STAT 222 Lecture 8-10
Chapter 5 Checking Model Assumptions

Yibi Huang

Assumptions to Check

In the model for multi-sample data

$$y_{ij} = \mu_i + \varepsilon_{ij}$$

we need to check the following.

- *form of the model*
- *outliers*: are there any unusual observations (outliers)?
- *independence*: do the errors ε_{ij} appear to be independent?
- *constant variance*: do the errors ε_{ij} have similar variances for each treatment?
- *normality*: do the errors ε_{ij} follow a normal distribution?

Many assumptions above are related to errors ε_{ij}, most of the model diagnostic methods are based on the *residuals*

$$\text{residual } e_{ij} = y_{ij} - \widehat{y}_{ij} = y_{ij} - \overline{y}_{i\bullet}.$$

Robustness of Validity

In the real world, data never exactly conform to these assumptions. Thankfully, the analysis in Ch3 & 4 work reasonably well if the reality doesn't deviate from the assumptions too much. This is called robustness of validity.

> *"All models are wrong, but some are useful." — George P.E. Box*

Residuals

Standardized Residuals = Internally Studentized Residuals

- The error ε_{ij} in the model $y_{ij} = \mu_i + \varepsilon_{ij}$ has mean 0 and SD σ
- The SD of $e_{ij} = y_{ij} - \overline{y}_{i\bullet}$ is actually $\sigma\sqrt{1 - \frac{1}{n_i}}$., not σ.

Standardized Residuals = Internally Studentized Residuals

- The error ε_{ij} in the model $y_{ij} = \mu_i + \varepsilon_{ij}$ has mean 0 and SD σ
- The SD of $e_{ij} = y_{ij} - \overline{y}_{i\bullet}$ is actually $\sigma\sqrt{1 - \frac{1}{n_i}}$., not σ.
- We hence standardize the ith raw residuals as follows, called the *standardized residual* or the *internally Studentized residual*

$$\textit{standardized residual } s_{ij} = \frac{e_{ij}}{\sqrt{\text{MSE}(1 - \frac{1}{n_i})}}.$$

- If the errors ε_{ij} are normal, s_{ij} is approximately $N(0, 1)$.
- Observations with $|s_{ij}| > 3$ are potential **outliers**.

Studentized Residuals = Externally Studentized Residuals

For each externally studentized residuals, we use the MSE obtained from the model that uses all the data *EXCEPT that observation*, denoted as $\text{MSE}_{(ij)}$.

- Subscript "(ij)" means "all but the jth observation in the ith group".

Studentized residuals or *externally studentized residuals* are defined as:

$$t_{ij} = \frac{e_{ij}}{\sqrt{\text{MSE}_{(ij)}(1 - \frac{1}{n_i})}}$$

- e_{ij} is still computed using all the data but $\text{MSE}_{(ij)}$ is computed excluding the jth observation in the ith group".
- t_{ij} has a t-distribution with $N - g - 1$ d.f. but s_{ij} does not have a t-distribution.

Which Residuals Should We Use?

Internally and externally studentized residuals are related as follows

$$t_{ij} = s_{ij}\sqrt{\frac{N-g-1}{N-g-1-s_{ij}^2}}$$

If an observation is not an outlier, $t_{ij} \approx s_{ij}$.
It makes little difference which one we used.

For potential outliers, it's **better using the externally studentized residuals**.

Check for Outliers

Example: Hodgkin's Disease

Hodgkin's disease is a type of lymphoma, a cancer originating from white blood cells called lymphocytes. The data file

```
hodgkins = read.table(
  "http://www.stat.uchicago.edu/~yibi/s222/Hodgkins.txt", header=T)
```

contains plasma bradykininogen levels (in μg of bradykininogen per ml of plasma) in 3 types of subjects

- normal,
- in patients with active Hodgkin's disease, and
- in patients with inactive Hodgkin's disease.

Example: Hodgkin's Disease

Hodgkin's disease is a type of lymphoma, a cancer originating from white blood cells called lymphocytes. The data file

```
hodgkins = read.table(
  "http://www.stat.uchicago.edu/~yibi/s222/Hodgkins.txt", header=T)
```

contains plasma bradykininogen levels (in μg of bradykininogen per ml of plasma) in 3 types of subjects

- normal,
- in patients with active Hodgkin's disease, and
- in patients with inactive Hodgkin's disease.

Response: The globulin bradykininogen is the precursor substance for bradykinin, which is thought to be a chemical mediator of inflammation.

Example: Hodgkin's Disease

Hodgkin's disease is a type of lymphoma, a cancer originating from white blood cells called lymphocytes. The data file

```
hodgkins = read.table(
  "http://www.stat.uchicago.edu/~yibi/s222/Hodgkins.txt", header=T)
```

contains plasma bradykininogen levels (in μg of bradykininogen per ml of plasma) in 3 types of subjects

- normal,
- in patients with active Hodgkin's disease, and
- in patients with inactive Hodgkin's disease.

Response: The globulin bradykininogen is the precursor substance for bradykinin, which is thought to be a chemical mediator of inflammation.

- Is this an experiment?

Example: Hodgkin's Disease

Hodgkin's disease is a type of lymphoma, a cancer originating from white blood cells called lymphocytes. The data file

```
hodgkins = read.table(
  "http://www.stat.uchicago.edu/~yibi/s222/Hodgkins.txt", header=T)
```

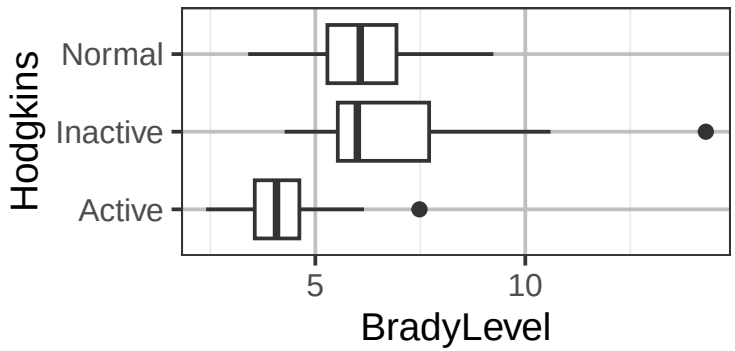
contains plasma bradykininogen levels (in μg of bradykininogen per ml of plasma) in 3 types of subjects

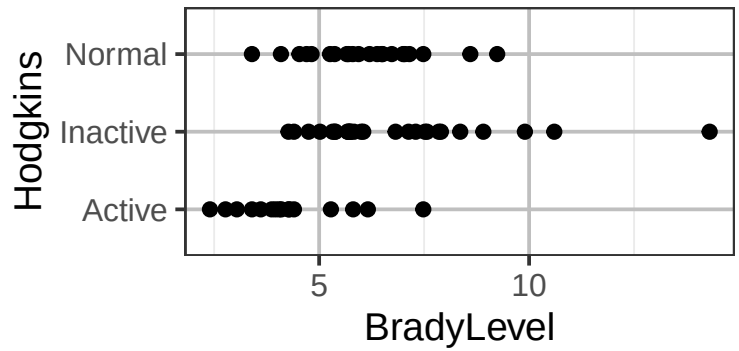
- normal,
- in patients with active Hodgkin's disease, and
- in patients with inactive Hodgkin's disease.

Response: The globulin bradykininogen is the precursor substance for bradykinin, which is thought to be a chemical mediator of inflammation.

- Is this an experiment?
- We can use ANOVA to compare means of several samples in an observational study.

```
library(ggplot2)
ggplot(hodgkins, aes(y=Hodgkins, x=BradyLevel)) + geom_boxplot()
ggplot(hodgkins, aes(y=Hodgkins, x=BradyLevel)) + geom_point()
```





The distribution within each group looks right skewed. Let's fit the ANOVA model anyway and take a look at the residuals.

```
brady1 = lm(BradyLevel ~ Hodgkins, data=hodgkins)
anova(brady1)
Analysis of Variance Table

Response: BradyLevel
           Df Sum Sq Mean Sq F value   Pr(>F)
Hodgkins    2  65.89   32.95   10.67 0.000104
Residuals 62 191.45    3.09
```

Residuals

Residual $e_{ij} = y_{ij} - \overline{y}_{i\bullet}$

```
round(brady1$res,2)
    1     2     3     4     5     6     7     8     9    10
-0.72 -0.30 -1.39 -0.39 -2.70  2.50  1.39  1.06  0.40 -2.01
   11    12    13    14    15    16    17    18    19    20
-0.15  0.28  3.15 -0.44 -1.56  0.41  0.90  0.10  0.94 -1.27
   21    22    23    24    25    26    27    28    29    30
 0.64 -0.84 -0.35 -1.27  0.97 -0.91 -0.21 -0.70  1.85  3.17
   31    32    33    34    35    36    37    38    39    40
-0.44 -0.04 -0.26 -1.91  1.50 -0.02 -1.54  0.09 -1.49  3.74
   41    42    43    44    45    46    47    48    49    50
-1.84  7.44  3.04 -2.59 -1.11 -1.12  0.99 -0.04  1.04  1.50
   51    52    53    54    55    56    57    58    59    60
-1.14 -0.86 -2.11 -1.03  0.44  0.66 -1.54 -0.81 -1.18  0.71
   61    62    63    64    65
-1.18  2.05 -1.47 -2.46  0.27
```

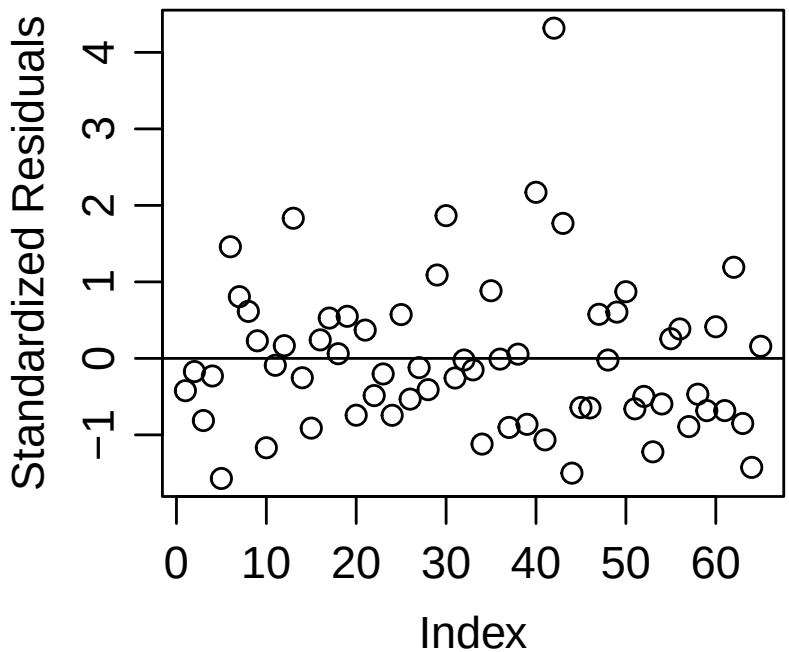
Observation #42 has the largest residual 7.44 (NOT standardized). Is it an outlier?

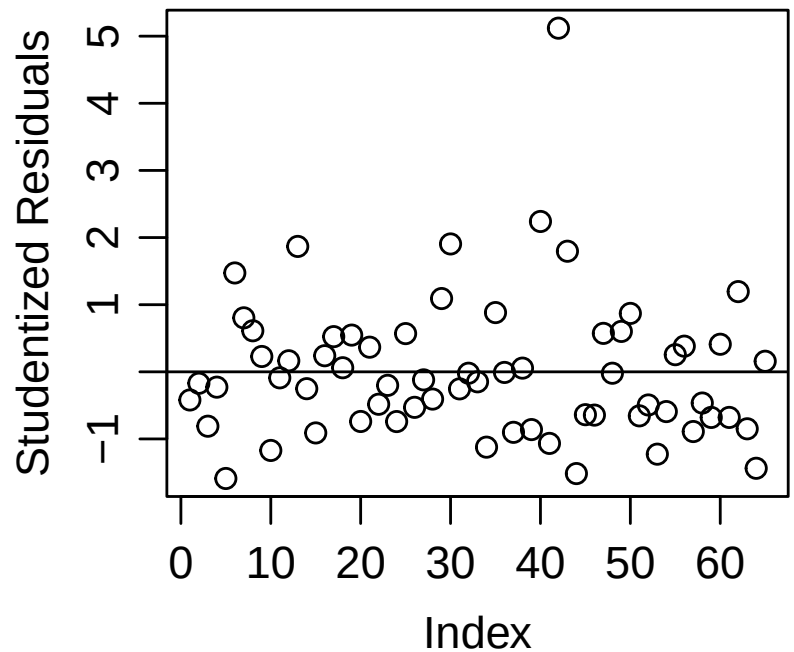
Standardized and Studentized Residuals in R

The `rstandard()` and `rstudent()` command can produce the standardized and studentized residuals in R

```
rstandard(brady1)         # Standardized residuals

         1          2          3          4          5          6
-0.4222885 -0.1718277 -0.8125414 -0.2300744 -1.5697484  1.4590797
(... some output is omitted ...)
        61         62         63         64         65
-0.6823674  1.1907600 -0.8505429 -1.4246593  0.1585103

rstudent(brady1)          # (externally) studentized residual

         1          2          3          4          5          6
-0.4194728 -0.1704770 -0.8102878 -0.2283089 -1.5889328  1.4727714
(... some output is omitted ...)
        61         62         63         64         65
-0.6793980  1.1948600 -0.8486212 -1.4368375  0.1572586
```

```
par(mai=c(.6,.6,.1,.1),mgp=c(2,.7,0))
plot(rstandard(brady1), ylab="Standardized Residuals")
abline(h=0)
plot(rstudent(brady1), ylab="Studentized Residuals")
abline(h=0)
```





There is a potential outlier with a studentized residual > 5.

Residual Plots

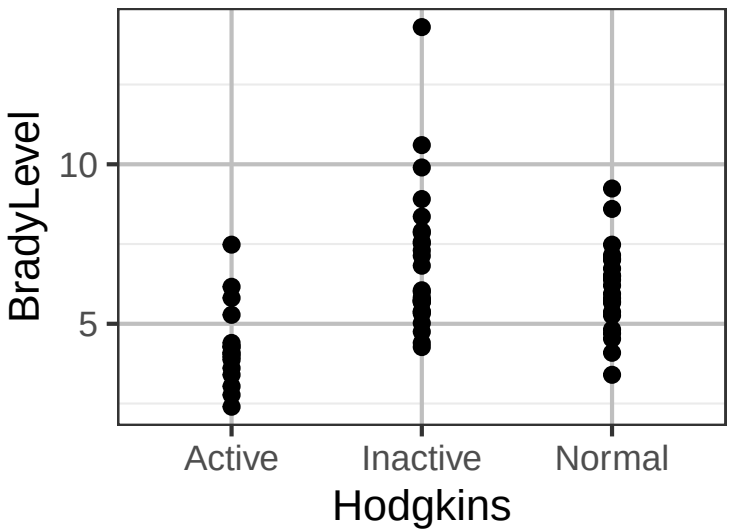
Residual Plots

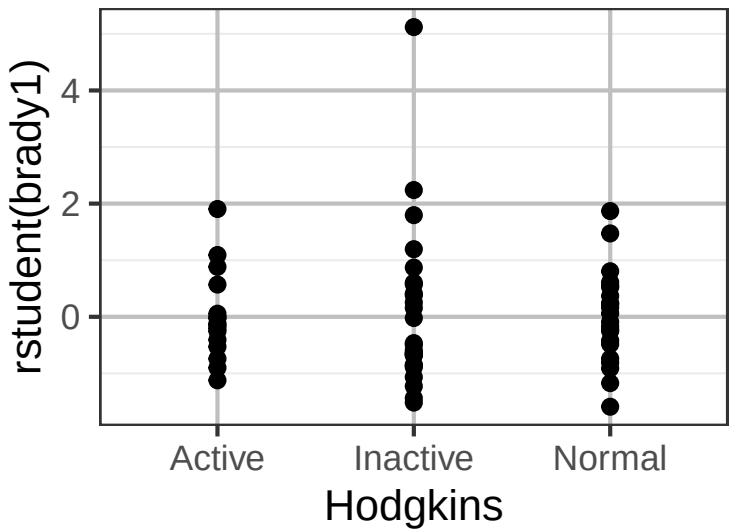
One can make various residuals plots to check model assumptions.

- residuals v.s. treatments
- residuals v.s. fitted values $\widehat{y}_i$
- residuals v.s. variables not in the model
- residuals v.s. time-order observations are made

Residuals v.s. Treatments/Groups

```
library(ggplot2)
ggplot(hodgkins, aes(x=Hodgkins, y=BradyLevel)) + geom_point()
ggplot(hodgkins, aes(x=Hodgkins, y=rstudent(brady1))) + geom_point()
```





If model assumptions are satisfied,

- studentized/standardized residuals should spread evenly above and below 0, almost all within ± 2 or ± 3
- variability of residuals in different treatments should be similar

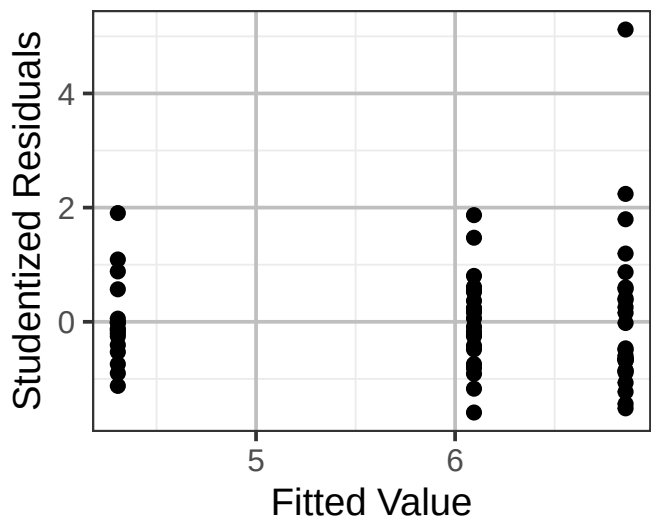
Residuals v.s. Fitted Values

For the multi-sample model $y_{ij} = \mu_i + \varepsilon_{ij}$, the fitted value is

$$\widehat{y}_{ij} = \widehat{\mu}_i = \bar{y}_{i\bullet}$$

```
ggplot(hodgkins, aes(x=fitted(brady1), y=rstudent(brady1))) +
  geom_point() +
  labs(x="Fitted Value", y="Studentized Residuals")
```

Why do points line up vertically?



Checking Constant Variance

Outline

- Why Is Non-Constant Variance a Problem?
- How to check constant variance — Residual Plots
- Remedies
  - Transforming the Response — Variance-Stabilizing Transformation
  - Brown-Forsythe Modified F-test — an alternative to ANOVA F-test
  - Welch Test for Contrasts w/o Constant Variance Assumption

Example — Milk Pasteurization (Oehlert's Text, p.143)

- Goal: to compare 3 pasteurization methods for milk
- Design: 15 samples of milk randomly assigned to the 3 trts
- Response: the bacterial load in each sample after treatment, determined via serial dilution plating
- Data: http://users.stat.umn.edu/~gary/book/fcdae.data/ex6.2

	Method 1	Method 2	Method 3
	26×10^2	35×10^3	29×10^5
	29×10^2	23×10^3	23×10^5
	20×10^2	20×10^3	17×10^5
	22×10^2	30×10^3	29×10^5
	32×10^2	27×10^3	20×10^5
Mean	25.8×10^2	27×10^3	23.6×10^5
SD	492	5874	536656

$\sqrt{\text{MSE}} = 309900$

Example — Milk Pasteurization (Oehlert's Text, p.143)

- Goal: to compare 3 pasteurization methods for milk
- Design: 15 samples of milk randomly assigned to the 3 trts
- Response: the bacterial load in each sample after treatment, determined via serial dilution plating
- Data: http://users.stat.umn.edu/~gary/book/fcdae.data/ex6.2

	Method 1	Method 2	Method 3
	26×10^2	35×10^3	29×10^5
	29×10^2	23×10^3	23×10^5
	20×10^2	20×10^3	17×10^5
	22×10^2	30×10^3	29×10^5
	32×10^2	27×10^3	20×10^5
Mean	25.8×10^2	27×10^3	23.6×10^5
SD	492	5874	536656
Size of Noise	100's	1000's	100,000's

$\sqrt{\text{MSE}} = 309900$

Example — Milk Pasteurization (Oehlert's Text, p.143)

- Goal: to compare 3 pasteurization methods for milk
- Design: 15 samples of milk randomly assigned to the 3 trts
- Response: the bacterial load in each sample after treatment, determined via serial dilution plating
- Data: http://users.stat.umn.edu/~gary/book/fcdae.data/ex6.2

	Method 1	Method 2	Method 3
	26×10^2	35×10^3	29×10^5
	29×10^2	23×10^3	23×10^5
	20×10^2	20×10^3	17×10^5
	22×10^2	30×10^3	29×10^5
	32×10^2	27×10^3	20×10^5
Mean	25.8×10^2	27×10^3	23.6×10^5
SD	492	5874	536656
Size of Noise	100's	1000's	100,000's

$\sqrt{\text{MSE}} = 309900$

$\Rightarrow$ Unequal Variability

Why Non-Constant Variability Causes Problems?

