') + K_1(t, t')\delta_{ij} + \delta_{ij}\delta_{t,t'}$$

provided that α has iid standard normal components, and the other three components are distributed as indicated.

6. Numerical precision: Adequate numerical precision is important, but ordinarily two significant digits are sufficient for standard errors. Parameter estimates should always be given with standard errors, or standard errors of differences in the case of factor levels. It is often sufficient to say that the standard error is 9–15%, or the standard error of pairwise differences is 0.35–0.45, if the range is not excessive. Parameter estimates should be accurate to 10% of a standard error. By convention, p-values are given as a percentage: rarely is there a need for more than two significant digits. The listing of excessively many uninformative digits in estimates and standard errors betrays a lack of statistical sense, and will be penalized.

7. Computer output: While it is necessary to demonstrate that you have mastered the computer system or statistical package, tailoring the computer output to the problem at hand is always necessary if only to demonstrate that you are the computer driver not the computer slave. (i) From the computer-generated analysis-of variance table, list only the parts that are relevant to your analysis. It is your job as statistician and expert to judge what is relevant and what is not. (ii) If the model matrix is non-standard and cannot be generated by a model formula, as for example the additive skew-symmetric formula $E(Y_{ij}) = \alpha_i - \alpha_j$, you need to explain what the structure of the matrix is. (iii) Do not quote a p-value without stating the hypothesis under test and how the value supports the stated conclusion. (iv) Parameters have a physical interpretation: do not pass up the opportunity to remind the reader what the physical interpretation of $\hat{\beta} = 0.684$ is in the context of the problem.

8. Physical units: Physical variables, unlike mathematical variables, always have units such as 'length in mm.,' 'temperature in °K,' 'mm. Hg.,' 'age in months,' or 'depth in fathoms.' If you lose sight of the units your conclusions are liable to be ridiculous. For a published example, see p. 105 in Andrews and Herzberg (1985) where, despite the fact that Adelaide borders on the Australian desert, its annual rainfall is given as 1530 mm., or an astonishing 60 in.

9. Grammar and style: Reports should be logically organized and written in grammatical English. In particular, each sentence should have one, and only one, main verb. Poor logical organization betrays a confused mind, and poor sentence structure indicates a lack of attention to detail.

10. Clarity and word usage: It is good to cultivate an awareness of grammar and word usage. Accurate word usage is important insofar as inaccurate or careless usage sows confusion; good grammar is important insofar as poor grammar betrays faulty logic.

    For example, some native English speakers who are employed as commentators at sports events seem not to understand the difference between the verbs *substitute* and *replace*. These words are also important in mathematics. Viewers are likely to be confused when a talking head recommends at the end of the first quarter that the starting quarterback be substituted! For the correct usage in the active voice, the coach may *substitute* a bench player for a starter or he may *replace* the starter with a substitute from the bench. In the passive voice, an active player may be *replaced*, in which case a bench player is *substituted*. To declare that an active player has been substituted on account of injury is to put the focus on the destination, implying that the coach's job is to ensure that the bench is well-supplied with injured players! Unnintended, perhaps, but possibly accurate.

    In a similar vein with relevance to genetics, the upstream region of a gene may be rich in certain motifs, meaning that those motifs are abundant in

the upstream region. The upstream region is enriched with motifs, but the motifs themselves are neither rich nor enriched anywhere. One could say that coal is abundant in Wyoming and fruit is plentiful in Florida, but coal is not enriched in Wyoming nor is fruit rich in Florida.

use versus usage: The line *What's the use of crying?* from the song *Smile* by Nat King Cole is a rhetorical query about the utility or futility of the act. Similarly, the phrase *cocaine use* refers to the act—its utility, its benefits or its prevalence. By contrast, the title *Modern English Usage* of Fowler's celebrated book refers to the manner in which the language is spoken or written, e.g., imaginatively, in long convoluted sentences, with flair, grammatically, clichéed, and so on. In the same vein, the phrase *cocaine usage* refers to the manner of ingestion. As a statistical factors, `cocaine usage` has levels `snorting`, `smoking`, `injection` and `other`; `cocaine use` has levels `never`, `infrequent`, `occasional` and `regular`.

Verbs for computational activities: Author A writes *I created a proportional-hazards model with covariates...*; author B writes *I ran a p-h model on the data...*; author C writes *I fitted the p-h model...*; author D writes *I trained the p-h model...*; author E writes *The p-h model was trained...*; author F writes *I learned the p-h model....* The proportional-hazards model is a set of probability distributions for survival times. Credit for its creation goes to Cox (1972), not to author A. Generally speaking, one *runs* computer code for an algorithm that is designed to pick the distribution that best fits the data. This activity is called model-fitting—or learning in CS circles. In a sense, the computer or the algorithm learns the best-fitting distribution, possibly using data from a training subsample, and shares that wisdom with the user. Grammatically speaking, if the p-h model is trained on the data, and learns from it, it would be more accurate for author F to write *The data taught the proportional-hazards model...*, or perhaps, *I used the data to teach the proportional-hazards model...*, but the semantic anomaly would then be too plain.