	Method 1	Method 2	Method 3
	26×10^2	35×10^3	29×10^5
	29×10^2	23×10^3	23×10^5
	20×10^2	20×10^3	17×10^5
	22×10^2	30×10^3	29×10^5
	32×10^2	27×10^3	20×10^5
Mean	25.8×10^2	27×10^3	23.6×10^5
SD	492	5874	536656

$\sqrt{\text{MSE}} = 309900$

Why Non-Constant Variability Causes Problems?

	Method 1	Method 2	Method 3
	26×10^2	35×10^3	29×10^5
	29×10^2	23×10^3	23×10^5
	20×10^2	20×10^3	17×10^5
	22×10^2	30×10^3	29×10^5
	32×10^2	27×10^3	20×10^5
Mean	25.8×10^2	27×10^3	23.6×10^5
SD	492	5874	536656

$\sqrt{\text{MSE}} = 309900$

95%C.I. for the mean of Method 1:

$$\begin{aligned}\bar{y}_{1\bullet} \pm t_{0.025,15-3}\frac{\sqrt{\text{MSE}}}{\sqrt{n_1}} &= 2580 \pm 2.179\frac{309900}{\sqrt{5}} \\ &= 2580 \pm 301965 \\ &= (-299385, 304545)\end{aligned}$$

which is far greater than the range of 5 observations for method 1 (2000-3200). What happened?

Checking Constant Variance Using Residual Plots

- Residuals v.s. Fitted Values
- Residuals v.s. Treatments
- Residuals v.s. Other Variables not in the model

If the constant variance assumption is true, residuals will evenly spread around the zero line.

Rule of thumb: ANOVA F-tests for CRD can tolerate non-constant variance to some extent, so do tests for contrasts.

It usually fine as long as

$$\frac{\text{maximum of } s_1, \ldots, s_g}{\text{minimum of } s_1, \ldots, s_g} \leq 2,\ 3 \text{ or even } 4,$$

where s_i is the sample SD of the ith group, especially when the group sizes n_i are (roughly) equal.

Example — Resin Glue Failure Time — Background

Question: How to measure the lifetime of things like computer disk drives, light bulbs, and glue bonds?
E.g., a computer drive is claimed to have a lifetime of 800,000 hours ($>$ 90 years).
Clearly the manufacturer did not have disks on test for 90 years; how do they make such claims?

Solution: *Accelerated life test*

Parts under stress (higher load, higher temperature, etc.) will usually fail sooner than parts that are unstressed. By modeling the lifetimes of parts under various stresses, we can estimate (extrapolate) the lifetime of parts that are unstressed.

- Example: resin glue failure time

Example — Resin Glue Failure Time

- Goal: to estimate the life time (in hours) of an encapsulating resin for gold-aluminum bonds in integrated circuits (operating at 120°C)[1]
- Method: accelerated life test
- Design: Randomly assign 37 units to one of 5 different temperature stresses (in Celsius)

$$175^\circ, 194^\circ, 213^\circ, 231^\circ, 250^\circ$$

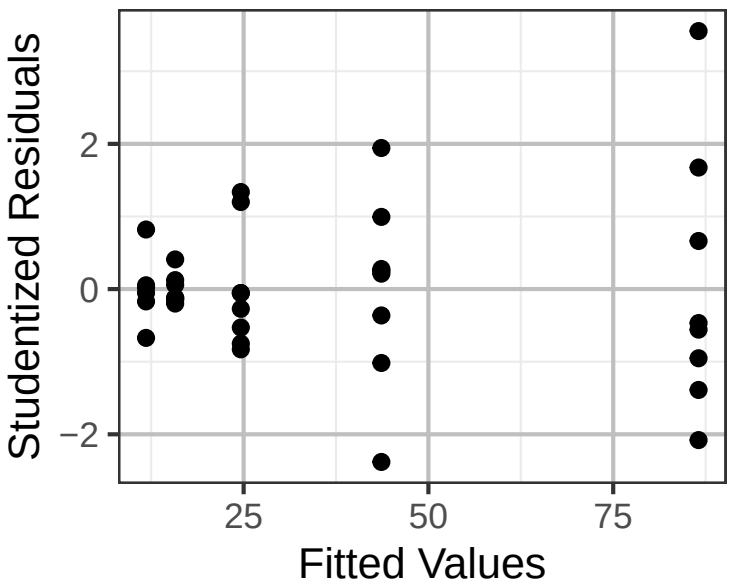
- Treatments: temperature in Celsius
- Response: time to failure in hours of the tested material

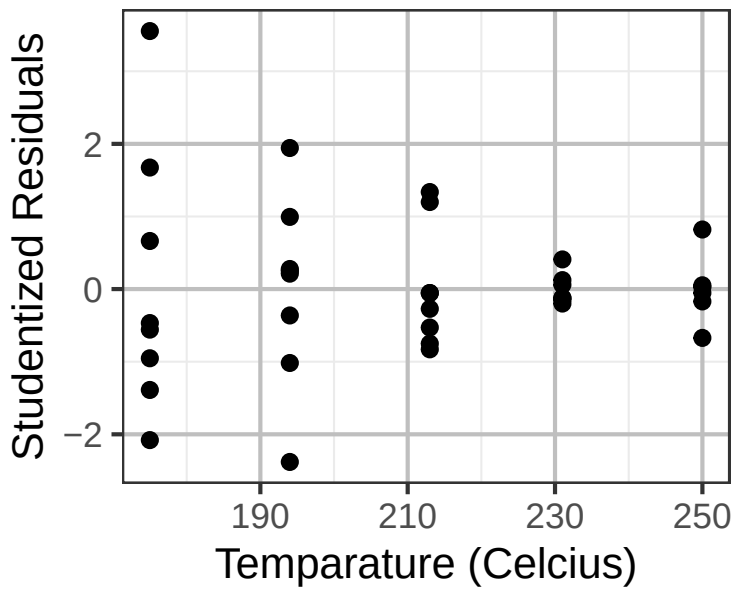
Temperature	Failure Time in Hours							
175°C	110.0	82.2	99.1	82.9	71.3	91.7	76.0	79.2
194°C	45.8	51.3	26.5	58.0	45.3	40.8	35.8	45.6
213°C	33.8	34.8	24.2	20.5	22.5	18.8	18.2	24.2
231°C	14.2	16.7	14.8	14.6	16.2	18.9	14.8	
250°C	18.0	6.7	12.0	10.5	12.2	11.4		

[1]Source: p. 448-449, *Accelerated Testing* (Nelson 2004). Original data is provided by Dr. Muhib Khan of AMD.

Example: Resin Glue Data

```
resin = read.table(
    "http://www.stat.uchicago.edu/~yibi/s222/resinlife.txt", h=T)
lm1 = lm(life ~ as.factor(temp), data=resin)
ggplot(resin, aes(x=fitted(lm1), y=rstudent(lm1)))+geom_point()+
  labs(x="Fitted Values", y="Studentized Residuals")
ggplot(resin, aes(x=temp, y=rstudent(lm1)))+geom_point()+
  labs(x="Temparature (Celcius)", y="Studentized Residuals")
```





- Variability of residuals increases with the fitted value, but decreases with the temperature

Remedies for Non-Constant Variance

Remedy 1: Variance-Stabilizing Transformation

If the SD σ (the spread of residuals) changes the mean μ (the fitted values), you can try a *variance-stabilizing transformation* of the response to make the variance (closer to) constant.

- if the SD is proportional to the fitted value, then

$$y \to \log(y)$$

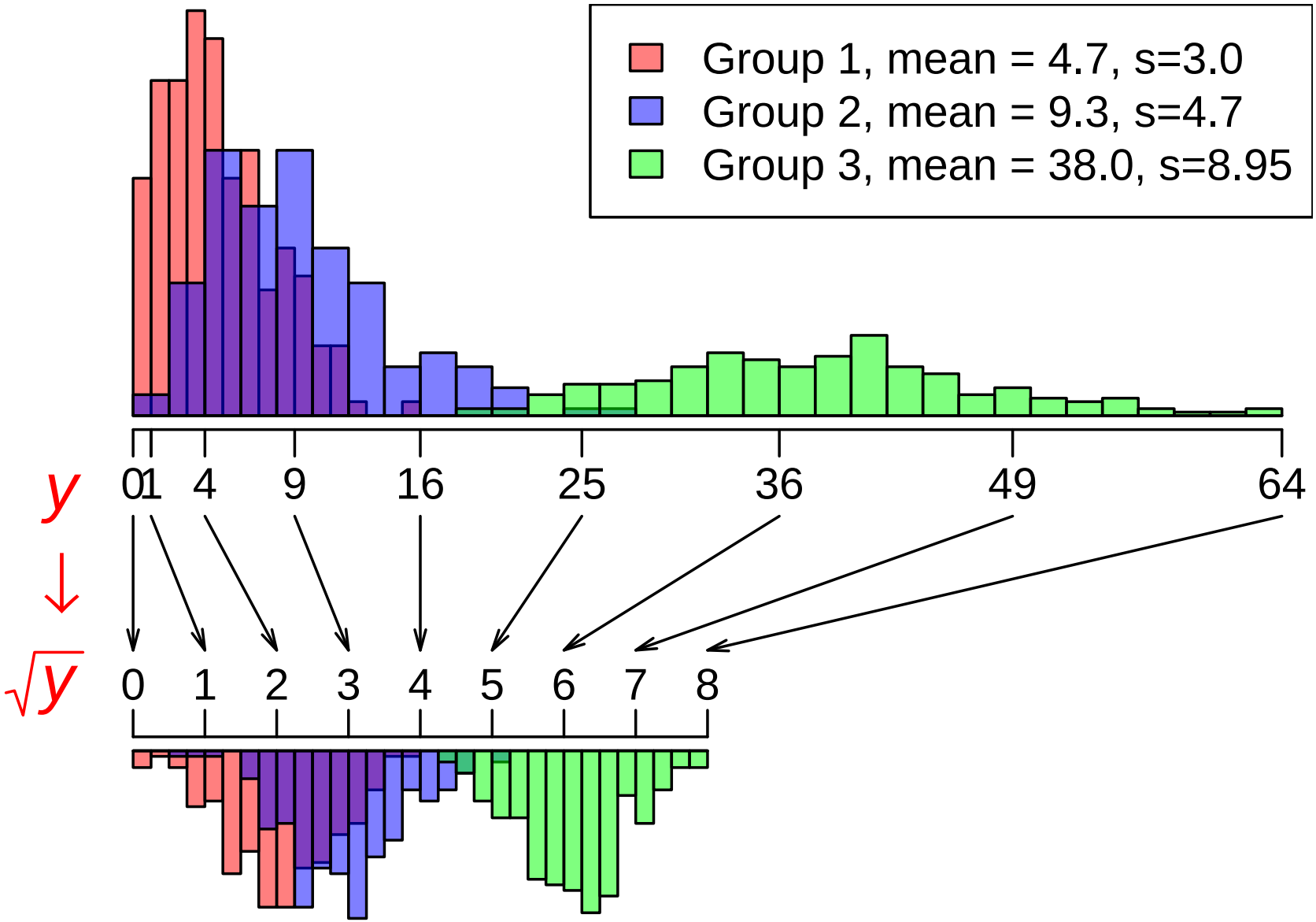
- if the SD is proportional to $\sqrt{\text{the fitted value}}$, i.e., the variance is proportional to the fitted value, then

$$y \to \sqrt{y}$$

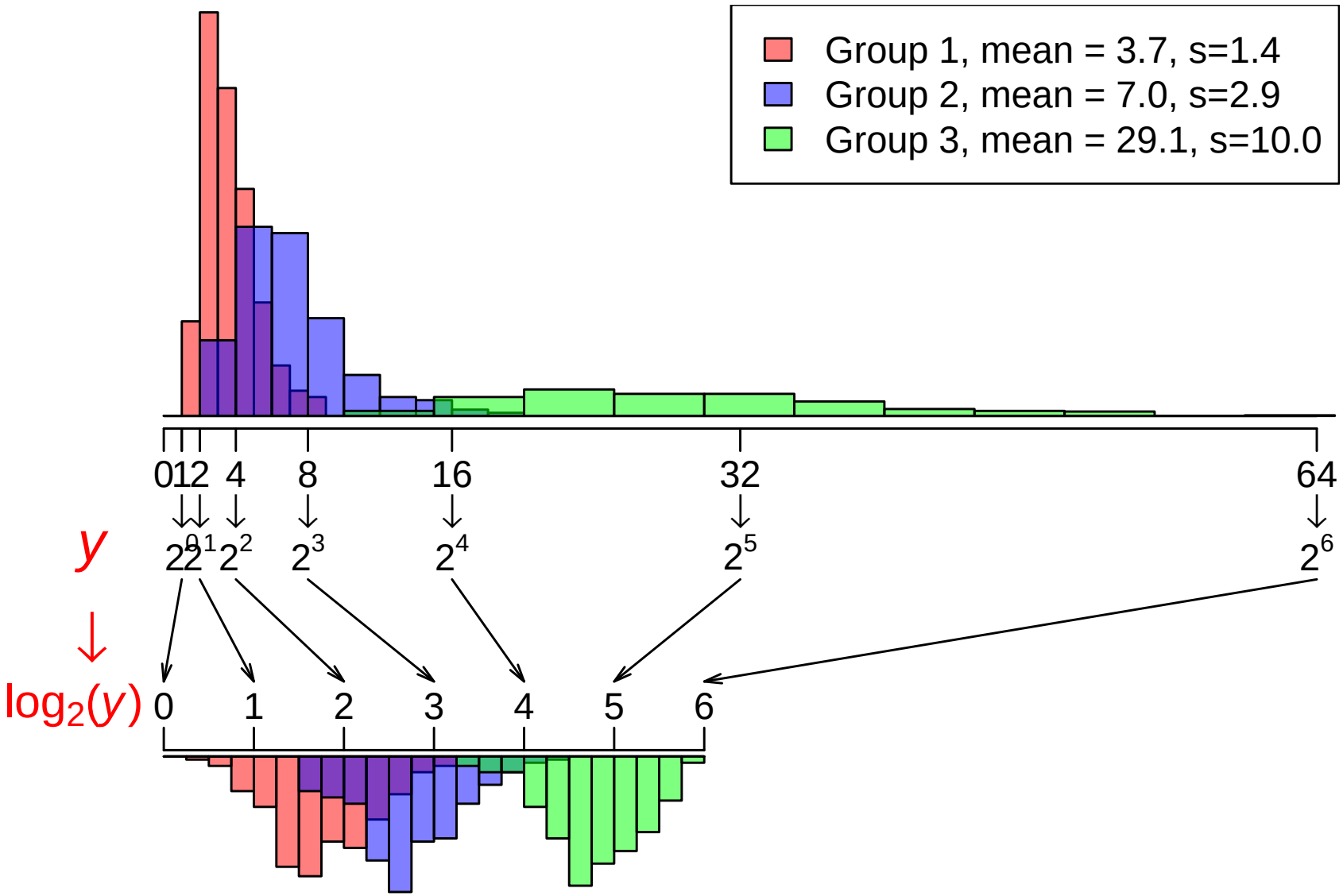
- In general, if the SD σ is proportional to (the fitted values)$^{\alpha}$, then the variance-stabilizing transformation is

$$y \to \begin{cases} y^{1-\alpha} & \text{for } \alpha \neq 1 \\ \log(y) & \text{for } \alpha = 1 \end{cases}$$

How Variance-Stabilizing Transformation Work? (1)

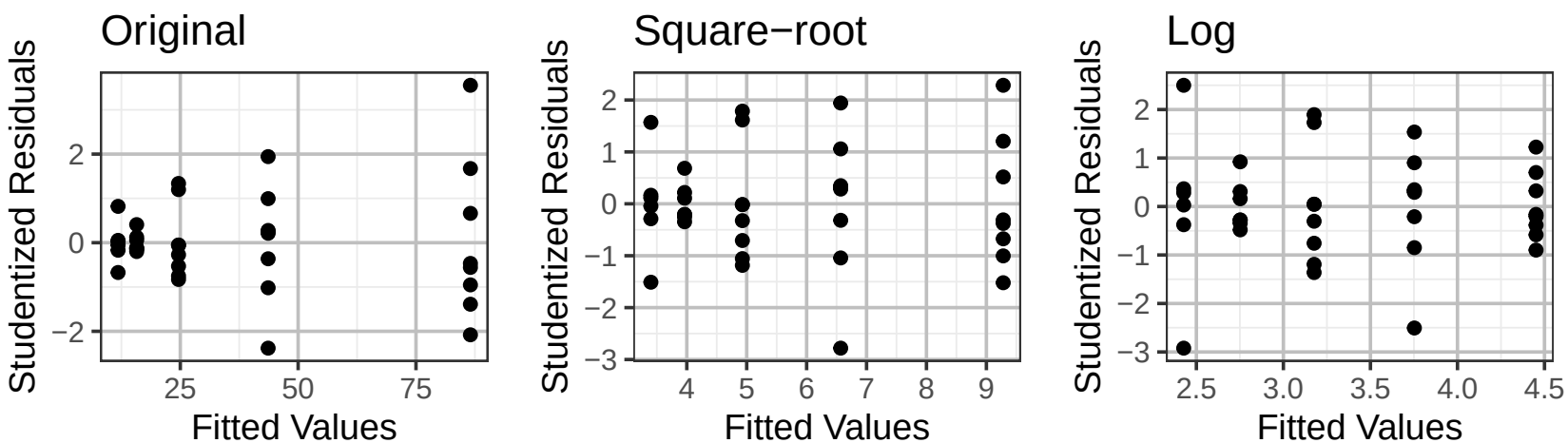


How Variance-Stabilizing Transformation Work? (2)



Example: Resin Glue

```
lm1 = lm(life ~ as.factor(temp), data=resin)
lm2 = lm(sqrt(life) ~ as.factor(temp), data=resin)
lm3 = lm(log(life) ~ as.factor(temp), data=resin)
ggplot(resin, aes(x=fitted(lm1), y=rstudent(lm1)))+geom_point()+
 labs(x="Fitted Values", y="Studentized Residuals", title="Original")
ggplot(resin, aes(x=fitted(lm2), y=rstudent(lm2)))+geom_point()+
 labs(x="Fitted Values", y="Studentized Residuals", title="Square-root"
ggplot(resin, aes(x=fitted(lm3), y=rstudent(lm3)))+geom_point()+
 labs(x="Fitted Values", y="Studentized Residuals", title="Log")
```



How to Select a Power Transformation?

In many cases, we don't have a good idea what is the value of α. We can still try power transformation of the response.

$$f_\lambda(y) = \begin{cases} y^\lambda, & \text{if } \lambda \neq 0 \\ \log(y), & \text{if } \lambda = 0. \end{cases}$$

How to select λ?

- Trial and error: try convenient power like $-1, -1/2, -1/3, 0, 1/3, 1/2, 2, \ldots$ and then check residual plots for each of them for the constant variance.
- *Box-Cox method*: See next page

Box-Cox Method

Box-Cox method is an automatic procedure to select the **best** power λ that make the residuals of the model

$$y_{ij}^{\lambda} = \mu_i + \varepsilon_{ij}$$

closest to normal.

Box-Cox Method

Box-Cox method is an automatic procedure to select the **best** power λ that make the residuals of the model

$$y_{ij}^{\lambda} = \mu_i + \varepsilon_{ij}$$

closest to normal.

- We usually round the optimal λ to a convenient power like

$$-1,\ -\frac{1}{2},\ -\frac{1}{3},\ 0,\ \frac{1}{3},\ \frac{1}{2},\ 1,\ 2, \ldots$$

since the practical difference of $y^{0.5827}$ and $y^{0.5}$ is usually small, but the square-root transformation is much easier to interpret.

Box-Cox Method

Box-Cox method is an automatic procedure to select the **best** power λ that make the residuals of the model

$$y_{ij}^{\lambda} = \mu_i + \varepsilon_{ij}$$

closest to normal.

- We usually round the optimal λ to a convenient power like

$$-1,\ -\frac{1}{2},\ -\frac{1}{3},\ 0,\ \frac{1}{3},\ \frac{1}{2},\ 1,\ 2, \ldots$$

  since the practical difference of $y^{0.5827}$ and $y^{0.5}$ is usually small, but the square-root transformation is much easier to interpret.
- A confidence interval for the optimal λ can also be obtained.

We usually select a convenient power λ^* in this C.I.

- Though Box-Cox is developed to select a power transformation making the residuals as *normal* as possible, it's been shown that the optimal λ is often close to the variance-stabilizing λ.

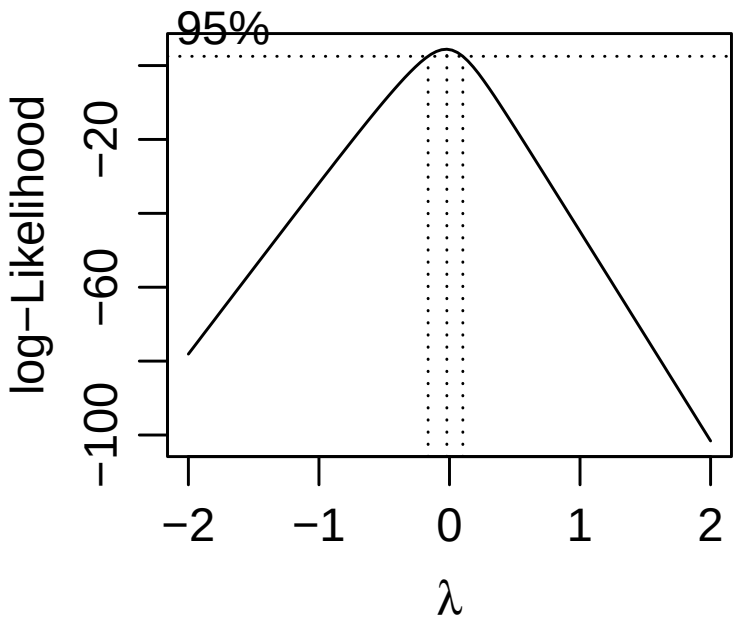
Box-Cox Method in R

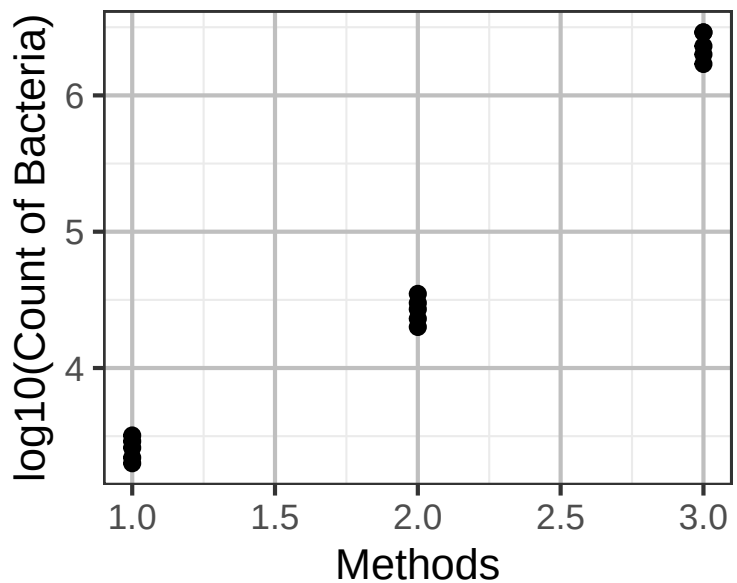
In R, one must first load the `MASS` library to use `boxcox()`.

The `boxcox()` function can be applied on a model formula or a `lm()` model.

Example – Count of Bacteria – Box-Cox

```
ex6.2 = read.table(
        "http://users.stat.umn.edu/~gary/book/fcdae.data/ex6.2", h=T)
library(MASS)
par(mai=c(.6,.6,.1,.1),mgp=c(2,.7,0))
boxcox(count ~ as.factor(method), data=ex6.2)
ggplot(ex6.2, aes(x=method, y=log10(count))) + geom_point() +
  labs(x="Methods", y="log10(Count of Bacteria)")
```

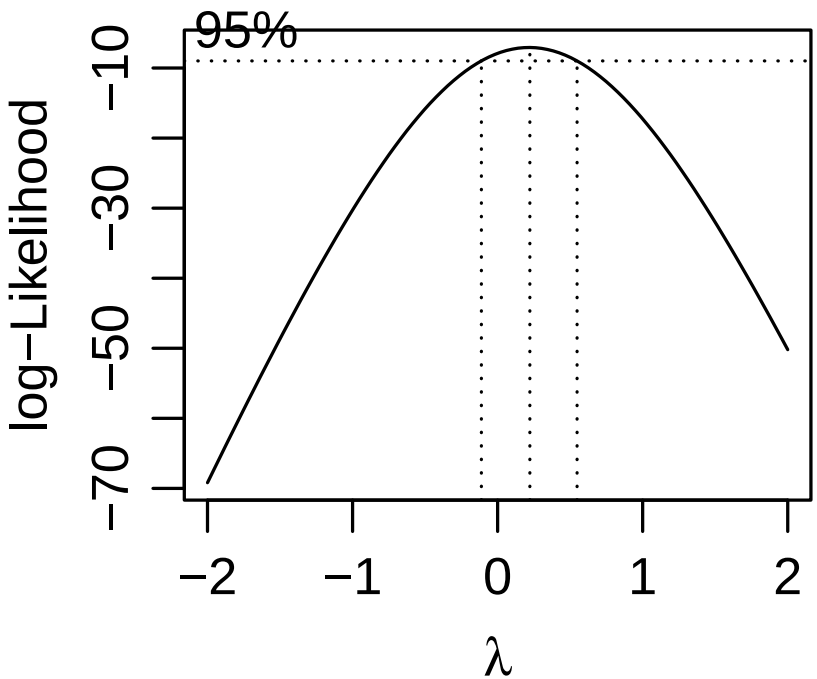




After log transformation, the 3 groups look even in variability.

Example: Resin Glue — Box-Cox

```
library(MASS)
par(mai=c(.6,.6,.1,.1),mgp=c(2,.7,0))
boxcox(life ~ as.factor(temp), data=resin)
```



The 95% C.I. for λ contains both 0 and $1/2$. As $\lambda = 1/2$ is very close to the boundary of the C.I, $\lambda = 0$ seems to be a better choice, which is consistent with the Arrhenius Law in Thermodynamics.

Drawbacks of Transformation

- Except for a few special transformation (log, $\sqrt{\ }$, inverse), the transformed response usually lacks natural interpretation (How to interpret $y^{0.1}$?)
- Unless having a good interpretation on the transformed response, think again before making transformations

Remember that ANOVA tests have some tolerance for non-constant variance. If

$$\frac{\text{maximum of } s_1, \ldots, s_g}{\text{minimum of } s_1, \ldots, s_g} \leq 2,\ 3 \text{ or even } 4,$$

where s_i is the sample SD of the ith group, don't worry too much about non-constant variance.
In that case, it is fine to leave the response untransformed.

```
tapply(resin$life, resin$temp, sd)
   175    194    213    231    250
12.896  9.558  6.379  1.661  3.647
tapply(sqrt(resin$life), resin$temp, sd)
   175    194    213    231    250
0.6802 0.7505 0.6223 0.2049 0.5306
tapply(log(resin$life), resin$temp, sd)
   175    194    213    231    250
0.1441 0.2393 0.2451 0.1013 0.3178
```

The ratio of the largest and smallest SD is

$$\begin{cases} 12.896/1.661 \approx 7.76 & \text{if untransformed} \\ 0.7505/0.2049 \approx 3.66 & \text{if square-root transformed} \\ 0.3178/0.1013 \approx 3.14 & \text{if log transformed.} \end{cases}$$

Both log and square-root transformation are acceptable, though log makes variance more even.

Brown-Forsythe Modified F-test

If one cannot find an appropriate transformation, **Brown-Forsythe modified F-test** is an alternative of the ANOVA F-test that doesn't rely on the constant variance assumption. The BF test statistic is

$$BF = \frac{\sum_{i=1}^{g} n_i(\overline{y}_{i\bullet} - \overline{y}_{\bullet\bullet})^2}{\sum_{i=1}^{g} s_i^2(1 - n_i/N)} = \frac{\text{SS}_{trt}}{\sum_{i=1}^{g} s_i^2(1 - n_i/N)}$$

in which s_i^2 is the sample variance in treatment i. Under the null hypothesis of equal treatment means, BF is approximately distributed as an F-distribution with $g - 1$ and ν degrees of freedom, where

$$\nu = \frac{(\sum_{i=1}^{g} d_i)^2}{\sum_{i=1}^{g} d_i^2/(n_i - 1)} \quad \text{in which} \quad d_i = s_i^2(1 - n_i/N).$$

Satterthwaite's Approx. for Contrasts w/ Unequal Variance

If one cannot find an appropriate transformation, try **Satterthwaite's Approximation test for a contrast** $\sum_{i=1}^{g} c_i \mu_i$, which doesn't rely on the constant variability assumption. The test statistic

$$t = \frac{\sum_{i=1}^{g} c_i \overline{y}_{i\bullet}}{\sqrt{\sum_{i=1}^{g} c_i^2 s_i^2 / n_i}} \text{ has an approx. t-distn w/ } df = \frac{\left(\sum_{i=1}^{g} c_i^2 s_i^2 / n_i\right)^2}{\sum_{i=1}^{g} \frac{1}{n_i - 1} \frac{c_i^4 s_i^4}{n_i^2}}$$

Specifically, for pairwise comparison $\mu_i - \mu_k$,

$$t = \frac{\overline{y}_{i\bullet} - \overline{y}_{k\bullet}}{\sqrt{\frac{s_i^2}{n_i} + \frac{s_k^2}{n_k}}} \text{ has an approx. t-distn w/ } df = \frac{\left(s_i^2/n_i + s_k^2/n_k\right)^2}{\frac{s_i^4}{n_i^2(n_i - 1)} + \frac{s_k^4}{n_k^2(n_k - 1)}}$$

This is a generalization of the two-sample test without the equal variance assumption.

Satterthwaite's Approximation for Tukey's Method

Without the equal variance assumption, for testing
$H_0 : \mu_i - \mu_k = 0$ v.s. $H_a : \mu_i - \mu_k \neq 0$, reject H_0 if

$$|t_0| = \frac{|\bar{y}_{i\bullet} - \bar{y}_{k\bullet}|}{\sqrt{\frac{s_i^2}{n_i} + \frac{s_k^2}{n_k}}} > \frac{q_{g,df,\alpha}}{\sqrt{2}}, \quad \text{where } df = \frac{\left(s_i^2/n_i + s_k^2/n_k\right)^2}{\frac{s_i^4}{n_i^2(n_i-1)} + \frac{s_k^4}{n_k^2(n_k-1)}}$$

Tukey's HSD becomes

$$\frac{q_{g,df,\alpha}}{\sqrt{2}} \times \sqrt{\frac{s_i^2}{n_i} + \frac{s_k^2}{n_k}}.$$

for the formula for the df is as shown above.

Check Independence

Check Independence

- Among the model assumptions, violation of the independence assumption causes severest problem. Most of our the analysis (ANOVA, test of contrasts, multiple comparisons, etc.) have little tolerance on dependence of errors
- There are various forms of dependence, *serial dependence* and *spatial dependence* are two common ones
- **Remedies for Dependence**
  - There isn't much we can do about dependence using our current machinery, since no simple transformation can remove dependence.
  - Analysis of dependent data requires tools like **time series** or **spatial statistics**, which is beyond the scope of this class

Residuals v.s. Time-Order Plot — Checking Time Dependence

- If there is a **time order** that the data are collected, please plot the residuals against the time order.
- If the time-plot of the residuals exhibit any pattern, . . .
- Better **keep track of the time order** the data are collected.

Example: Balloon Experiment (p.66 Textbook)

- Goal: To determine whether balloons of different colors are similar in terms of the time taken for inflation to a diameter of 7 inches.
- Four colors were selected from a single manufacturer.
- An assistant blew up the balloons and the experimenter recorded the times (to the nearest 1/10 second) with a stop watch. The data, in the order collected, are given in next page where the codes 1, 2, 3, 4 denote the colors pink, yellow, orange, blue, respectively.
- Any flaw in the design of the experiment?

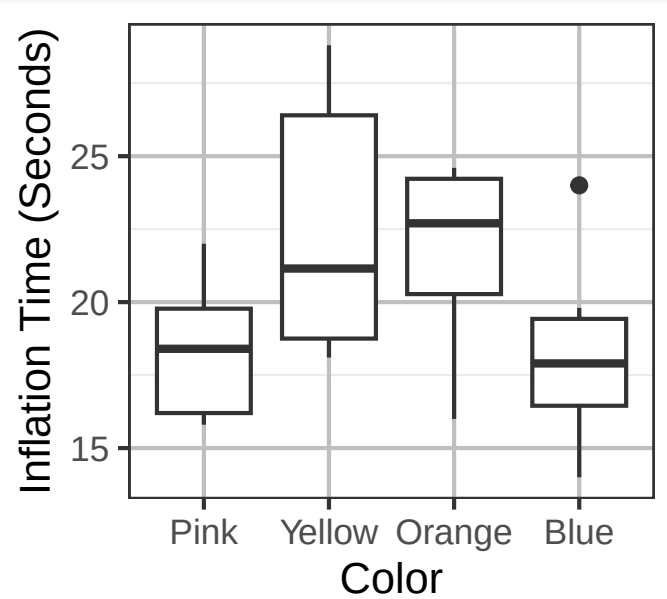
Example: Balloon Experiment — Data

Times (in seconds) for the balloon experiment

Time Order	1	2	3	4	5	6	7	8
Coded color	1	3	1	4	3	2	2	2
Inflation Time	22.4	24.6	20.3	19.8	24.3	22.2	28.5	25.7
Time Order	9	10	11	12	13	14	15	16
Coded color	3	1	2	4	4	4	3	1
Inflation Time	20.2	19.6	28.8	24.0	17.1	19.3	24.2	15.8
Time Order	17	18	19	20	21	22	23	24
Coded color	2	1	4	3	1	4	4	2
Inflation Time	18.3	17.5	18.7	22.9	16.3	14.0	16.6	18.1
Time Order	25	26	27	28	29	30	31	32
Coded color	2	4	2	3	3	1	1	3
Inflation Time	18.9	16.0	20.1	22.5	16.0	19.3	15.9	20.3

```
balloon = read.table(
  "http://www.stat.uchicago.edu/~yibi/s222/balloon.txt", h=T)
balloon$Color = factor(balloon$Color,
                       labels=c("Pink", "Yellow", "Orange", "Blue"))
lmball = lm(Time ~ Color, data=balloon)
anova(lmball)
Analysis of Variance Table

Response: Time
          Df Sum Sq Mean Sq F value Pr(>F)
Color      3    128    42.6    3.94  0.018
Residuals 28    303    10.8
```

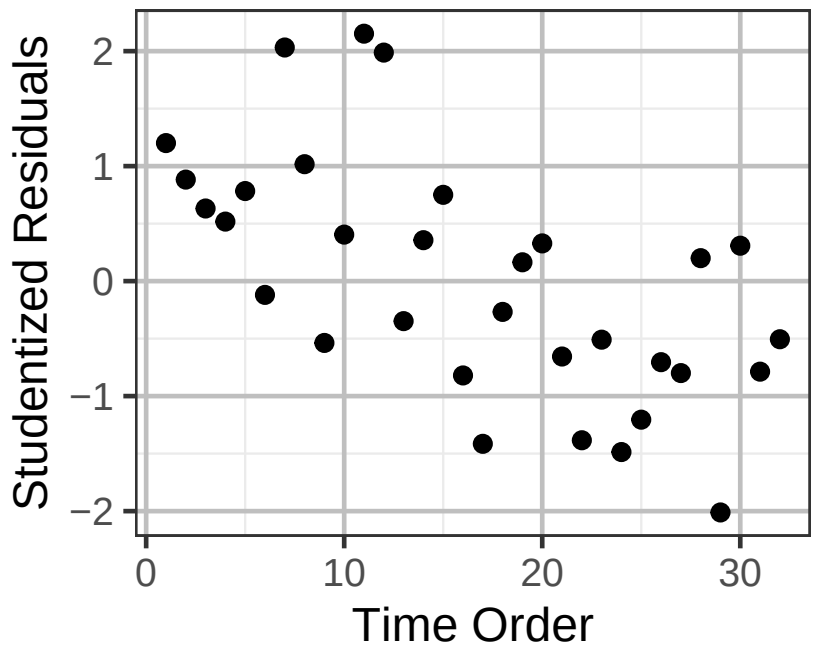
```
library(ggplot2)
ggplot(balloon, aes(x=Color, y=Time)) +
  geom_boxplot() +
  labs(y="Inflation Time (Seconds)")
```



Balloon Experiment — Residual Time Plot

```
ggplot(balloon, aes(x=Order, y=rstudent(lmball))) + geom_point() +
  labs(y="Studentized Residuals", x="Time Order")
```

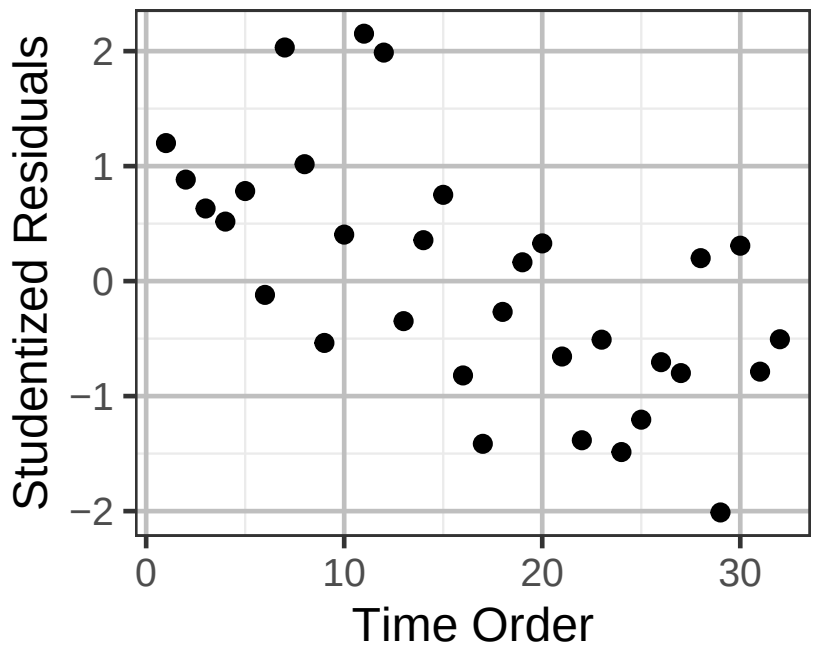
A clear downward drift in the residuals as time went by. Why?



Balloon Experiment — Residual Time Plot

```
ggplot(balloon, aes(x=Order, y=rstudent(lmball))) + geom_point() +
  labs(y="Studentized Residuals", x="Time Order")
```

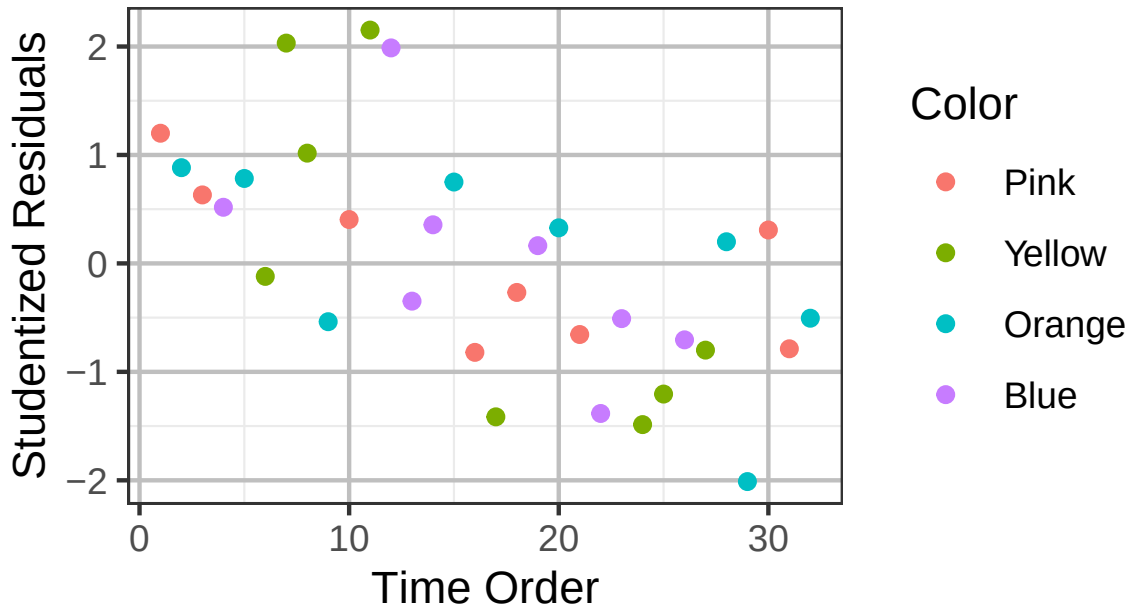
A clear downward drift in the residuals as time went by. Why?



A single assistant blew up all the balloons in the experiment, and has become more skillful (required less time to inflate the balloon to the required size) as the he blew up more balloons.

Balloon Experiment — Why Randomize?

```
ggplot(balloon, aes(x=Order, y=rstudent(lmball), color=Color)) +
  geom_point() + labs(y="Studentized Residuals", x="Time Order")
```



Why is it important to *randomize*?

What would happen if the assistant blew all the 8 pink balloons first, then the 8 yellow ones, the 8 orange ones, and the 8 blue ones at last?