11. Appropriate adjectives: Some computational tasks are easy, while other are hard; some algorithms are efficient for the task while other are inefficient. Likewise for a software implementation of an algorithm. Simulation is easy for some distributions, less so for others. Maximum-likelihood estimation for some models admits a computationally efficient algorithm, not so for other models.

    A task may be easy or it may be hard, but it is neither efficient nor inefficient. A model as a set of probability distributions may be finite or infinite, finite-dimensional or infinite-dimensional; it may be suited to the task or it may not, but it is neither easy nor hard, efficient nor inefficient.

12. Verb tense: Reports should be written in the present tense. If you wish to refer to a past event, by all means use the past tense; likewise for future events. If you switch from from one tense to another mid-paragraph readers will notice, and if a good reason is not apparent, the result will

be confusion. It is best to keep the bulk of your report in the present tense, including references to later sections: *An open-air experiment was conducted during the period 2012–2015; the data from that experiment were analyzed and conclusions are presented in sections 4–5.*

Present tense: *Anthropogenic emissions lead to global climate warming.* Past tense: *Anthropogenic emissions led to global climate warming.* Past perfect tense: *Anthropogenic emissions have led to global climate warming.* Both versions of the past tense, but particularly the first, suggest (probably incorrectly) that anthropogenic emissions no longer have the effect that they had in the past. That incorrect implication may be deliberate if the writer is a White-House hack seeking to justify the U.S. exodus from the Paris Accord, but it is a distraction for the discerning reader. Future tense: *Anthropogenic emissions will lead to global climate warming.* The future tense suggests that emissions did not have this effect in the past.

13. Numbers in text: Small integers 0–10 or 0–12 are usually spelled out when they occur in text. *A 3^{4-1} design has 27 observational units indexed by four factors with three levels each. Zero is one of the dose levels.*

14. Quantities; number, amount, volume: *a great number of tired tourists, diving dolphins, ornery kangaroos, football supporters...; amount of cash in low denominations, amount of food, alcohol, etc.; volume of crude oil, undelivered mail, mining sludge, ripe tomatoes, cheap alcohol...; mass of water, mass of humanity.*

20.2 Coaching tips II

These remarks are the instructor's responses following a Statistics consulting presentation by students on 17 April, 2018. The experiment was done on mice, and the design was factorial with three factors; the observations were cell counts, all large integers.

1. Transformation: In applications of this sort where the observation is a cell count, or any large count of objects, it is much more natural for treatment effects to be multiplicative than additive. Why so? If the mean cell count for controls in the three genotypes are 1000, 2000, 3000, and the treatment effect is -0.69, or a 50% reduction, the cell means for treated mice in the same genotype classes will be 500, 1000, 1500. So the average reduction is 1000. An additive model with an additive reduction of 1000 will have treatment means 0000, 1000, 2000 for the three groups. Usually, this sort of thing—no cells at all in one group—is very implausible; negative counts are even less likely. So the conclusion is that the log scale should be the first option for analysis, the go-to choice, but not necessarily the final choice.

2. Experimental units versus observational units: In this experiment, the observational units are mice, and all responses are measured post-mortem.

However, all mice in one litter have the same genotype and were given the same treatment. This is a classic distinction. It is not possible in this design for two mice in the same litter to be given different treatments. Accordingly, the mice are the observational units, and the litters are the experimental units, i.e., each litter is one experimental unit. It is one of the few universally-agreed rules of experimental design and analysis that you cannot have more degrees of freedom for the estimation of treatment-effect variances than there are experimental units available for analysis (27 in this case). One way to proceed is to reduce each litter to the litter average, and to do the standard factorial decomposition on the litter averages. My preference is to average the counts and then take the log, but you could take logs first and then average. The operations are not commutative.

3. Random effects: The use of litter averages is not ideal because litters vary in size, probably from one to six or thereabouts. A linear analysis weighted by litter size is not correct either—that weighting is too extreme. A better option is to use a random-effects model in which each litter is associated with an independent additive Gaussian variable with constant variance independent of litter size. Since each random effect is associated with the contribution of one experimental unit, the question of significance testing for a zero between-litter variance is not something that arises naturally. There is simply no reason to expect zero additional variance per experimental unit, so the litter effect must be retained whether it is statistically significant or not. Remember that it is the experimental units that govern the degrees of freedom for treatment-effect estimation: the number of observational units is entirely irrelevant, even if infinite.