The residuals shows a clear downward linear time trend. Clearly, the assistant got more experienced and required less time to blew up a balloon, as he blew up more.

If all the 8 pink balloons were inflated first, we would obtain a higher mean inflation time for pink balloons than if they were inflated last. The effect of time order and balloon colors would be confounded. We won't be able to tell whether the difference in mean inflation time is due to the different color or due to the time-order.

Example: Standard Gravity

The National Bureau of Standards performed 8 series of experiments in 1924-35 to determine g, the standard gravity.

The data are given in the table below (in deviations from $9.8 m/s^2 \times 10^5$, e.g., the first measurement of g is 9.80076 m/s^2), with series 1 representing the earliest set of experiments and series 8 the last.

Series	Measurements												
1	76	82	83	54	35	46	87	68					
2	87	95	98	100	109	109	100	81	75	68	67		
3	105	83	76	75	51	76	93	75	62				
4	95	90	76	76	87	79	77	71					
5	76	76	78	79	72	68	75	78					
6	78	78	78	86	87	81	73	67	75	82	83		
7	82	79	81	79	77	79	79	78	79	82	76	73	64
8	84	86	85	82	77	76	77	80	83	81	78	78	78

Weird ANOVA F-Test Result

```
g = c(76,82,83,54,35,46,87,68,87,95,98,100,109,109,100,81,75,68,67,
  105,83,76,75,51,76,93,75,62,95,90,76,76,87,79,77,71,
   76,76,78,79,72,68,75,78,78,78,78,86,87,81,73,67,75,82,83,
   82,79,81,79,77,79,79,78,79,82,76,73,64,
   84,86,85,82,77,76,77,80,83,81,78,78,78)
series = c(rep(1,8),rep(2,11),rep(3,9),rep(4,8),rep(5,8),rep(6,11),
             rep(7,13),rep(8,13))
lmg = lm(g ~ as.factor(series))
anova(lmg)
                  Df Sum Sq Mean Sq F value Pr(>F)
as.factor(series)  7   2819     403    3.57 0.0024
Residuals         73   8239     113
```

ANOVA rejects the H_0 of the 8 series having equal means. What does this mean? Will you conclude that

(a) g had changed in the 8 series of measurements, or
(b) the ANOVA F-test failed?

Weird ANOVA F-Test Result

```
g = c(76,82,83,54,35,46,87,68,87,95,98,100,109,109,100,81,75,68,67,
  105,83,76,75,51,76,93,75,62,95,90,76,76,87,79,77,71,
   76,76,78,79,72,68,75,78,78,78,78,86,87,81,73,67,75,82,83,
   82,79,81,79,77,79,79,78,79,82,76,73,64,
   84,86,85,82,77,76,77,80,83,81,78,78,78)
series = c(rep(1,8),rep(2,11),rep(3,9),rep(4,8),rep(5,8),rep(6,11),
             rep(7,13),rep(8,13))
lmg = lm(g ~ as.factor(series))
anova(lmg)
                  Df Sum Sq Mean Sq F value Pr(>F)
as.factor(series)  7   2819     403    3.57 0.0024
Residuals         73   8239     113
```

ANOVA rejects the H_0 of the 8 series having equal means. What does this mean? Will you conclude that

(a) g had changed in the 8 series of measurements, or
(b) the ANOVA F-test failed?

If your answer is (a), how do you explain the change of g?

If your answer is (b), why the ANOVA F-test failed?

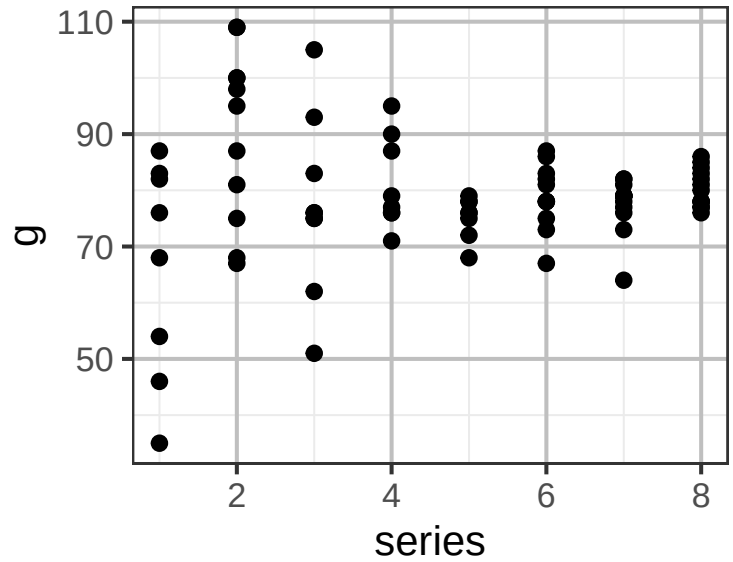
Example: Standard Gravity

The National Bureau of Standards (NBS) (now called the National Institute of Standards and Technology (NIST))is the government agency that measures things. The following statement is taken from the NIST website:

> *Founded in 1901, NIST is a non-regulatory federal agency within the U.S. Department of Commerce. NIST mission is to promote U.S. innovation and industrial competitiveness by advancing measurement science, standards, and technology in ways that enhance economic security and improve our quality of life.*

Thus, it is safe to assume that the NBS scientists were trying hard to measure the same quantity g (e.g., all experiments were done in the same location) throughout all 8 series of experiments.

```
grav = data.frame(series, g)
ggplot(grav, aes(x=series, y=g)) +
  geom_point()
```



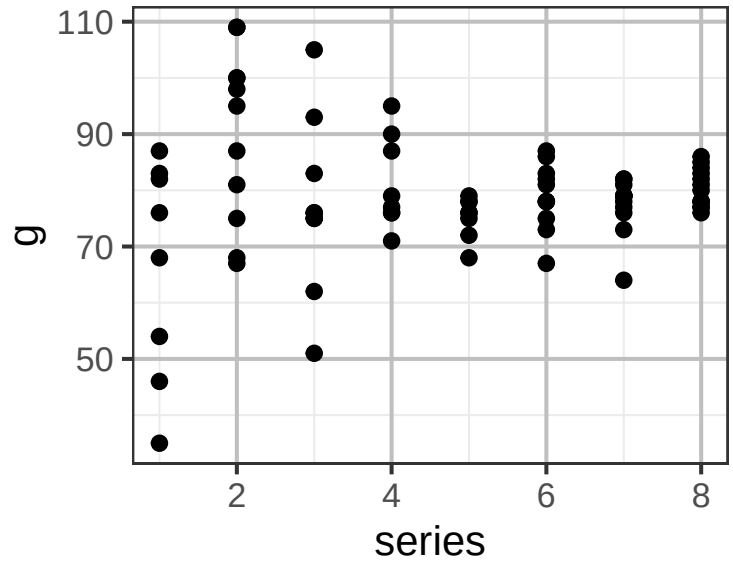
Variance decreases with series, which makes sense since the accuracy of measurement improved as time went by.

The ANOVA F-test here may be unreliable since the constant variance assumption is not met

```
round(tapply(g, series, sd), 2)
    1     2     3     4     5     6     7     8
19.25 15.29 15.76  8.30  3.65  5.84  4.74  3.36
```

Will a transformation work here? Box-Cox?

```
grav = data.frame(series, g)
ggplot(grav, aes(x=series, y=g)) +
  geom_point()
```



Variance decreases with series, which makes sense since the accuracy of measurement improved as time went by.

The ANOVA F-test here may be unreliable since the constant variance assumption is not met

```
round(tapply(g, series, sd), 2)
    1     2     3     4     5     6     7     8
19.25 15.29 15.76  8.30  3.65  5.84  4.74  3.36
```

Will a transformation work here? Box-Cox?

No. The variance-stabilizing transformation works only when the variability increases or decreases with the mean. Here the means of the 8 series are nearly the same.

Brown-Forsythe Modified F-test

In view of the nonconstant variance, let's try the Brown-Forsythe modified F-test

$$BF = \frac{SS_{Trt}}{\sum_{i=1}^{g} s_i^2(1 - n_i/N)}$$

The numerator $SS_{Trt} = 2819$ can be found in the ANOVA table above. The denominator is found using R (see the codes below) to be 888.5747

```
sds = tapply(g, series, sd); sds
     1      2      3      4      5      6      7      8
19.250 15.293 15.756  8.297  3.655  5.839  4.737  3.355
ni = c(8, 11, 9, 8, 8, 11, 13, 13)
di  = sds^2*(1-ni/sum(ni))
BFbottom = sum(di)
BFbottom
[1] 888.6
BF = 2819/BFbottom
```

The BF-statistic is thus $\text{BF} = \dfrac{2819}{888.5747} = 3.1725.$

Under the null hypothesis of equal group means, BF has an approx. F-distribution with df1$= g - 1$ and df2 given below

$$\text{df2} = \frac{(\sum_{i=1}^{g} d_i)^2}{\sum_{i=1}^{g} d_i^2/(n_i - 1)} \quad \text{in which} \quad d_i = s_i^2(1 - n_i/N)$$

where the second degrees of freedom df2 is calculated as 29.46 in the R code below.

```
df2 = (BFbottom)^2/sum(di^2/(ni-1))
df2
[1] 29.46
pf(BF, 8-1, df2, lower.tail=F)          # P-value of the BF-test
[1] 0.01263
```

However, the BF-test, not relying on the constant variance assumption, also rejects the null hypothesis of equal mean at a P-value 0.0126. Why the BF-test also failed?

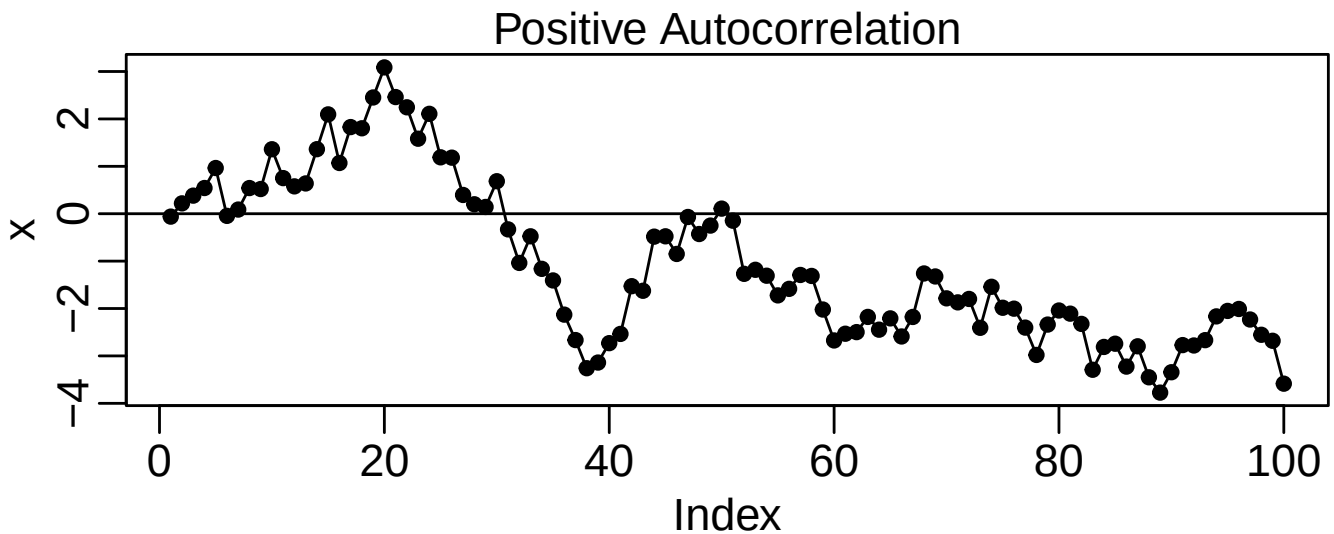
Tools for Checking Serial Dependence

1. *Time plot*: a plot of residuals v.s. the order they are measured)
2. *Lag Plots*
3. *Autocorrelation* & *Autocorrelation Function*

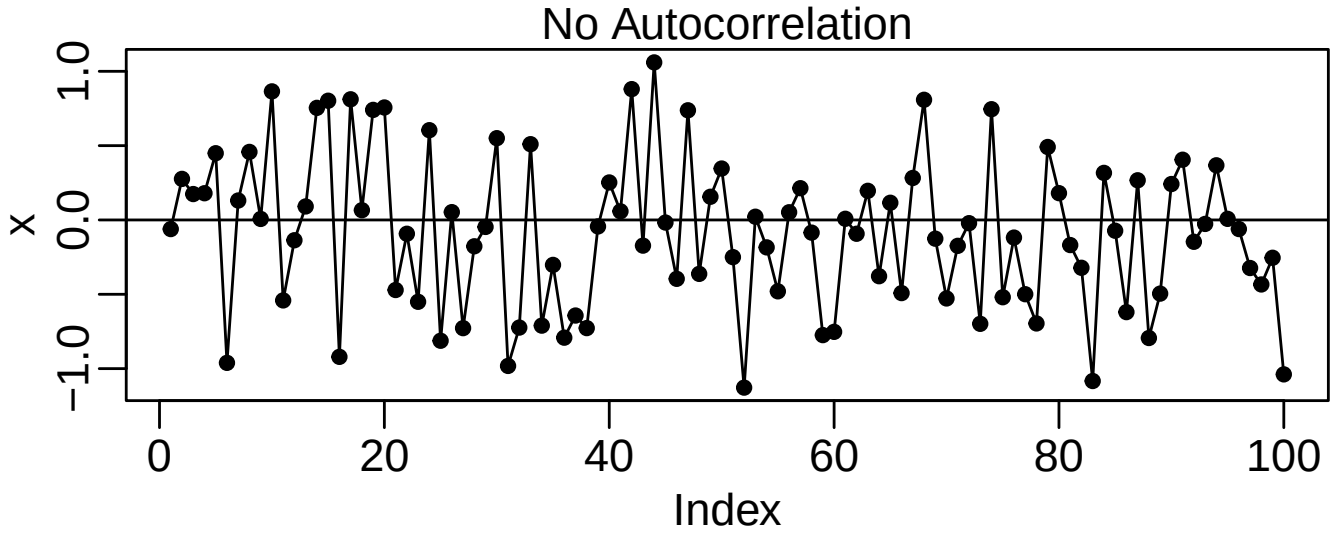
Time Plot of Residuals

A time plot of residuals is a plot of residuals v.s. the (time) order they are recorded, where the points are connected by a line.

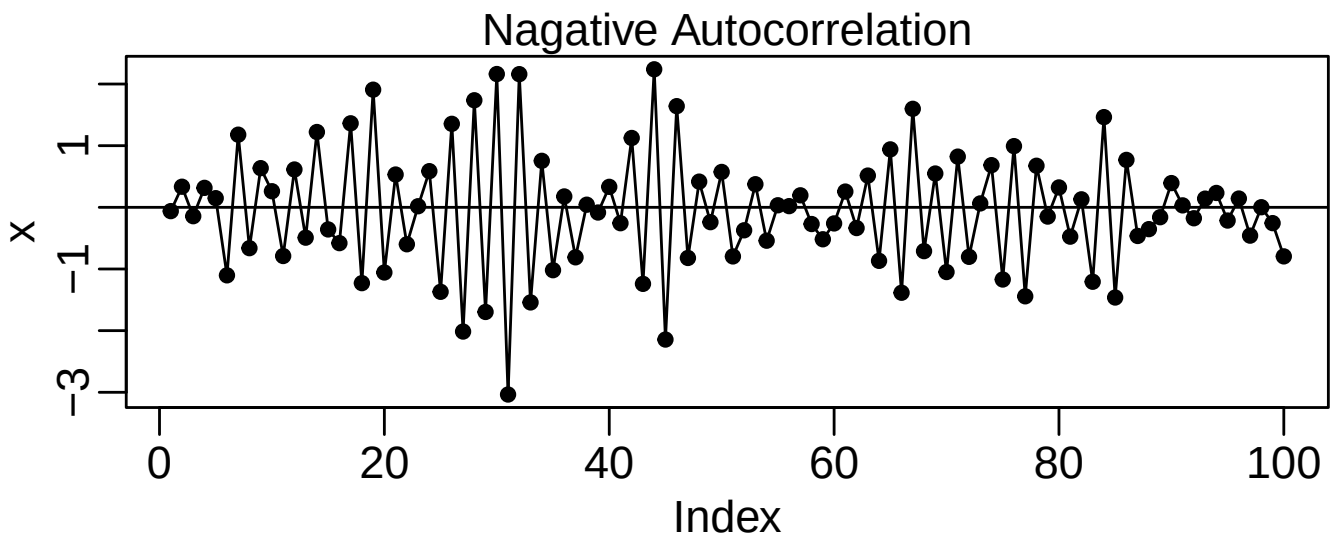
- Better keeping track of the time order observations are recorded so we can make a time-plot
- A *smooth* time-plot is a sign of *positive* serial dependence, since a smooth time plot means successive residuals are too close together



If *no* autocorrelation, the time plot has more up-and-downs.

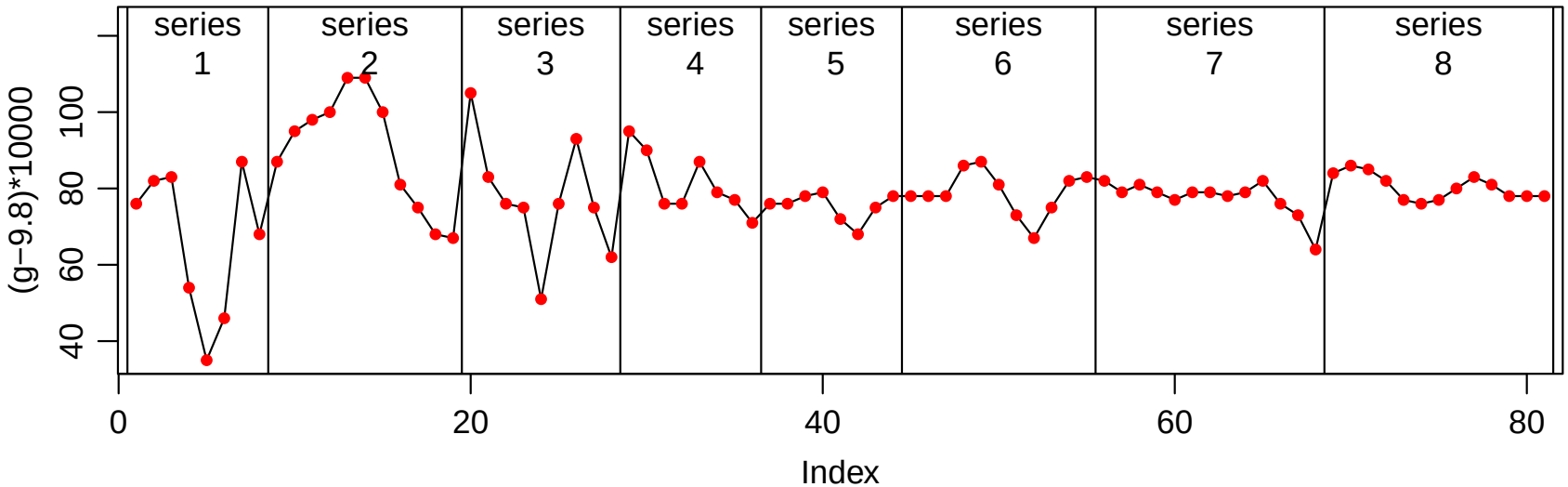


Time plot of data w/ *negative* autocorrelation tend to *alternate* regularly between positive and negative values.



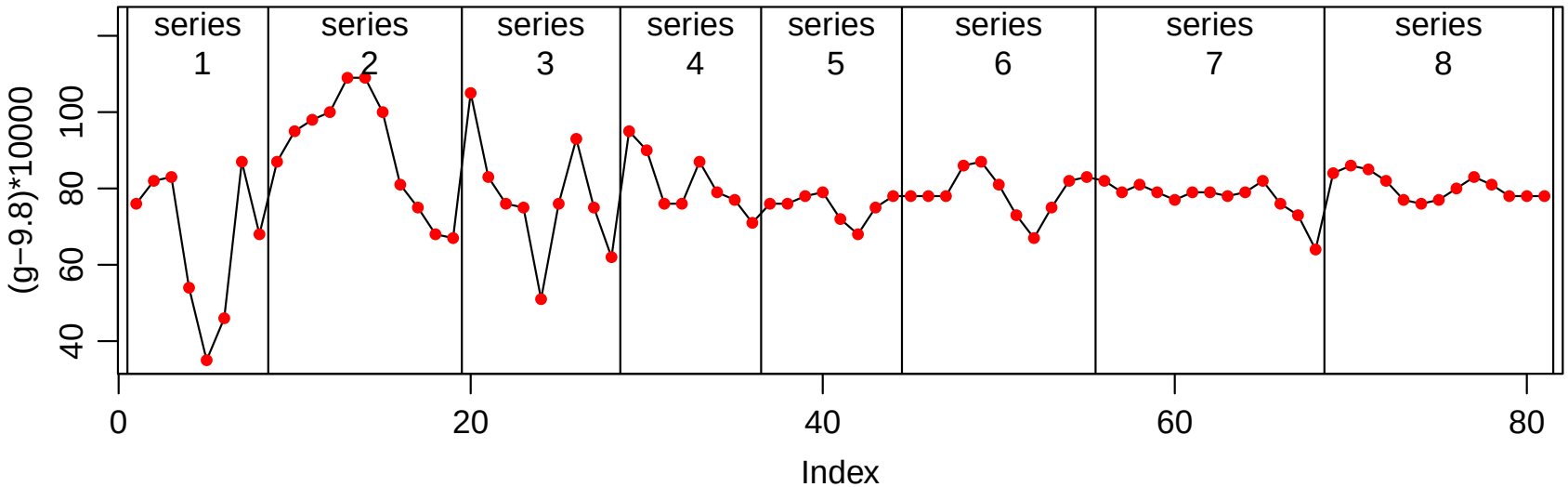
Time Plot of Gravity Data

The observations in each series are in fact given in time order taken. We can thus make a time plot.



Time Plot of Gravity Data

The observations in each series are in fact given in time order taken. We can thus make a time plot.



The time plot look *smooth* — measurements tend to stay close to the previous ones, which indicates a positive serial correlation.

Scientists in NIST might unconsciously match their results with the previous measurement, which was often regarded as the most accurate one till then.

Lag Plots — Residuals Against Lag-k Residuals

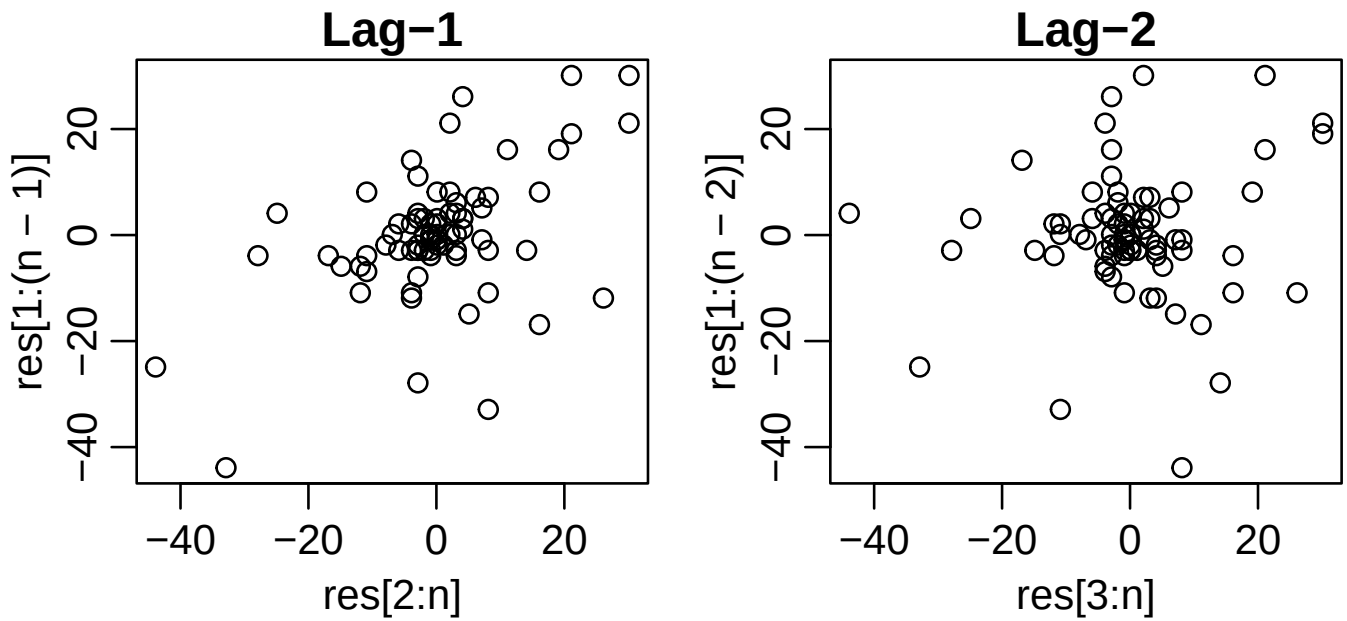
If successive residuals are correlated, we would observe a positive correlation when we plot the residuals $(e_1, \ldots, e_{n-1})$ against the next ones $(e_2, \ldots, e_n)$ (Lag 1).

- or $(e_1, \ldots, e_{n-k})$ against the lag-k residuals $(e_{1+k}, \ldots, e_n)$
- Any trend in the plot is a sign of autocorrelation.

Lag 1	Lag k
(e_1, e_2)	(e_1, e_{1+k})
(e_2, e_3)	(e_2, e_{2+k})
(e_3, e_4)	(e_3, e_{3+k})
$\vdots$	$\vdots$
(e_{n-1}, e_n)	(e_{n-k}, e_n)

Lag-1 and Lag-2 Plots of Gravity Data

```
lm0 = lm(g ~ 1)    # Null model: All series had the same mean
res = lm0$res      # Residual of null model
n = length(g)
par(mai=c(0.56,0.56,0.25,0.1), mgp=c(1.7,0.6,0))
plot(res[2:n], res[1:(n-1)], main="Lag-1")
plot(res[3:n], res[1:(n-2)], main="Lag-2")
cor(res[2:n],res[1:(n-1)])
[1] 0.5002
cor(res[3:n],res[1:(n-2)])
[1] 0.1233
```



Autocorrelation

The **lag-k autocorrelation** of the residuals $(e_1, \ldots, e_n)$ is defined as

$$\hat{\rho}_k = \frac{\sum_{t=k+1}^{n} e_t e_{t-k}}{\sum_{t=1}^{n} e_t^2}, \quad k = 1, 2, 3, \ldots,$$

which is slightly different from the "correlation" of $(e_1, \ldots, e_{n-k})$ v.s. $(e_{1+k}, \ldots, e_n)$,

$$\frac{\sum_{t=k+1}^{n} e_t e_{t-k}}{\sqrt{\sum_{t=1}^{n-k} e_t^2 \sum_{t=k+1}^{n} e_t^2}}$$

The R command `acf()` (autocorrelation function) in R can calculate lag-k autocorrelation.

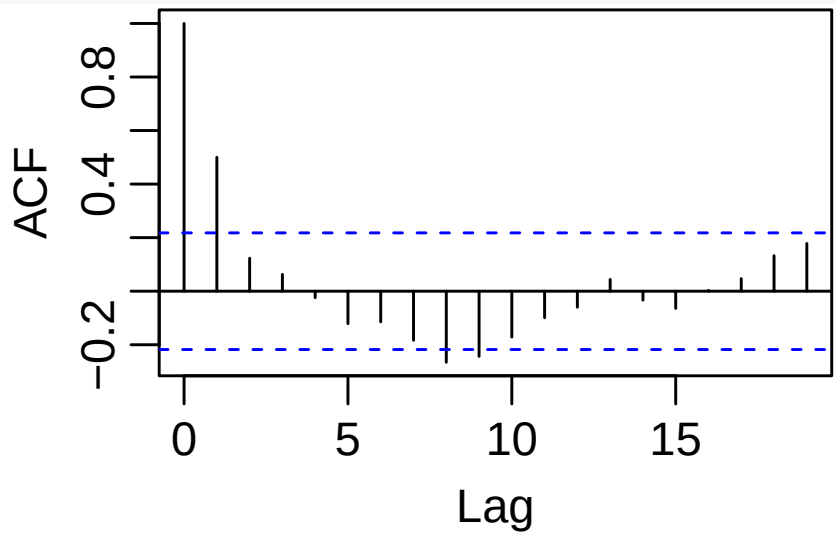
```
acf(res, lag.max =5, plot=FALSE)

Autocorrelations of series 'res', by lag

    0     1     2     3     4     5
1.000 0.500 0.123 0.063 -0.024 -0.121
```

Autocorrelation Function and the Plot

The `acf()` command can produce the **autocorrelation plot**, which is a plot of lag-k autocorrelations against k.

E.g., here is the acf plot of the residuals of gravity data.

```
acf(res)
```



The horizontal dash lines marks the levels autocorrelations to be significantly different from 0.

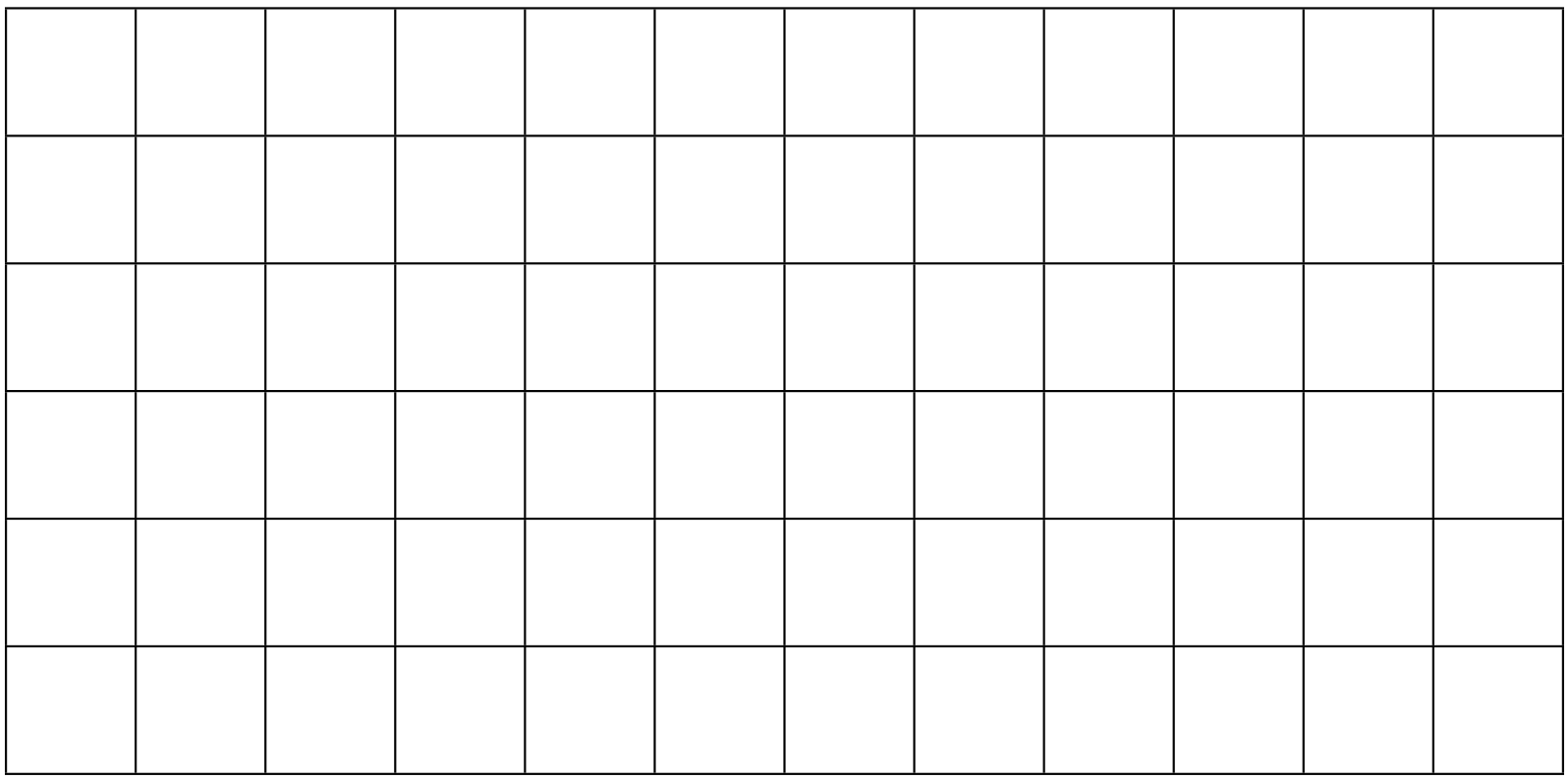
Effect of Dependent Errors

For the gravity data, as adjacent observations are positively correlated,

$\Rightarrow$ observations within each group/series tend to be too close to each other,
$\Rightarrow$ small within group variability
$\Rightarrow$ MSE underestimates the actual variance σ^2 of noise
$\Rightarrow F = \frac{\text{MS}_{trt}}{\text{MSE}}$ would be too large $\Rightarrow$ more likely to reject H_0

Spatial Dependence

Spatial dependence can arise when the experimental units are arranged in space, like plants in a farm. Spatial dependence occurs when units that are closer together are more similar than units farther apart.



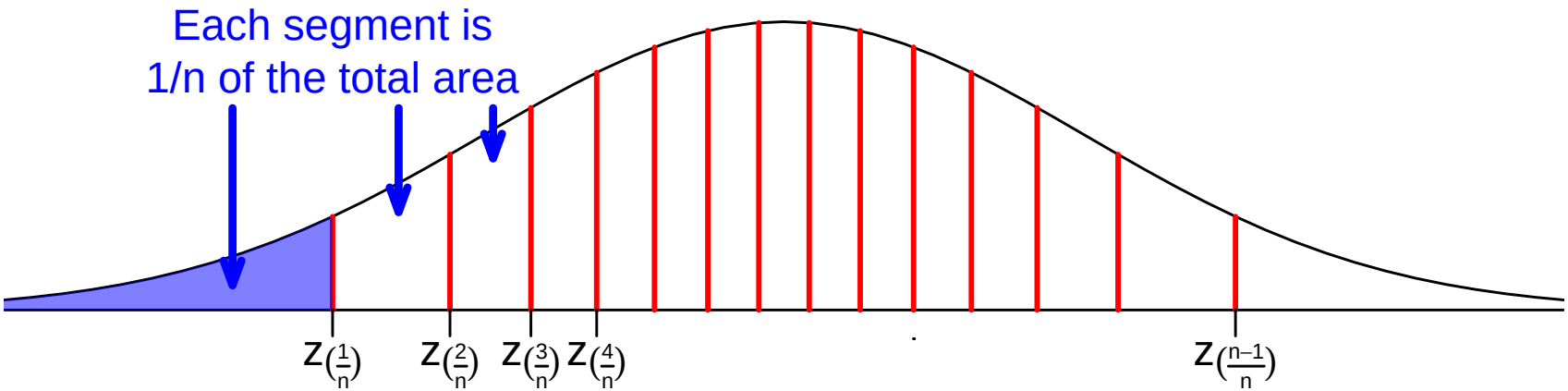
Checking Normality

How to Check the Normality Assumption

- Histogram of the residuals: if normal, should be bell-shaped
  - Pros: simple, easy to understand
  - Cons: for a small sample, histogram may not be bell-shaped even though the sample is from a normal distribution
- *Normal probability plot* of the residuals
  - aka. *normal QQ plot*,
    QQ stands for "quantile-quantile"
  - best tool to assess normality
  - See next slide for details

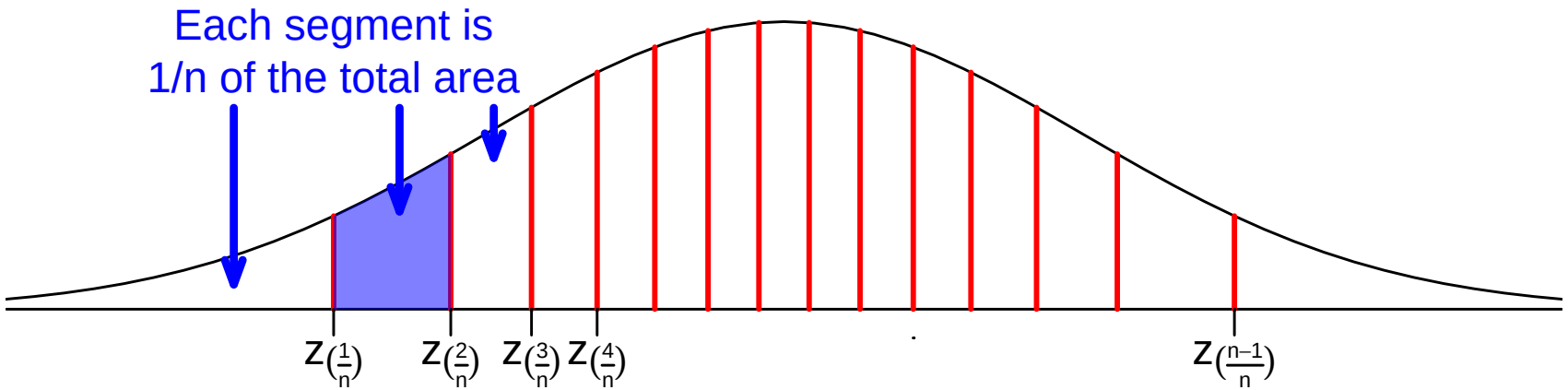
Ideas Behind the Normal Probability Plot (1)

- Data: $y_1, y_2, \ldots, y_n$
- Sorted Data: $y_{(1)} \leq y_{(2)} \leq \ldots \leq y_{(n)}$, call the **Sample Quantiles**
- **Theoretical Quantiles** of the $\mathrm{N}(0,1)$: $z_{(\frac{1}{n})}, z_{(\frac{2}{n})}, \ldots, z_{(\frac{n-1}{n})}$, where, $z_{(\frac{k}{n})}$ is a value such that $\mathrm{P}(Z \leq z_{(\frac{k}{n})}) = k/n$ for $Z \sim N(0,1)$.



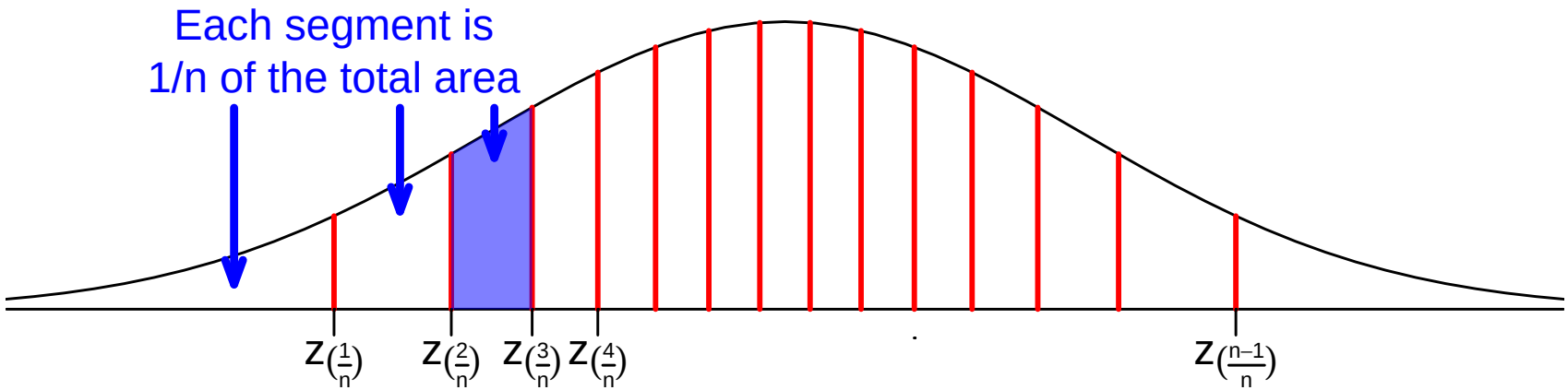
Ideas Behind the Normal Probability Plot (1)

- Data: $y_1, y_2, \ldots, y_n$
- Sorted Data: $y_{(1)} \leq y_{(2)} \leq \ldots \leq y_{(n)}$, call the **Sample Quantiles**
- **Theoretical Quantiles** of the $\mathrm{N}(0,1)$: $z_{(\frac{1}{n})}, z_{(\frac{2}{n})}, \ldots, z_{(\frac{n-1}{n})}$, where, $z_{(\frac{k}{n})}$ is a value such that $\mathrm{P}(Z \leq z_{(\frac{k}{n})}) = k/n$ for $Z \sim N(0,1)$.



Ideas Behind the Normal Probability Plot (1)

- Data: $y_1, y_2, \ldots, y_n$
- Sorted Data: $y_{(1)} \leq y_{(2)} \leq \ldots \leq y_{(n)}$, call the **Sample Quantiles**
- **Theoretical Quantiles** of the $\mathrm{N}(0,1)$: $z_{(\frac{1}{n})}$, $z_{(\frac{2}{n})}$, $\ldots$, $z_{(\frac{n-1}{n})}$, where, $z_{(\frac{k}{n})}$ is a value such that $\mathrm{P}(Z \leq z_{(\frac{k}{n})}) = k/n$ for $Z \sim N(0,1)$.