4. Model selection: In this design, there are three factors $3 \times 2 \times 2$, where genotype is a three-level classification factor. Ordinarily, in the analysis of a factorial design, the main effects of all three factors will be retained in the 'final model', regardless of significance. This is sound scientific practice, and there are many reasons for it. Comparability with other studies of the same phenomenon in similar or different circumstances is paramount. For three factors, regardless of the number of levels for each, there are only 9 factorial subspaces that include all three main effects,

$$A + B + C, \quad A * B + C, \quad \ldots, \quad A * B + B * C, \quad \ldots, \quad A * B * C.$$

Of these, only about five are likely to be seriously contemplated: $A+B+C$ (additivity, no interactions), $A * B + C$, $A + B * C$, $B + A * C$ (one interaction only, but additivity for the other), $A*B*C$ (no additivity anywhere). The lesson: whatever you learned in class about model selection in regression is not relevant here. Subset selection is not such a big issue in most scientific work involving factorial designs, and standard covariate selection algorithms are an outright menace in this setting. However, if you are using a random-effects model, as you ought, you do need to be careful to use a proper likelihood-ratio statistic (NOT REML) for the comparison

of two nested factorial models. This is one of the few instances where a technical measure-theoretic issue impinges on statistical methodology. If you are unsure about the technicality, just ask.

5. Coding of factors: Unless the factor levels are ordered or have additional structure, the fitted model should be independent of the coding. For example a factor with two levels coded "M" and "F" might represent sex—or it might represent parent. In one case the "F" level stands for 'Female' in the other case it stands for 'Father'. It is clearly unacceptable for the fitting or selection procedure to depend on the letter or character string used to represent each level. Each factorial model is a vector subspace; although the labelling of the basis vectors must depend on the coding, the subspace itself is invariant with respect to coding. The coding determines the basis vectors but not the subspace. The factorial models are essentially the only subspaces that have this property, which is most naturally expressed in terms of algebraic representation theory. A model-selection procedure that is code-dependent is a plague to be avoided.

6. Graphs and tables: A graph is helpful for presentation only if it illustrates an important effect clearly. A graph of residuals may be helpful for model checking, and may be mentioned in presentation, but it is seldom included as part of the report. In most factorial designs, whether balanced or otherwise, one-way and two-way tables of averages are often useful as a partial summary of conclusions.

7. Higher-order interactions: How do you present the conclusions comprehensibly if high-order interaction is present? If the additive model is a satisfactory fit, you can report estimates of main effect contrasts in the usual way—pooling higher-order interaction sums of squares to obtain an estimate of variance. There is little need to encourage bad scientific behaviour by giving undue emphasis on p-values, but you should report the degrees of freedom of the variance estimate, particularly if it is small. If, as appears to be the case here, there is a high-order interaction, it is best to partition the units into sub-classes by genotype, and to report the treatment effects separately, but in parallel, for each genotype. Since there are four treatment combinations for each genotype, you can report the three contrasts with some reference level. Show these numbers in a 3×4 table, one row per genotype with standard errors but absolutely no p-values.

8. Rules-of-thumb for summary statistical tables:

   (a) Report effect estimates and standard errors only. Ratios are OK. No asterisks please!

   (b) Always label the effects in an informative way so that the reference level for each factor is clear.

   (c) Always report the reference level of each factor with zero as the estimate.

(d) Always report the variance-component estimates.

(e) Estimates and regression coefficients: four significant decimal digits maximum.

(f) Standard errors: three significant digits maximum.

(g) F-ratios: two digits maximum.

(h) p-values: best avoided completely, but two digits maximum as a percent if absolutely needed.

(i) If you must report a p-value, be sure to state the null hypothesis being tested.

9. Baseline: Baseline refers to a point in time just prior to randomization and treatment assignment. Notionally, the probability model for the outcomes is registered at baseline, so all information needed to determine outcome probabilities (including the randomization outcome) must be revealed at that time. Any variable recorded at or pre-baseline is called a baseline variable. A block factor is an example of a baseline variable. Age at recruitment in a clinical trial is a baseline variable.

10. Covariate: A covariate is a function on the units that is known in advance and recorded pre-baseline for the in-sample units. Typical covariates in a clinical trial include age, sex, and medical history. In the case of a vital response variable such as blood pressure, cholesterol or blood serum level, the baseline value is usually recorded as part of the recruitment interview. As such, the initial response and other baseline variables may be used to determine eligibility for inclusion in the study, particularly if the study focuses on high-risk patients. Thus, the baseline response value is, or may be treated as, a covariate, which is regarded as fixed in the probability model.