Ideas Behind the Normal Probability Plot (2)

- If $Y \sim N(\mu, \sigma^2)$, then

$$P(Y \leq \mu + \sigma z_{(\frac{k}{n})}) = P\Big(\underbrace{\frac{Y-\mu}{\sigma}}_{\sim N(0,1)} \leq z_{(\frac{k}{n})}\Big) = \frac{k}{n}$$

  We expected k/n of the observations to be $\leq \mu + \sigma z_{(\frac{k}{n})}$
- We observe k/n of the observations are $\leq y_{(k)}$.
- If the data are indeed $N(\mu, \sigma^2)$, we expect

$$y_{(k)} \approx \mu + \sigma z_{(\frac{k}{n})}$$

- If one plots the Sample Quantiles $y_{(1)} \leq y_{(2)} \leq \ldots \leq y_{(n)}$ against the Theoretical Quantiles $z_{(\frac{1}{n})}$, $z_{(\frac{2}{n})}$, $\ldots$, $z_{(\frac{n-1}{n})}$, the points would fall on the straight line

$$y = \mu + \sigma z.$$

  if the data follow $N(\mu, \sigma^2)$

A Technical Remark

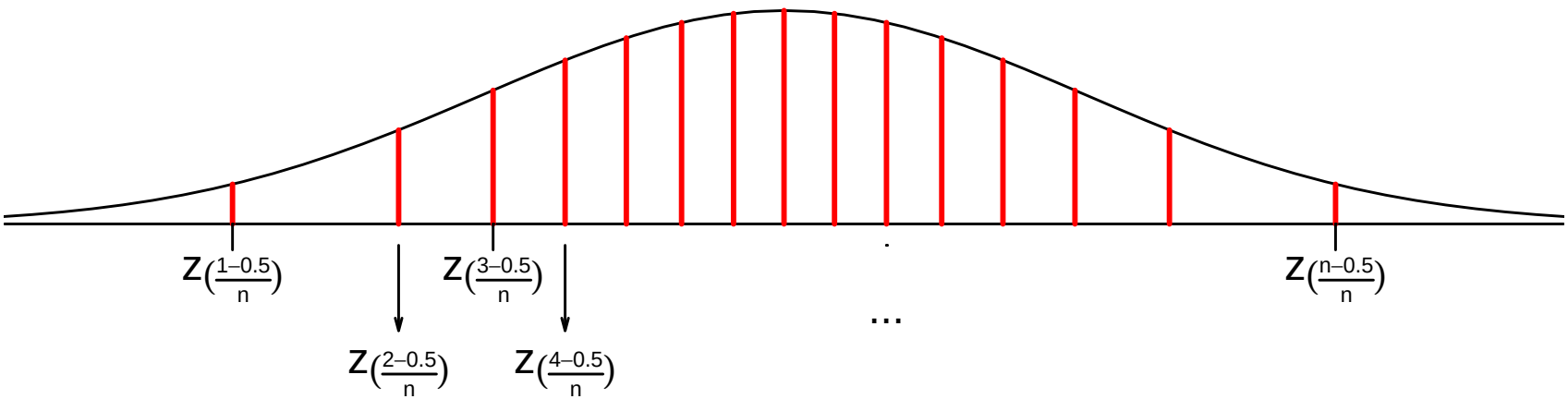
R actually uses the Theoretical Quantiles:

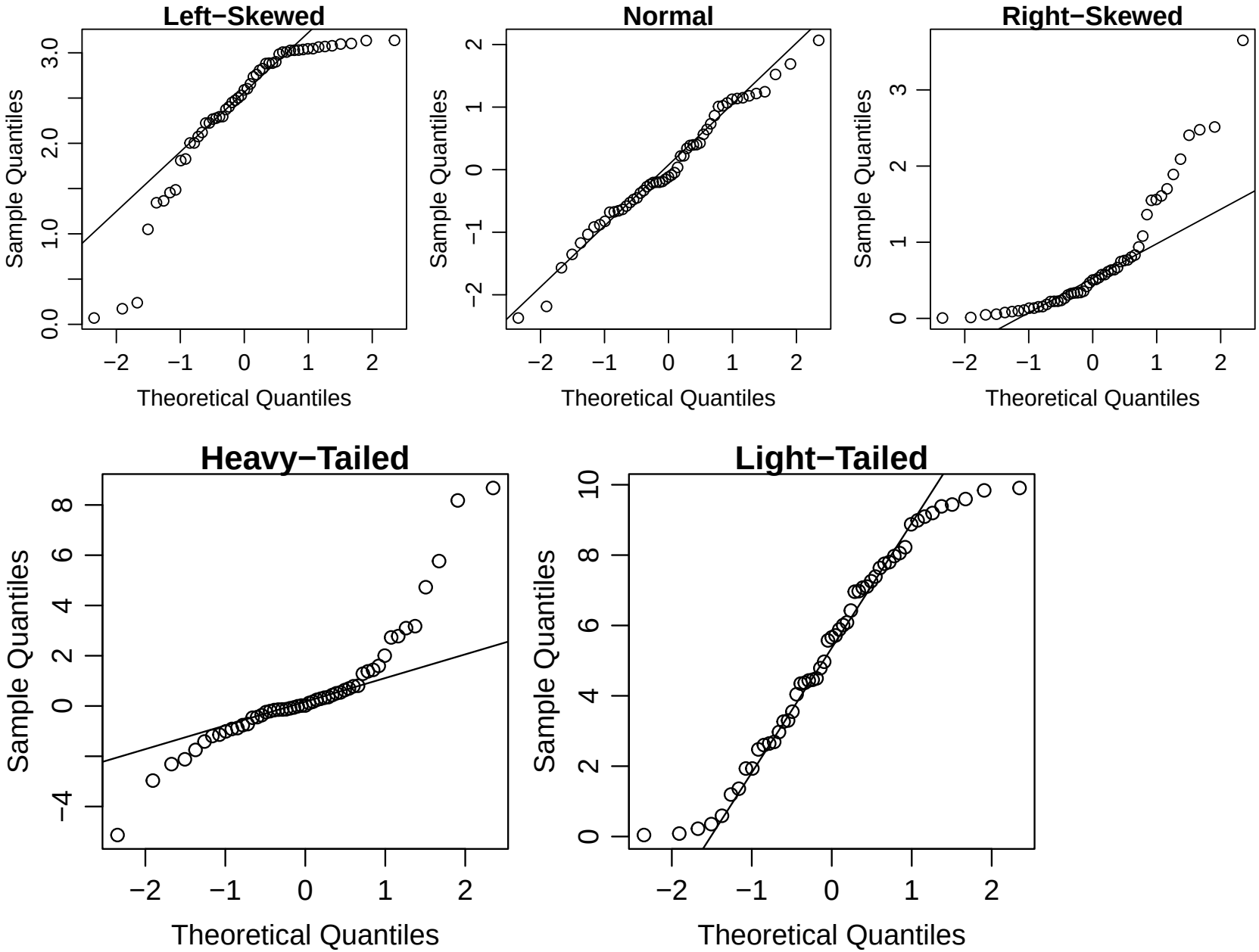
$$z_{(\frac{1-0.5}{n})}, \quad z_{(\frac{2-0.5}{n})}, \quad z_{(\frac{3-0.5}{n})}, \quad \ldots, \quad z_{(\frac{n-0.5}{n})}$$

instead of

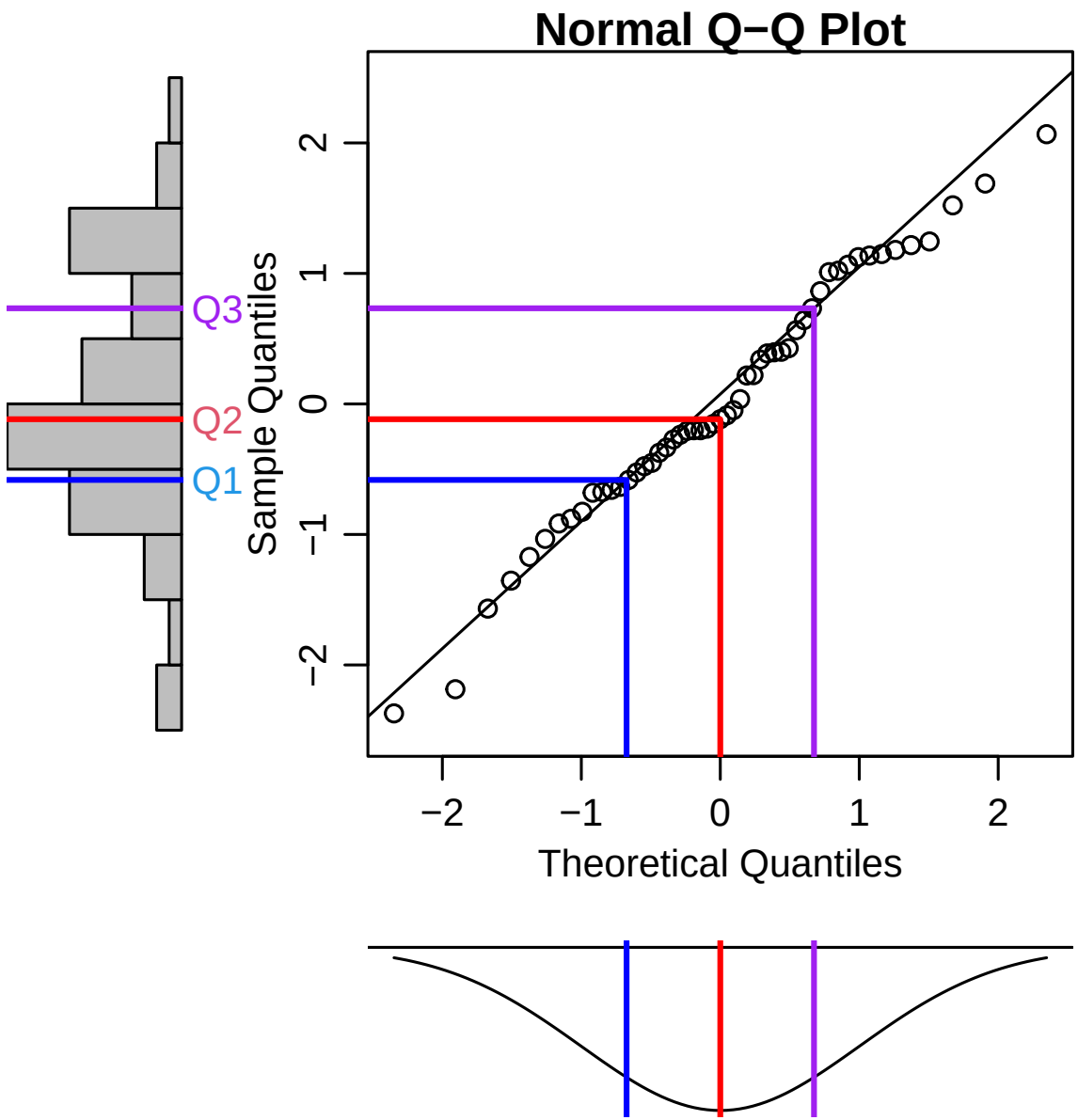
$$z_{(\frac{1}{n})}, \quad z_{(\frac{2}{n})}, \ldots, z_{(\frac{n-1}{n})}, z_{(\frac{n}{n})},$$

since $z_{(n/n)} = \infty$.

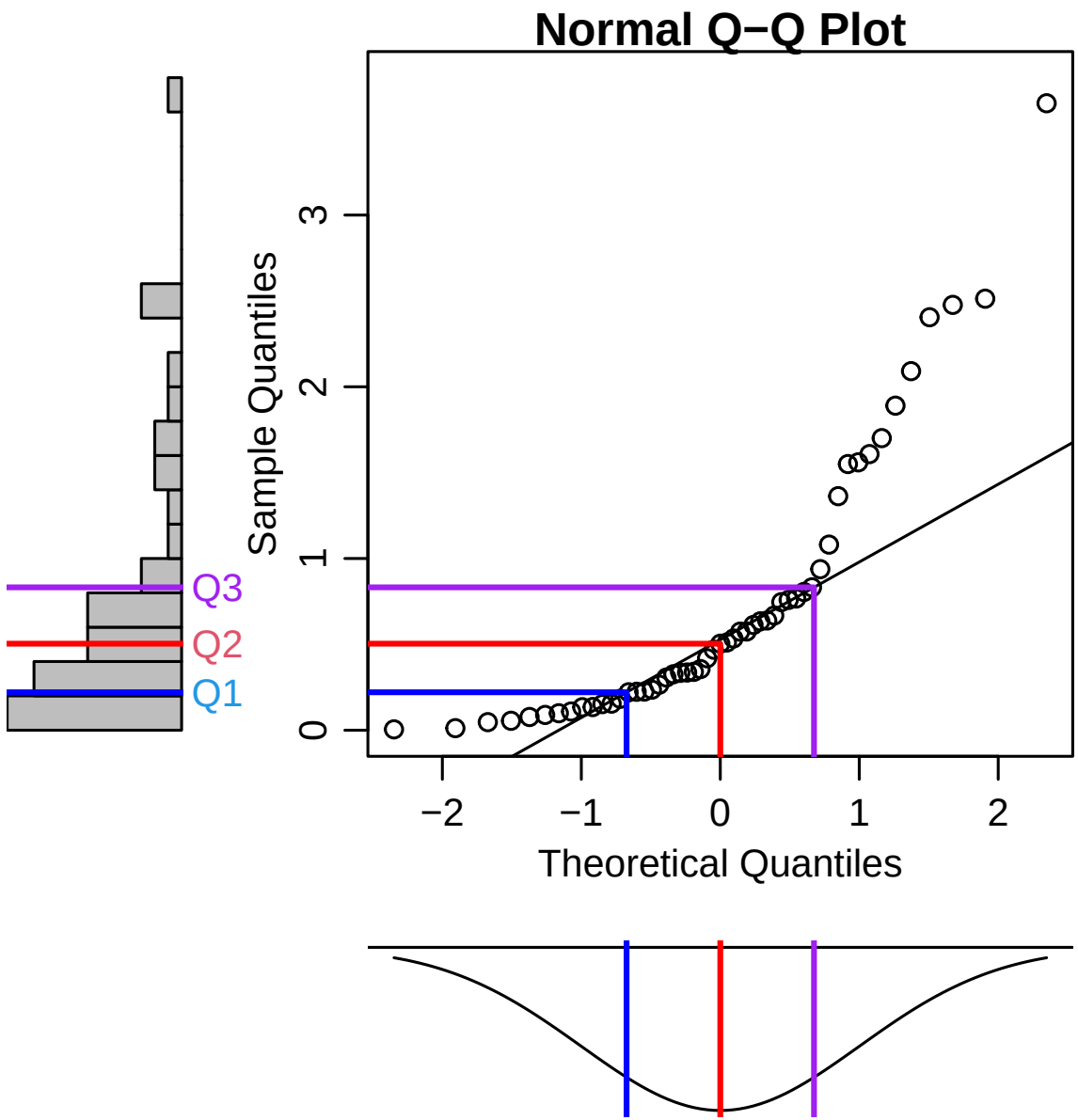



Left−Skewed
Normal
Right−Skewed
Heavy−Tailed
Light−Tailed
Sample Quantiles
Theoretical Quantiles