11. Treatment assignment: Treatment assignment is determined by randomization at baseline. Usually the treatment is not assigned independently to experimental units, but is subject to design conditions such as balance and equi-replication within blocks. In general, the treatment assignment probabilities, most obviously the joint probabilities for two or more observational units, may depend on block sizes, covariates and other baseline variables. The treatment assignment vector is technically a random variable, not a covariate or baseline variable.

12. Response: Many studies have multiple responses per observational unit, for example birth weight and gestational period in a study of the effect of certain interventions in a medical setting. For such a setting, each observational/experimental unit i is a mother/baby, and the response $i \mapsto (t_i, w_i)$ is bivariate. Treatment (e.g., folic acid supplement) may have an effect on the baby's weight at birth; it may also have an effect on the probability of a premature birth. So there are at least two treatment effects to be considered. The effect of treatment on birth weight is ordinarily

defined as the difference of average weights or log-weights; the effect of treatment on gestational period is defined likewise. But, in general, the full story is the effect of treatment on the joint distribution: gestational period and birth weight are strongly correlated. To estimate the effect of treatment on birth weight, it is not legitimate to include gestational period as a 'covariate' in the one-dimensional model for birth weight.

Chapter 21

Question and answer

21.1 Scientific investigations

21.1.1 Observational unit

Q1. Who made the world?
A1. God made the world.

Q2. Was it an experiment?
A2. We have every reason to believe so. It is the best explanation we have for the current state of pestilence and political chaos.

Q3. Where did He start?
A3. If it was an experiment, He started at the baseline.

Q4. What is the baseline?
A4. A point in time prior to all experience—the most recent point in time prior to the revelation of protocols and the implementation of randomization.

Q5. Does anything exist before the baseline?
A5. Yes, every scientific investigation has a protocol—written or unwritten.

Q6. What is the protocol?
A6. The protocol is a declaration of purpose, timeline, strategy and tactics.

Q7. Tell me about the individual parts.
A7. The purpose is the phenomenon to be investigated, the response and the target population. Strategy tactics and timeline refer to the study design, the sample, and the measurement process.

Q8. What is the target population?
A8. The target population is the set of observational units.

Q9. Does the population exist pre-baseline.
A9. The observational units are declared pre-baseline, so they must exist.

Q10. What does it mean for something to exist?

A10. Existence means occurrence as a feature or component in the mathematics as declared in the protocol.

Q11. Does mathematical existence have any connection with reality?

A11. Assuming that we can agree on the meaning of reality, everything of interest that exists in reality, and every relevant event that could possibly occur in reality, must have a counterpart in the mathematics. Reality in that counterpart sense is a subset of mathematics.

Q12. Isn't that asking a lot from mathematics?

A12. Yes and no. The phrase 'everything of interest' implies compartmentalization or restriction to objects and events that are considered relevant to the investigation.

Q13. What objects and events are relevant to God's experiment?

A13. Only God can answer that. There are claims that His protocols has been revealed, but I haven't read them.

Q14. Is every observational unit a physical object?

A14. Every observational unit is a mathematical object, which may or may not correspond to a physical object.

Q15. Tell me more about that.

A15. The NW3 weather station near Hampstead is a physical object of sorts, but the observational units in a meteorological series are site-time pairs. A data analyst is usually content to represent the sample by certain floating-point pairs of numbers in an electronic computer. But the mathematical system contains uncountably many units that cannot be represented in an electronic computer.

Q16. How many observational units are there?

A16. Usually the number in the population is infinite. But the sample is always finite.

Q17. So the sample is a finite random subset of the population?

A17. Finite, yes.

Q18. And random?

A18. The status of the sample as a fixed subset or a random subset is part of what is revealed implicitly or explicitly by the protocol.

Q19. Can you give me an example.

A19. The protocol identifies the baseline, the population of interest, the sample or sampling scheme, and the response variable. Suppose the baseline is Dec 31, 1899. The protol declaration *daily noon temperature at Kew, Greenwich and Hampstead, Jan 1, 1900 to Dec 31, 2020* identifies the response and the sample points as a fixed finite set.

Q20. And the population?

A20. Usually it is not necessary to be persnickety about the population, so any larger space-time domain suffices. In the absence of a compelling argument to the contrary, the entire space-time product set serves as the population.

Q21. Didn't you say that the population must exist at baseline? Does Jan 8, 2022 exist at baseline?

A21. Yes, I did. And yes, the ordered pair (Kew, Jan 8, 2022) exists today just as it did in AD 1899. But I did not say that every unit must be accessible or observable immediately after baseline.

Q22. The space-time product set is uncountable in both dimensions. Isn't that excessive and unnecessarily extensive for statistical work?

A22. Maybe so, but imperialism is inscribed in the DNA of mathematics. Besides, if you restrict the population, you forego the opportunity to make inferences about the disenfranchised parts.

21.1.2 Clinical trials

Q1. What is the role of the protocol for a clinical trial?

A1. Patient eligibility is one crucial protocol declaration.

Q2. And what are the implications of eligibility criteria?

A2. The population consists of all eligible patients—patients who were eligible yesterday, patients who are eligible today, and most certainly those who will be eligible tomorrow. The recruitment scheme for patients is also part of the protocol. Of necessity, the sample is a subset of patients who are eligible today. Usually the sample is also restricted geographically.

Q3. What's the point of including dead folk?

A3. Why not? They don't charge for service or rent.

Q4. Why include patients who are not yet born?

A4. If you are interested only in the current population, so be it. That's OK for short-term planners and short-sighted politicians. If a goal is to say something about the effect of a COVID vaccine or global warming, you may wish to cast the net liberally by including future generations.

Q5. What hope is there of saying anything useful about the effect of a vaccine on future generations?

A5. The purpose of casting a wide net is not to say something *useful* about future generations, but to be in a position to say anything at all. Similarly for patients who have the misfortune to be foreigners or aliens.

Q6. Patients in a clinical trial are usually recruited sequentially as they present themselves at the medical centre. Is a sequentially-recruited sample fixed or random?

A6. That appears to be a philosophically complicated question, but, ... (reaching

into his pocket), here is a sample of six pennies. Is it a fixed subset of all pennies or a random subset?

Q7. It's obviously a random subset.

A7. And if I say that it is a fixed subset, can anyone prove otherwise?

Q8. Perhaps not, but I would not believe you.

A8. And you might well be right to be skeptical. But the question is meaningless without mathematical context. It can be answered only as a mathematical question.

Q9. So how do you formulate sequential recruitment mathematically?

A9. First, you must retain in the mathematics anything that is essential for the context. All else can and must be discarded. One obvious difficulty is that there is no master-list of eligible patients—not even a comprehensive list of patients who are eligible today. So either you pretend that there is a master list, or you figure out a way to cope without it.

Q10. How does your mathematics cope without a master-list?

A10. One solution is to record eligible patients as a point process by date of presentation—with follow-up to monitor disease progress. The sample is the subset that presents at a given medical centre in a given time window.

Q11. So, is such a sample fixed or random?

A11. Well, the window is fixed, but the sample as a set of time points is random and locally finite.

Q12. And what about the patients? Can they be a random sample?

A12. That's complicated because there is no master-list that can be identified as the set of eligible patients. There is only a window and a set of presentation times, which we use to label patients in the sample.

Q13. So, what is it? Fixed or random?

A13. It is neither a fixed subset nor a random subset because there is no concept in the mathematics of a population of patients. Disease occurrence is a process and recruitment is a process.

Q14. That doesn't seem to fit in with the general framework.

A14. Maybe so. The recruitment process is a marked point process that is observed in a fixed temporal window. Each patient has his own baseline, which is the time of recruitment. The marks, which include age and sex plus current and future health, are random variables. However, any marks that are revealed at recruitment are pre-baseline values. That includes the sample size or window length. In my opinion, the point-process sample is best treated as a fixed subset of an infinite population.

Q15. Only patients who have access to a qualified physician are included in your description of the population. What about those who are eligible but do not have access, either for reasons of geography or economics?

A15. Whether the sample is fixed or random, it can usually be guaranteed that there are units in the population that have zero probability of being

included in the sample. If the outcome (the effect of COVID vaccine) were very different depending on geography or economics, we would certainly want to know. Practically speaking, it is best to recruit broadly and to record adequate baseline information.

Q16. I want to revisit a remark that you made earlier about mathematicians being imperialistic in outlook. Could you elaborate on that?

A16. Far be it from me to say anything derogatory about mathematicians or statisticians, either individually or as a group. I did say that mathematics was imperialistic in outlook.

Q17. That sounds like criticism to me. What do you mean by it?

A17. I may have said it in a cynical tone of voice, but I meant it in a positive and approving way. Mathematics has always been imperialistic, and it should be imperialistic. When Pythagoras discovered his theorem, he declared it to be a universal truth holding not only for Greek triangles but also for Egyptian and Assyrian triangles as well. That sort of imperialism is good. Maybe catholic ($\kappa\alpha\theta o\lambda\iota\kappa o\varsigma$) or universal would be a better word.