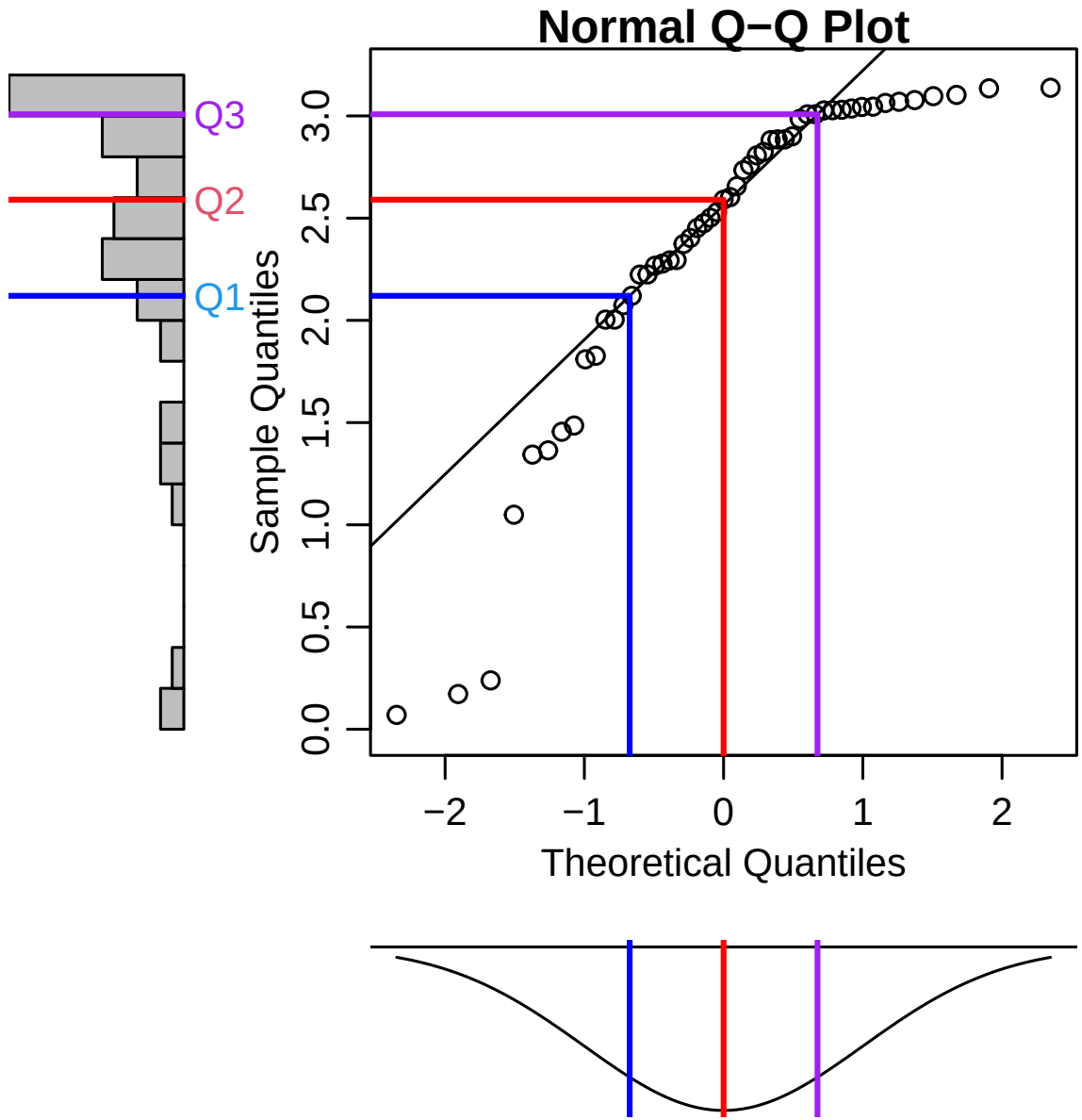
Normal QQ Plot — Normal Data



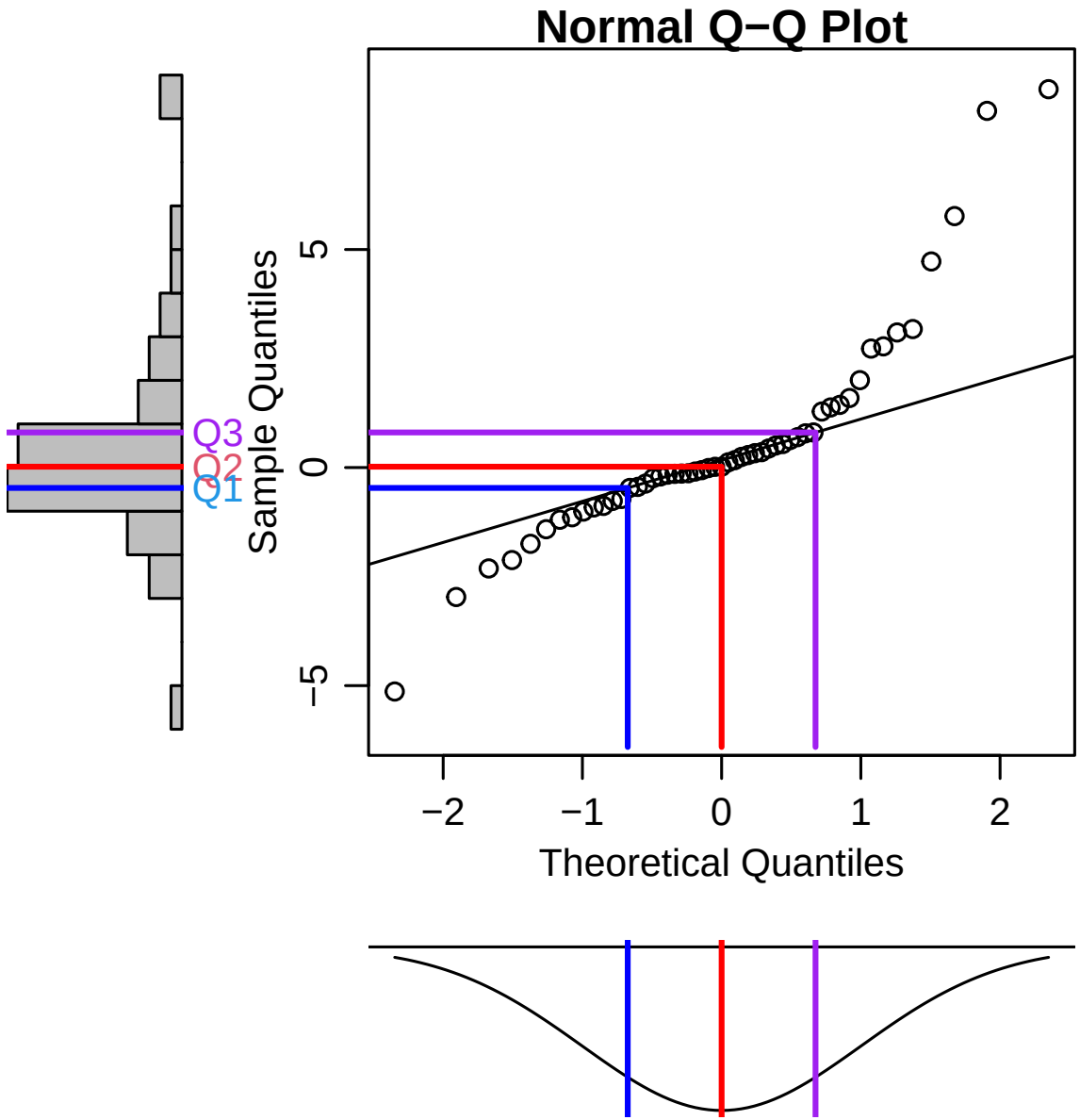
Normal QQ Plot — Right-Skewed Data



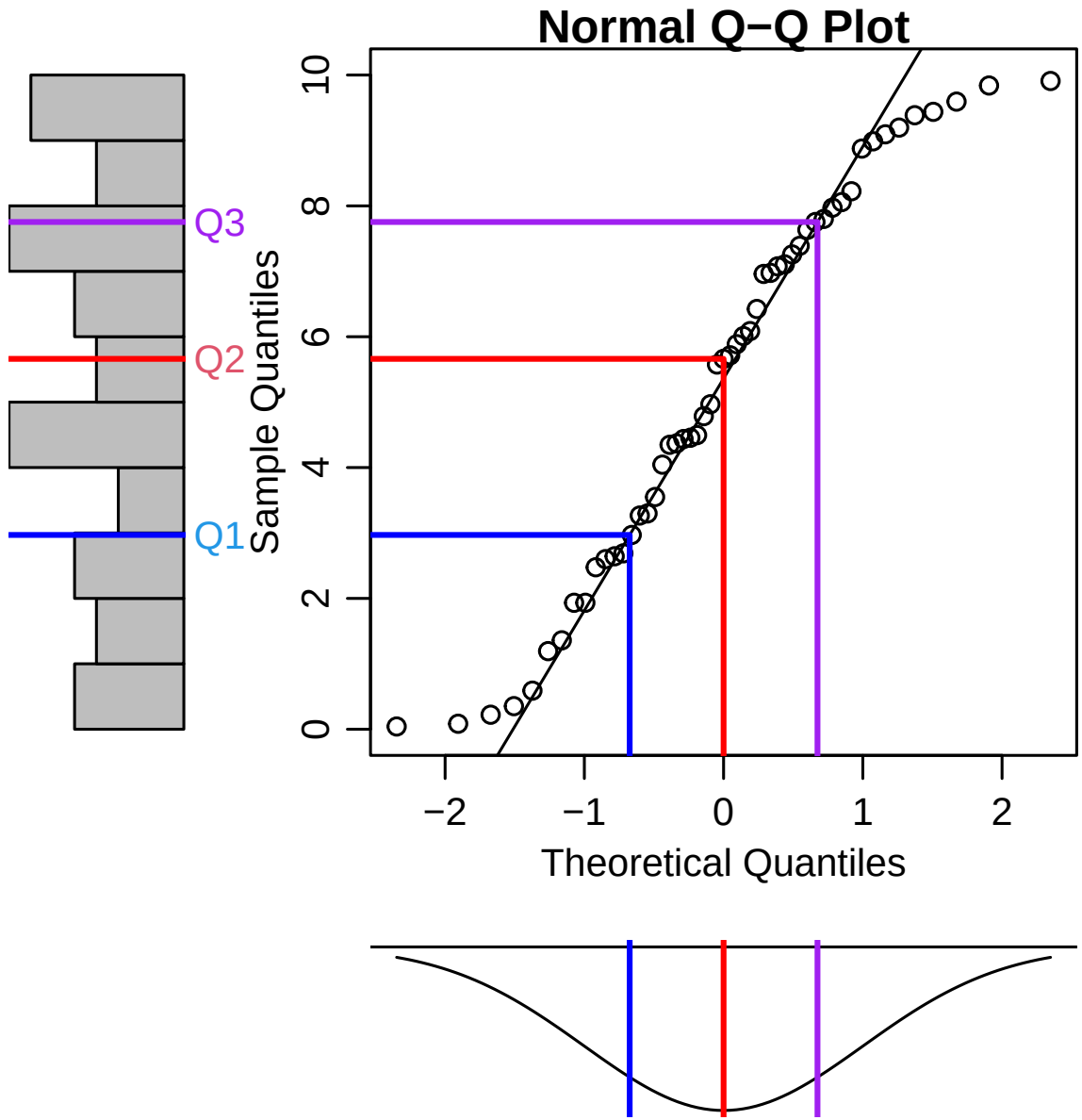
Normal QQ Plot — Left-Skewed Data



Normal QQ Plot — Heavy-Tailed Data



Normal QQ Plot — Light-Tailed Data

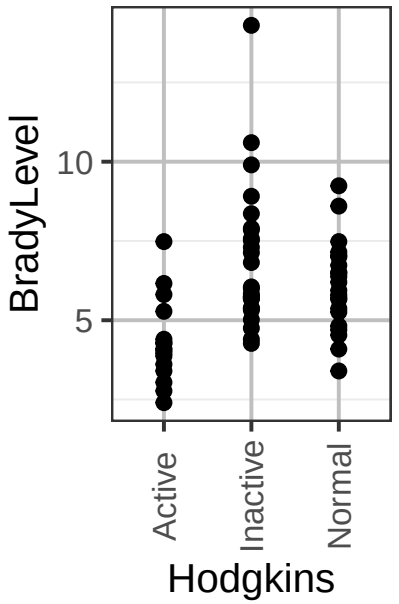


Recall Hodgkin's Disease Data

```
hodgkins = read.table(
  "http://www.stat.uchicago.edu/~yibi/s222/Hodgkins.txt", h=T)
```

Plasma bradykininogen levels in 3 types of subjects

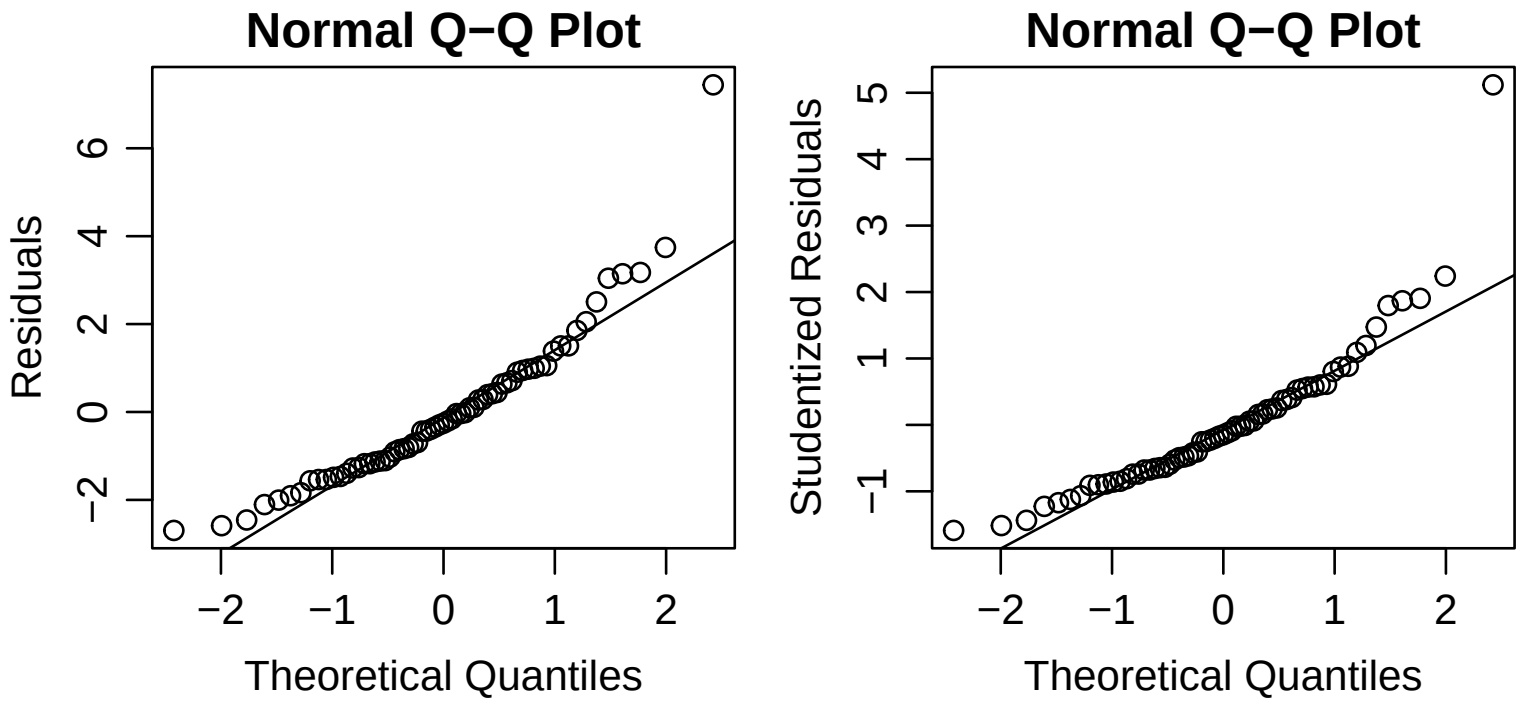
- normal,
- patients with active Hodgkin's disease,
- patients with inactive Hodgkin's disease.




```
brady1 = lm(BradyLevel ~ Hodgkins, data=hodgkins)
anova(brady1)
Analysis of Variance Table

Response: BradyLevel
           Df Sum Sq Mean Sq F value Pr(>F)
Hodgkins    2   65.9    32.9    10.7 0.0001
Residuals 62  191.4     3.1
```

Normal QQ Plot for The Hodgkin Data

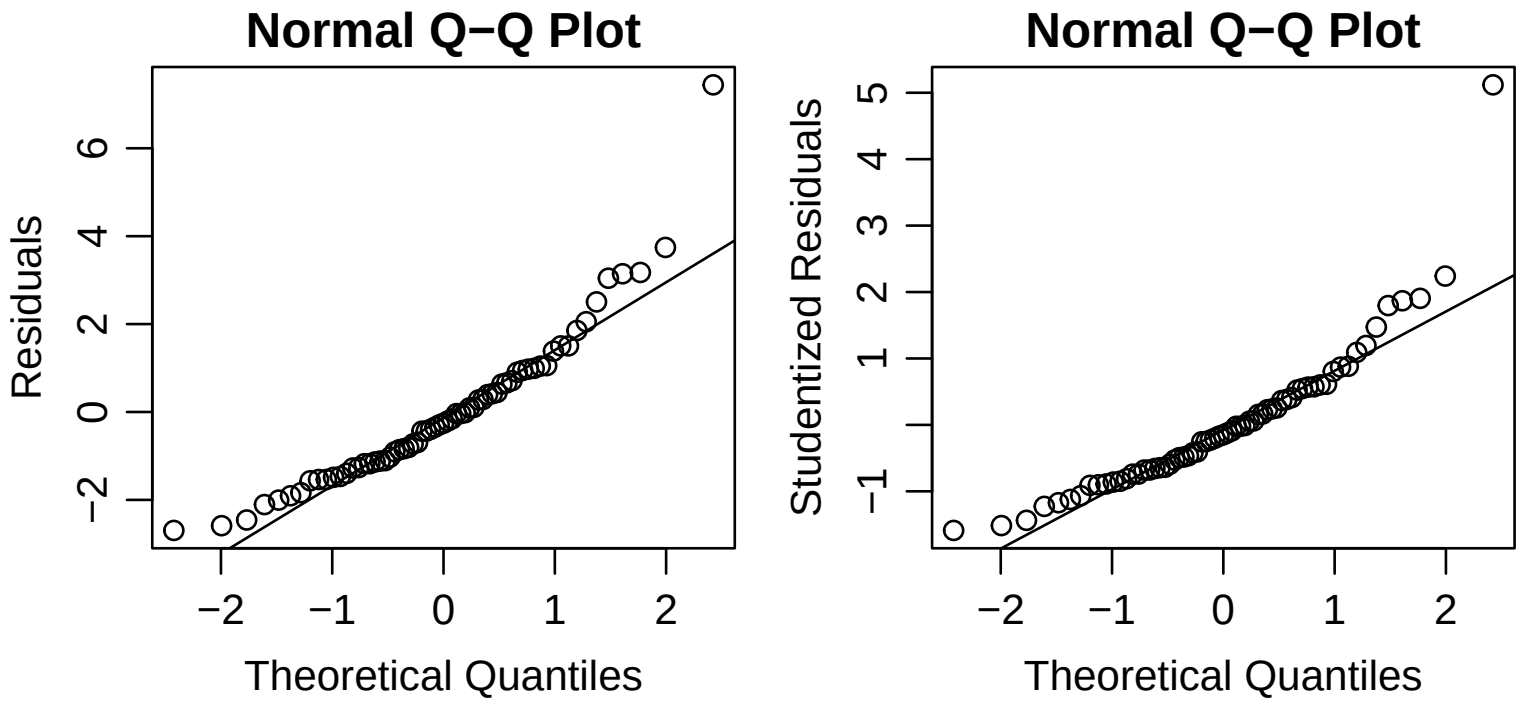
```
qqnorm(brady1$res, ylab="Residuals")
qqline(brady1$res)
qqnorm(rstudent(brady1), ylab="Studentized Residuals")
qqline(rstudent(brady1))
```



Does the distribution of the residuals look normal?

Normal QQ Plot for The Hodgkin Data

```
qqnorm(brady1$res, ylab="Residuals")
qqline(brady1$res)
qqnorm(rstudent(brady1), ylab="Studentized Residuals")
qqline(rstudent(brady1))
```



Does the distribution of the residuals look normal?

Somewhat right-skewed, a potential outlier with $t_{ij} > 5$.

Remedies for Non-Normality

Remedies for Non-Normality

Skewness can often be ameliorated by **transforming the response** — often a power transformation.

$$f_\lambda(y) = \begin{cases} y^\lambda, & \text{if } \lambda \neq 0 \\ \log(y), & \text{if } \lambda = 0. \end{cases}$$

- If **right-skewed**, try taking square root, logarithm, or other powers $\lambda < 1$

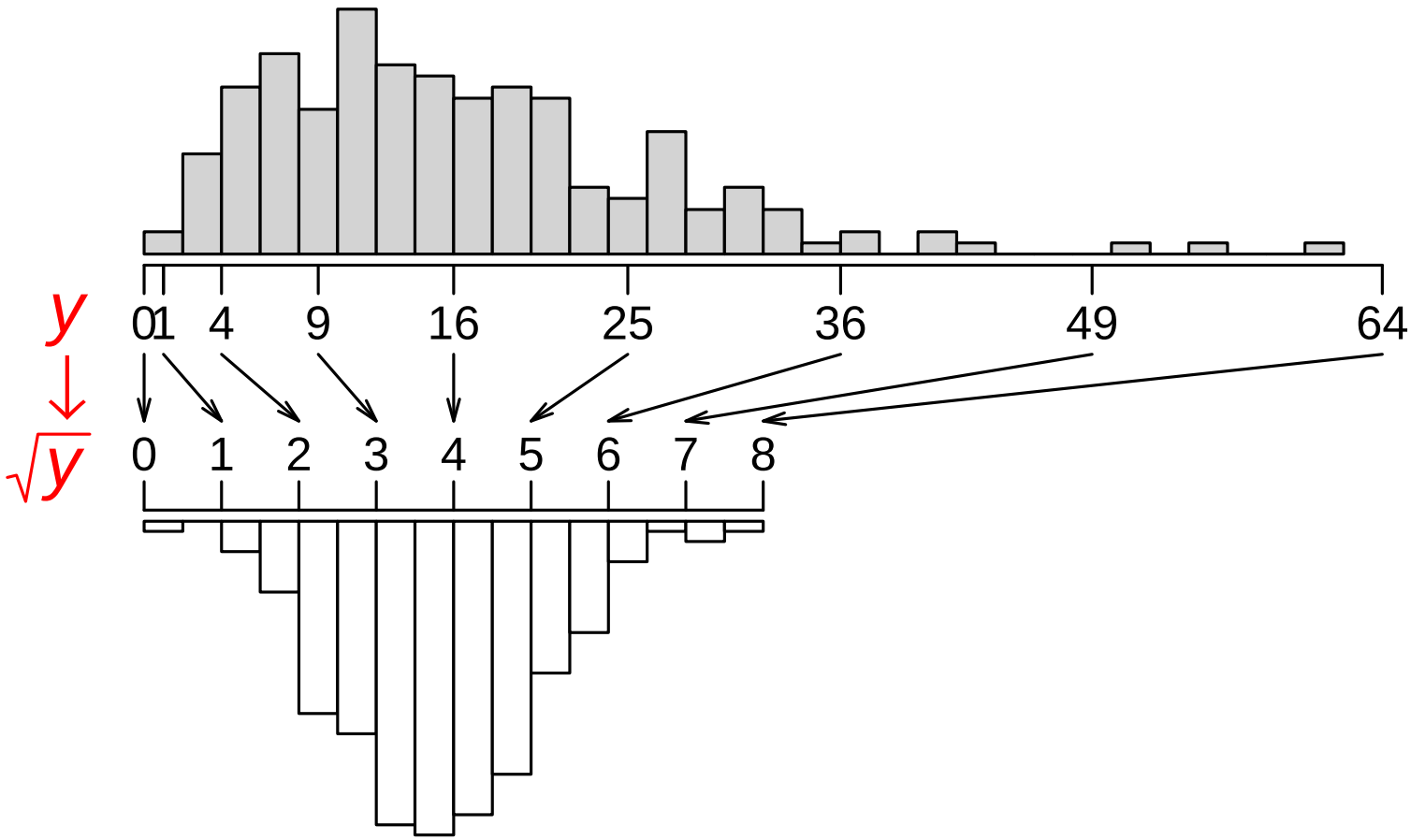
$$y \longrightarrow 1/y, \log(y), \sqrt{y}, \text{or } y^\lambda \text{ with } \lambda < 1$$

- If **left-skewed**, try squaring, cubing, or other powers $\lambda > 1$

$$y \longrightarrow y^2, y^3, \text{or } y^\lambda \text{ with } \lambda > 1$$

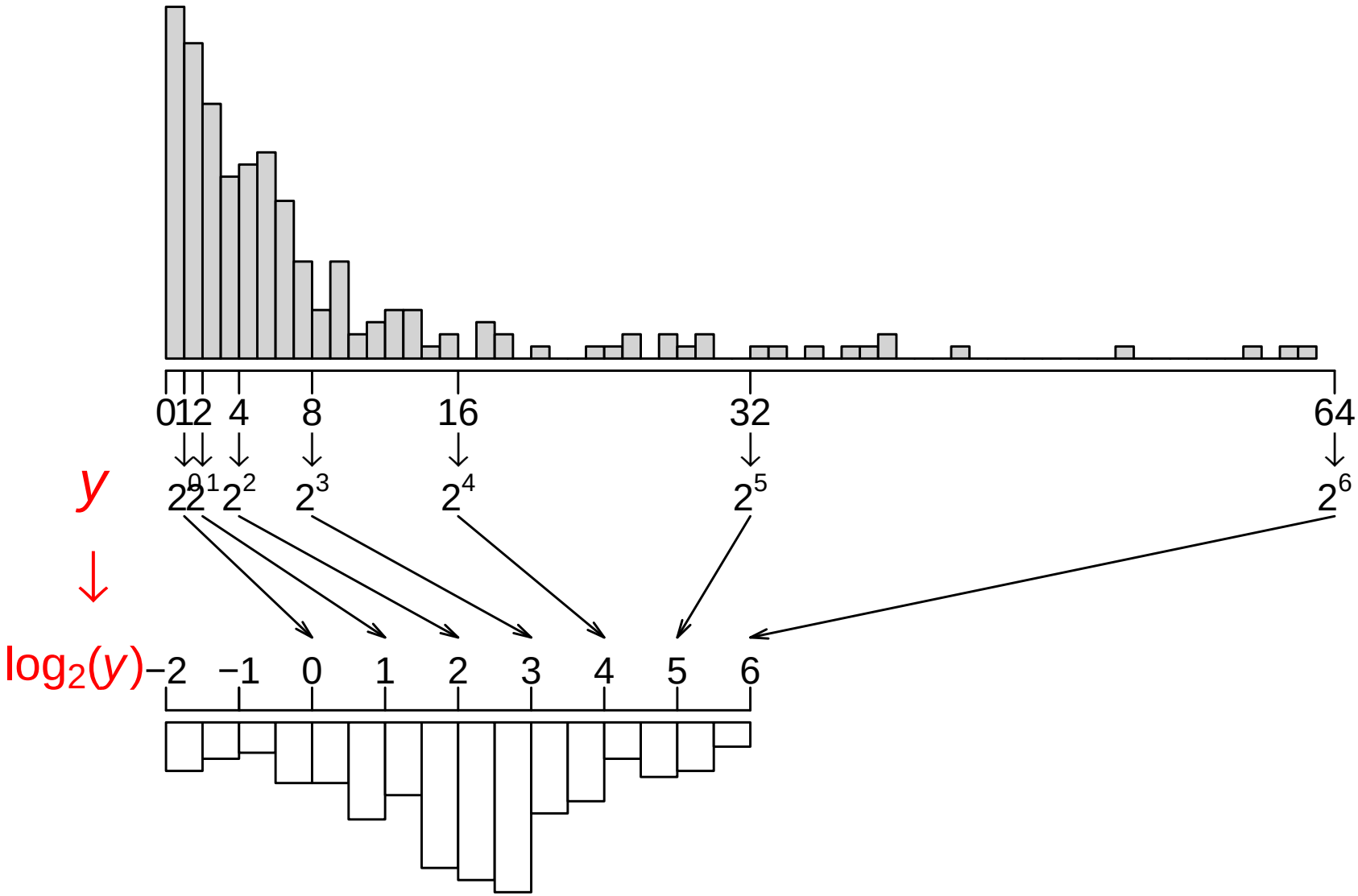
Square-Root Transformation Can Reduce Right-Skewness

The square-root transformation can shorten the upper tail and extend the lower tail, of a distribution and hence can reduce right-skewness.