Q18. It is hard to see the relevance of catholicism to applied statistics.

A18. On the contrary. Random samples and finite-population models for clinical trials are a case in point. There is nothing mathematically wrong with a finite population if that is your universe. But the philosophy is all bad. It is democratic, short-sighted and inward-looking.

Q19. How so?

A19. To arrange matters so that every individual in the population has strictly positive inclusion probability is an undeniable democratic idea. But it comes at enormous cost to subsequent generations who are not accessible today, and must be excluded. It is also contrary to the spirit of scientific catholicism, which recoils at restrictions. I would go further to say that any medical statistician who restricts the population to the current generation is mathematically derelict in his or her duty of care to subsequent generations.

Q20. But surely the finiteness assumption can't make much difference to procedures and conclusions.

A20. It absolutely makes a difference to conclusions because, if you don't admit that the next generation exists in your population, you forego the opportunity to say anything about the effect of treatment tomorrow.

Q21. What reason is there to say that the effect of treatment today must be the same as the effect tomorrow?

A21. I do not claim that the effect is constant over generations. But I do insist on the right to make that comparison. As do you, apparently. If you were to conclude that today's data are irrelevant for tomorrow's patients, that would be fine by me. But if you don't admit the existence of tomorrow, you can't even say that.

Q22. But medical recommendations are seldom explicitly time-constrained.

A22. True enough. In that case your actions imply that today's data are rele-

vant, and that the effect is constant or neatly so. Horvitz-Thompson sampling theory with positive inclusion probabilities implies the opposite.

Q23. So, how does philosophy affect procedures?

A23. Mathematically speaking, you can't have it both ways. If you want to say something about the effect of treatment in the future, you must have future generations in the population. If you insist on a finite population with a random sample and strictly positive sample-inclusion probabilities, future generations must be excluded, and you forego the opportunity to address the critical question. I'm a statistical catholic, so you know where I stand.

21.1.3 Agricultural field trials

Q1. Can you say a little about agricultural field trials?

A1. By comparison with clinical trials, field trials are very simple.

Q2. How so?

A2. Each observational unit is a plot in the field. The protocol specifies the varieties or cultivars to be tested by growing on the sample, which consists of 36 plots situated at the western end of Hoos field. That's all there is to it. No recruitment or random sampling of plots. Only random assignment of varieties to plots in the sample.

Q3. My impression was that random samples were the norm in all statistical work. Wouldn't it be better to use a random sample of plots from several fields?

A3. Try that on the farm manager! But you might make a case for a more extensive design replicated in several distant blocks differing in soil composition or weather pattern. A variety that performs well at Rothamsted might fare poorly in Rotherham or Rothesay.

Q4. I have an image of each sample unit as a rectangular plot, all sample plots being neatly arranged by rows and columns separated by access paths. What does the population of 'all plots' look like.

A4. The population is a family of planar subsets.

Q5. Are they all the same size and shape?

A5. Not at all. It is not necessary to include all planar subsets, but a mathematician instinctively aims to include all Borel subsets. That's a big set, big enough for most purposes, but maybe not big enough for all purposes. The units in a long-term field experiment also have a temporal component.

Q6. That seems far too big. Besides, plots cannot overlap.

A6. A catholic statistician must always think big. If the response is yield, there is no concern about overlap: yield is an additive set function.

Q7. But you cannot have different treatments on overlapping plots.

A7. That's a good reason for picking a sample of non-overlapping plots.

Q8. If your sample of plots is a non-random subset, where does the probability come from?

A8. Probability comes from the mathematical framework that is implicit or explicit in the protocol. Exchangeability gives rise to probabilities. Randomization also gives rise to probabilities.

Q9. What is the role of randomization analysis?

A9. Randomization is usually associated with the uniform distribution on a finite group acting on the sample units. Re-randomization enables you to generate new 'pseudo-samples' having the same distribution as the original. For any non-invariant statistic, you can compute its randomization distribution. This is a useful way to determine where the observed treatment effect occurs in the spectrum of treatments effects to be expected under randomization.

Q10. So the set of units in the randomization analysis is the finite sample of plots?

A10. Certainly the sample is finite.

Q11. Isn't the randomization population the same as the sample?

A11. In a purely arithmetical sense, yes!

Q12. Is there any other sense?

A12. There must always be a wider statistical sense.

Q13. To what end?

A13. Presumably you want to say something about the likely effect of treatment on other plots of a similar type in the population.

Q14. Couldn't you just take the finite-population estimate, patch it together with the randomization distribution or bootstrap distribution, and apply that to other plots.

A14. If you had no principles or concerns about mathematical honesty, you could do whatever you liked.

Q15. Isn't that what every statistician does? Are we all dishonest?

A15. It is true that many statisticians do exactly that—and very often it is the right thing to do.

Q16. So what's the problem?

A16. The problem is one of honesty in mathematics. If you refuse to acknowledge extra-sample plots, the statement about treatment effect is meaningless. If you acknowledge their existence you have to establish a connection between yields on the in-sample plots and yields on extra-sample plots. That step requires an assumption such as exchangeability.

Q17. In that case, what is the role of randomization analysis?

A17. Randomization analyses and bootstrap analyses are logically sound and useful statistical tools. On its own—restricted to the finite sample of plots—randomization is a basis for arithmetic and distribution-theory. It is not otherwise a basis for statistical inference.

21.1.4 Covariates

Q1. Apart from the observational units, what else exists before the baseline?
A1. Covariates are recorded pre-baseline.

Q2. What is a covariate?
A2. A variable recorded pre-baseline.

Q3. What is a variable?
A3. A variable is a function on the observational units.

Q4. What types of covariate are there?
A4. Qualitative variables or classification factors, and quantitative variables such as age or calendar date or spatial position.

Q5. Are there any other types of covariate?
A5. Yes, relationships can also be recorded at baseline.

Q6. What is a relationship?
A6. A relationship is a function on pairs of observational units.

Q7. Can you give examples.
A7. A block factor is an equivalence relation; there are also genetic relationships, familial relationships, temporal relationships and metric relationships.

Q8. What is a metric relationship?
A8. A metric is a symmetric non-negative function on pairs that satisfies the triangle inequality.

Q9. Are any covariates recorded post-baseline?
A9. No. Every post-baseline variable is a random outcome subject to the rules of probability.

Q10. What happens at baseline?
A10. The protocol is announced, units are assembled, treatment is assigned by randomization, and nature or Tyche takes over.

Q11. Who is Tyche?
A11. Tyche is the Greek goddess of chance—Fortuna to the Romans.

Q12. Is treatment a covariate?
A12. No, it is not.

Q13. Why not?
A13. Treatment is the outcome of randomization as specified by protocol.

Q14. Is treatment assigned independently to units in the sample?
A14. Not necessarily. A balanced design has non-independent assignments.

Q15. Is the treatment assignment distribution the same for every unit?
A15. Not necessarily. In principle, the treatment assignment probability may vary from one covariate sub-group to another as specified by protocol. But this practice is not common and is not encouraged.

Q16. What is the purpose of randomized treatment assignment?
A16. Randomization is a panacea. It has many purposes.

Q17. Tell me one specific purpose.
A17. Concealment of treatment assignment promotes integrity in human trials.

Q18. Can you elaborate?
A18. Where human subjects are involved, the integrity of the experiment is at risk if the treatment assignment is revealed prematurely, either to the patient or to the physician. Concealment helps to limit the possibilities for subverting the design.

Q19. Any other purposes?
A19. To see if God is paying attention.

Q20. What has God got to do with it?
A20. Concealment means that treatment assignment is known only to the controlling statistician, who must pay attention to events as they unfold.

Q21. Any other purpose?
A21. To help convince skeptics by levelling the playing field for treatment comparisons.

Q22. Tell me about the role of exchangeability?
A22. Exchangeability is the fundamental axiom of statistical modelling.

Q23. What does exchangeability imply?
A23. It implies that two units having the same covariate value must have the same response distribution. Implicitly or explicitly, that's usually part of the protocol.

Q24. Is exchangeability a mathematical theorem?
A24. No, it is an axiom of applied statistics. You can think of it as a bill of rights or a guarantee of equality under the law. If two units are to have different response distributions, there needs to be a demonstrable reason for that difference.

Q25. What is the purpose of a covariate in a randomized study?
A25. There are three inter-related purposes.

(i) to accommodate sub-group effects (sex, age,...);
(ii) to improve precision of the treatment estimate;
(iii) to check for interaction.

Q1. What is the baseline for a matched-pairs design?
A1. The time when the units are assembled or declared, just pre-randomization.

Q2. What covariates are available in the matched-pairs design?
A2. In the simplest setting, only the block factor indicating the pairs.

Q3. What was the baseline for the hypertension study?
A3. Jan 1 when the patients were first measured to determine eligibility.

Q4. Was that pre-randomization?
A4. Yes. Eligibility precedes randomization.

Q5. What covariates are available in the hypertension study?
A5. The block factor and the initial values.

Q6. Which is the correct method of analysis?
A6. ANCOVA with adjustment for initial values.

21.1.5 The effect of treatment

Q1. I've read that each patient in a randomized trial has two potential outcomes or counterfactual responses, only one of which can be recorded. Is that correct?
A1. Yes, in the sense that only one response is observed per patient. Otherwise, no.