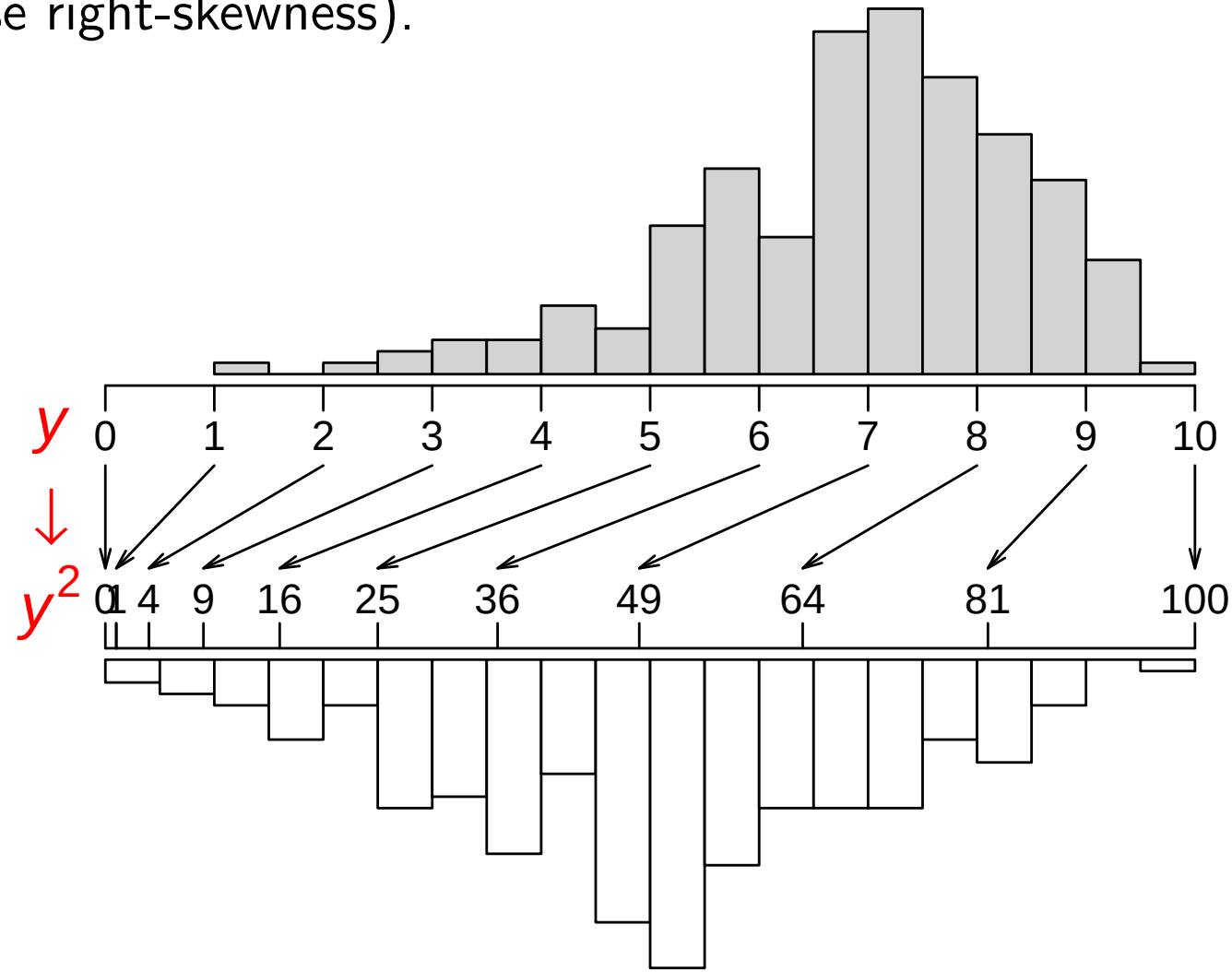
Log Transformation Reduces Right-Skewness Even More!

Logarithm can shorten the upper tail and extend the lower tail even more



Square Transformation Can Reduce Left-Skewness

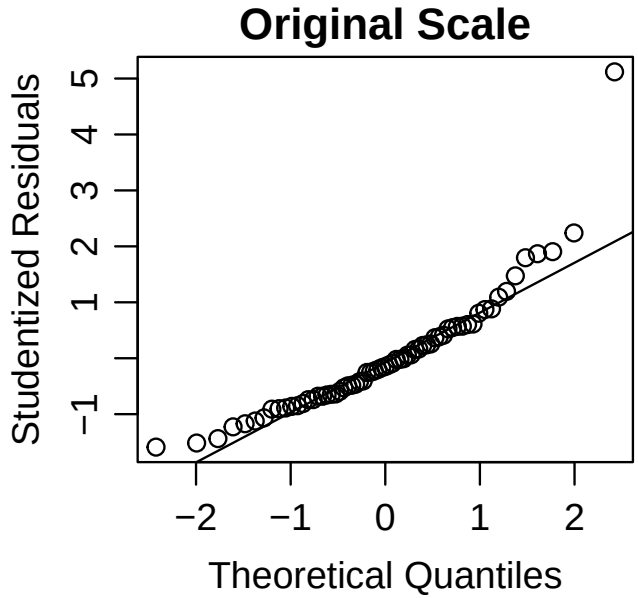
The square transformation ($y \to y^2$) can extend the upper tail and shorten the lower tail, and hence can reduces left-skewness (and increase right-skewness).

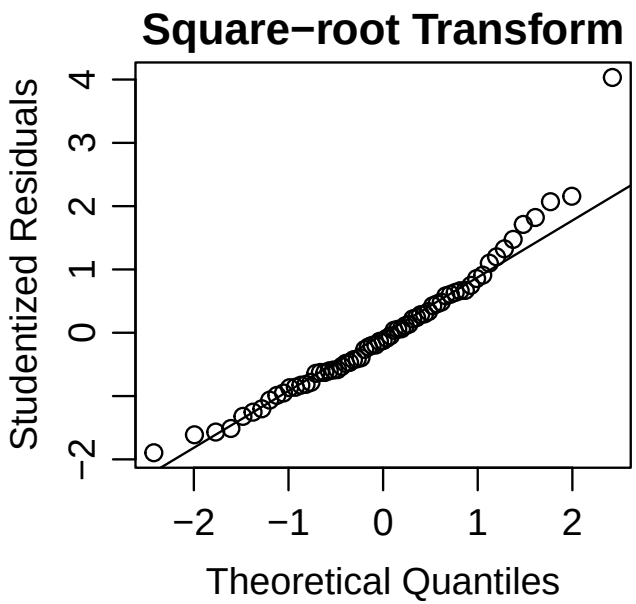


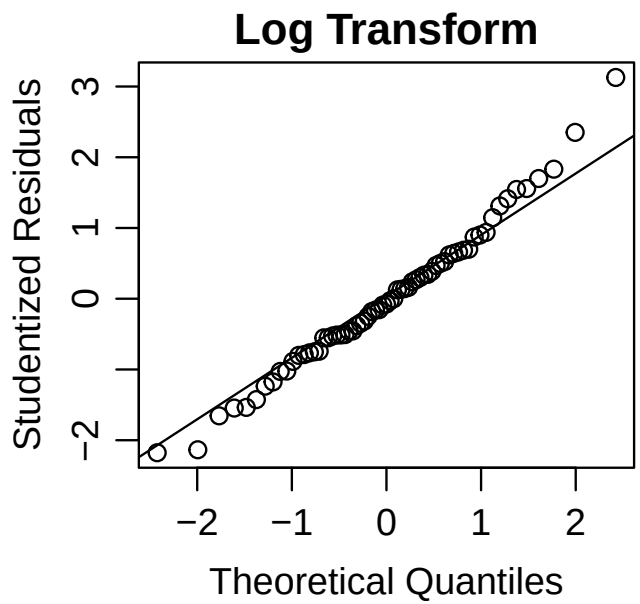
Example: Hodgkin's Disease – QQ Plots

Log-transformation makes residuals less R-skewed, and the outlier less extreme.

The square-root transformation also reduces R-skewness but not as much as the log transformation.







R-Codes for Making the Plots on the Previous Slide

```
brady1 = lm(BradyLevel ~ Hodgkins, data=hodgkins)
brady2 = lm(sqrt(BradyLevel) ~ Hodgkins, data=hodgkins)
brady3 = lm(log(BradyLevel) ~ Hodgkins, data=hodgkins)

qqnorm(rstudent(brady1), main="Original Scale")
qqline(rstudent(brady1))

qqnorm(rstudent(brady2), main="Square-root Transform")
qqline(rstudent(brady2))

qqnorm(rstudent(brady3), main="Log Transform")
qqline(rstudent(brady3))
```

Box-Cox Method

Box-Cox method is an automatic procedure to select the "best" power λ that make the residuals of the model

$$y_{ij}^{\lambda} = \mu_i + \epsilon_{ij}$$

closest to normal.

Box-Cox Method

Box-Cox method is an automatic procedure to select the "best" power λ that make the residuals of the model

$$y_{ij}^{\lambda} = \mu_i + \epsilon_{ij}$$

closest to normal.

- We usually round the optimal λ to a convenient power like

$$-1,\ -\frac{1}{2},\ -\frac{1}{3},\ 0,\ \frac{1}{3},\ \frac{1}{2},\ 1,\ 2, \ldots$$

since the practical difference of $y^{0.5827}$ and $y^{0.5}$ is usually small, but the square-root transformation is much easier to interpret.

Box-Cox Method

Box-Cox method is an automatic procedure to select the "best" power λ that make the residuals of the model

$$y_{ij}^{\lambda} = \mu_i + \epsilon_{ij}$$

closest to normal.

- We usually round the optimal λ to a convenient power like

$$-1,\ -\frac{1}{2},\ -\frac{1}{3},\ 0,\ \frac{1}{3},\ \frac{1}{2},\ 1,\ 2, \ldots$$

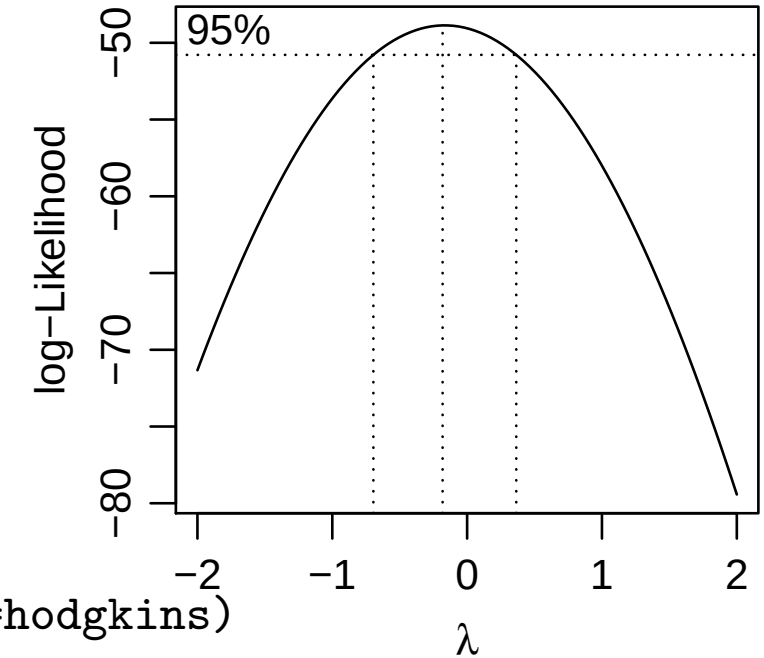
  since the practical difference of $y^{0.5827}$ and $y^{0.5}$ is usually small, but the square-root transformation is much easier to interpret.
- A confidence interval for the optimal λ can also be obtained.

We usually select a convenient power λ^* in this C.I.

Example: Hodgkin's Disease – Box-Cox

In R, one must first load the `MASS` library to use `boxcox()`.

The argument of the `boxcox()` function can be a model formula, an `lm()` model.

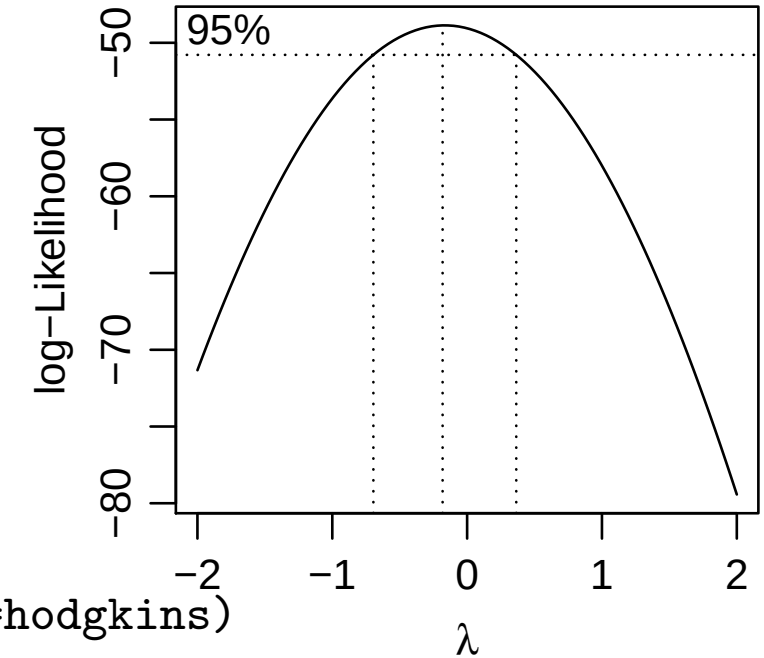



```
library(MASS)
boxcox(brady1)
boxcox(BradyLevel ~ Hodgkins, data=hodgkins)
```

Example: Hodgkin's Disease – Box-Cox

In R, one must first load the `MASS` library to use `boxcox()`.

The argument of the `boxcox()` function can be a model formula, an `lm()` model.



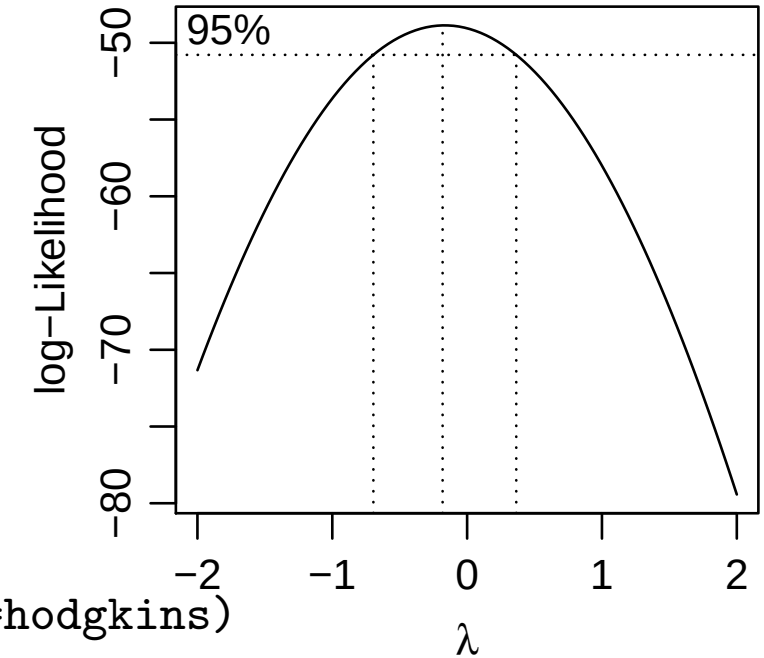

```
library(MASS)
boxcox(brady1)
boxcox(BradyLevel ~ Hodgkins, data=hodgkins)
```

The middle dash line marks the optimal λ, the right and left dash line mark the 95% C.I. for the optimal λ.

Example: Hodgkin's Disease – Box-Cox

In R, one must first load the `MASS` library to use `boxcox()`.

The argument of the `boxcox()` function can be a model formula, an `lm()` model.




```
library(MASS)
boxcox(brady1)
boxcox(BradyLevel ~ Hodgkins, data=hodgkins)
```

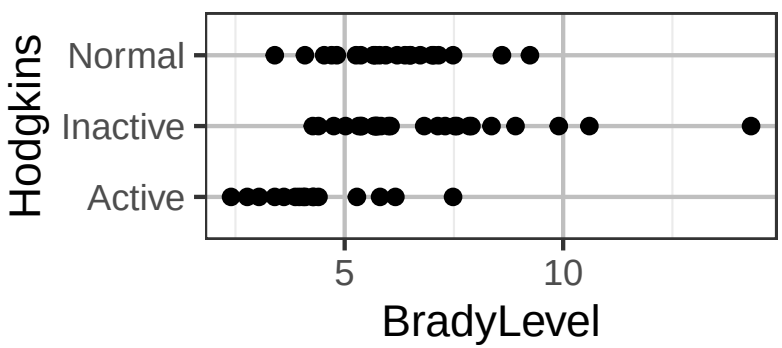
The middle dash line marks the optimal λ, the right and left dash line mark the 95% C.I. for the optimal λ.

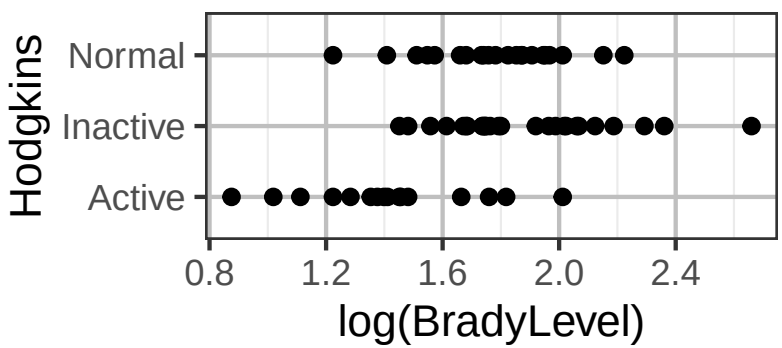
For the plot, we see the optimal λ is around -0.2, and the 95% C.I. contains 0. For simplicity, we use the log-transformed `BradyLevel` as our response.

Example: Hodgkin's Disease

From the 2 ANOVA tables below, observe the differences of the 3 groups of patients become more significant after a log transformation, since the outlier become less extreme and do not inflate the MSE as much.

```
Response: BradyLevel
           Df  Sum Sq Mean Sq F value    Pr(>F)
Hodgkins    2  65.893  32.946   10.67 0.0001042 ***
Residuals  62 191.449   3.088

Response: log(BradyLevel)
           Df Sum Sq Mean Sq F value    Pr(>F)
Hodgkins    2 2.2526 1.12631  15.436 3.628e-06 ***
Residuals  62 4.5238 0.07297
```





Log-Scale is Commonly Used for Concentrations

In fact, for measurements of concentration, the log scale is more commonly used than the original scale. For example,

- The concentrations of 10.1 and of 10.001 are nearly indistinguishable
- However, there is a huge difference between concentration of 0.1 and 0.001 since 0.1 is 100 times as high as 0.001.
- In the original scale, (10.1, 10.001) and (0.1, 0.001) differ by the same amount, 0.099.
- In log scale,

$$\log_{10} 0.1 - \log_{10} 0.001 = 2 \quad \text{far greater than}$$
$$\log_{10} 10.1 - \log_{10} 10.001 \approx 0.0043$$