Q2. Where do counterfactuals fit in?
A2. Anything relevant that exists in reality must have a counterpart in the mathematics; some things that do not exist in reality may occur in the mathematics. Existence in that sense is a question of mathematical style and taste. Counterfactuals are employed by many authors, and for those authors they exist. But they are not needed, and they do not exist in these notes.

Q3. If it is simply a matter of mathematical style, what is there to argue about?
A3. If the conclusions are the same either way, we can only argue about style. That's plenty.

Q4. You've argued that treatment is a random variable and not a covariate. What are the implications of that distinction?
A4. To be clear, it is the vector of treatment assignments that is random. Each patient or observational unit has a joint treatment-response distribution, which implies a conditional distribution given the treatment.

Q5. What is a conditional distribution?
A5. A conditional distribution associates with each treatment level a probability distribution on the state space. In effect, each patient has one joint and two conditional distributions, one for control and one for active treatment. Two distributions, one response.

Q6. How does this viewpoint fit in with counterfactuals?
A6. The most extreme counterfactual framework associates with each patient two real numbers, a C-value and a T-value so that the conditional distribution is degenerate at one or other point.

Q7. So the two points of view coincide if all conditional distributions are degenerate?
A7. Up to a point. Exchangeability is a fundamental assumption, and degeneracy is hard to reconcile with exchangeability.

Q8. What is exchangeability?

A8. Exchangeability in these notes means that two units having the same baseline covariate value, also have the same response distribution.

Q9. In what way does exchangeability contradict degeneracy?

A9. Pick two units having the same covariate value, and suppose they get the same treatment. Degeneracy plus exchangeability implies that they must have exactly the same response, which will certainly not happen for all such pairs.

Q10. So you have to abandon one or the other?

A10. You have to abandon something. I prefer to keep exchangeability and abandon degeneracy.

Q11. But that's not the only escape route, is it?

A11. Regrettably, no. You could compromise by considering a less extreme model for counterfactuals with non-degenerate distributions. Or, if compromise is not part of your vocabulary, you could take the anarchist route.

Q12. I like the sound of that. What is the anarchist route?

A12. The anarchist argues that covariate values are so numerous that no pairs exist having the same value. The exchangeability argument is then demolished by fiat.

Q13. What options does that leave?

A13. Very few, and none that are palatable, which explains the pejorative label.

Q14. How do you address the anarchist?

A14. It is hard to engage with the anarchist, so I address only the argument. If you look at it as a mathematical statement, you can include in the population infinitely many duplicate units for each covariate value, so the statement is obviously false.

Q15. Does that take care of the argument.

A15. Not quite, I'm afraid. Any continuously-varying covariate such as age or height has non-countably many values, but no anarchist would consider those to be counter-examples capable of undermining exchangeability. The more damaging examples occur in so-called personalized medical studies, where each patient starts off with an entire DNA sequence as a covariate.

Q16. What makes the second more damaging than the first?

A16. Topology, I suppose. In the first case, you might be persuaded that the response distribution varies continuously with age or height. But it hard to make a similar argument for continuity on the space of DNA sequences.

Q17. Can anything useful be done in that setting.

A17. That's complicated. It all depends on what you mean by 'useful' and 'that setting'.

Q18. What other means is there of escape from anarchy?

A18. The option of ignoring the covariate entirely is always available.

Q19. That sounds defeatist and contrary to catholic dogma, isn't it.

A19. It goes against the grain, but the avoidance of chaos is also a virtue. Remember that every unit in the population has a unique identifier, but you're not expected to include that as a covariate. Decisions must always be made about the relevance of various pieces of information, the vast majority of which is discarded. So if you are given information so voluminous that you have no idea how to use, you can simply decline it. You might subsequently learn how to use it. That's called progress.

Q20. Let's move back to treatment. What do you mean by the effect of treatment?

A20. The effect of treatment is to modify the response distribution by group action. The effect is to change the control distribution for each patient to the corresponding active distribution.

Q21. What do you mean by the treatment effect?

A21. The treatment effect is a specific group element, a parameter if you like.

Q22. Is the effect of treatment the same for everyone?

A22. Yes, in the sense that it is the same group action on distributions. But no, the particular group element need not be the same for everyone. It may vary from one covariate subset to another; in (5.2), the treatment effect is linear in time.

Q23. If the treatment effect is not the same for everyone, should we report the average treatment effect?

A23. The question presumes that the group is closed under averages, which is not necessarily the case. If the treatment effect for males is not the same as that for females, it would be better to report one effect for each sex. Same for population subsets determined by any covariate.