Chapter 7. Chi Squared Tests.

Example 7.A (Weldon's Dice). In 1894 the biologist Frank Weldon sent to several British statisticians the results of a large number of tosses of a set of dice. Weldon and his assistants had tossed a set of 12 ordinary six-sided dice and counted the number of the 12 exhibiting an up-face of 5 or 6. They had repeated this process 26,306 times; the results are shown in column two of Table 7.1.

Table 7.1. Weldon's Dice Data. The results derived from 26,306 rolls of a set of 12 ordinary dice, compared to the counts expected under the hypothesis that the dice are fair.

No. of Dice X showing 5 or 6	Observed	Theory	Difference
0	185	203	–18
1	1149	1217	–68
2	3265	3345	–80
3	5475	5576	–101
4	6114	6273	–159
5	5194	5018	176
6	3067	2927	140
7	1331	1254	77
8	403	392	11
9	105	87	18
10	14	13	1
11	4	1	3
12	0	0	0
Total	26,306	26,306	0

The chance that a fair die will show 5 or 6 is 1/3; if 12 dice are tossed independently of one another, the number X of 5's or 6's will have a Binomial (12, 1/3) distribution. The "Theory" column gives the expected counts under this hypothesis. For example, for X = 2 we have $26{,}306 \times \binom{12}{2} \left(\tfrac{1}{3}\right)^2 \left(\tfrac{2}{3}\right)^{10} = 3{,}345.366 \cong 3{,}345$. Weldon thought the agreement was acceptable; one of his correspondents, Karl Pearson, disagreed. Pearson would reject the "fair dice" hypothesis. This testing problem is not a simple problem; if only one question is asked ("does the observed count for the outcome X=2 agree with the theory?") then an answer can be fashioned without difficulty along the lines of Example 6.D, with n = 26,306 and $\theta = \binom{12}{2} \left(\tfrac{1}{3}\right)^2 \left(\tfrac{2}{3}\right)^{10} = 0.12717$. To that question, Weldon's answer would be the correct one (the observed difference is but 1.48 standard deviations away from "Theory"). But there are evidently 13 different tests to be performed simultaneously, and they are not independent of one another – the sum of all the differences is zero.

The problem is one of testing a simple hypothesis (the dice are fair and tossed independently) against a composite alternative (they are not). The data are the list of observed values, and they are dependent (since the total 26,306 is considered fixed). In

order to specify a model we will need to introduce a multivariate distribution appropriate to this situation (and to others, such as the contingency table problem of Example 6.C).

7.1. The Multinomial Distribution.

The multinomial distribution is, as its name suggests, a generalization of the binomial. Suppose an experiment may be viewed as consisting of n independent trials, where in each trial there are $k \geq 2$ possible outcomes $A_1, A_2, \ldots, A_k$. The outcomes are assumed to be mutually exclusive (no two could occur on the same trial) and exhaustive (one of them must occur each trial). On each single trial we suppose the probabilities of the k outcomes are the same and denote these probabilities by $\theta_1, \theta_2, \ldots, \theta_k$, where of course $\theta_1 + \theta_2 + \ldots + \theta_k = 1$. After completing n trials, the counts for each of the k outcomes are recorded as $X_1, X_2, X_3, \ldots X_k$, where $X_1 + X_2 + \ldots + X_k = n$. The list ($X_1$, $X_2, X_3, \ldots X_k$) is said to have a Multinomial $(n; \theta_1, \theta_2, \ldots, \theta_k)$ distribution.

Example 7.B (Binomial). The simplest example of a multinomial experiment is in fact a binomial experiment: Here k = 2, A_1 = "Success", A_2 = "Failure", X_1 = #Successes, X_2 = #Failures. Here X_2 is redundant since it can be calculated as $X_2 = n - X$, and we do not bother listing it.

Example 7.C (Roulette Wheel). An American Roulette Wheel spun for n times will result each turn in the ball dropping into one of 38 slots, thus producing each time one of the k = 38 different outcomes, A_1 = "1", … , A_{36} = "36", A_{37} = "0", A_{38} = "00", with probabilities θ_i, i = 1, … , 38. If the wheel is perfectly balanced $\theta_i = 1/38$ for each i. the count X_i is the number of times the ball lands in the slot A_i, and $X_1 + X_2 + \ldots + X_{38} = n$.

Example 7.A (Weldon's Dice, continued). Weldon's Dice are an example of a multinomial experiment. Each of the n = 26,306 rolls of 12 dice is one trial, and there are k = 13 possible outcomes in a trial, with A_i = "exactly i of the 12 dice show a 5 or a 6", and under Weldon's hypothesis $\theta_i = \binom{12}{i}\left(\frac{1}{3}\right)^i\left(\frac{2}{3}\right)^{12-i}$, for i = 0, 1, 2, … , 12. Here X_i is the number out of the 26,306 trials where a "5 or 6" occurs i times, and $X_0 + X_1 + X_2 + \ldots + X_{12} = 26{,}306$.

The list ($X_1, X_2, X_3, \ldots X_k$) has a multivariate distribution, and the components are clearly dependent: they sum to n and if one or more is known that restricts the possible values of those that remain. For an extreme example, if $X_1 = n$, all the remaining X's must be 0. The multivariate probability function can be derived by the same argument (slightly generalized) used in Chapter 1 to derive the binomial probability function. The distribution is given by

$$p(x_1, x_2, \cdots, x_k \mid \theta_1, \theta_2, \cdots, \theta_k) = \frac{n!}{x_1! x_2! \cdots x_k!}\theta_1^{x_1}\theta_2^{x_2}\cdots\theta_k^{x_k} \text{ if } x_1 + x_2 + \cdots + x_k = n$$
$$= 0 \text{ otherwise.} \qquad (7.1)$$

The marginal distribution of a single component X_i (say X_3) can be found by a simple observation: Just call A_3 a "success" and all other outcomes are grouped together as "failures"; then X_3 is the count of the number of successes in n trials with probability

θ_3 of success on each trial, and probability $1-\theta_3 = \theta_1+\theta_2+\theta_4\ldots+\theta_k$ of failure. And so X_3 has a Binomial (n, θ_3) distribution! In general,

$$p(x_i \mid \theta_1,\theta_2,\cdots,\theta_k) = \frac{n!}{x_i!(n-x_i)!}\theta_i^{x_i}(1-\theta_i)^{n-x_i} \text{ if } 0 \le x_i \le n \qquad (7.2)$$
$$= 0 \text{ otherwise.}$$

Then, from Section 3.11, we have

$$E(X_i) = n\theta_i \text{ and } Var(X_i) = n\theta_i(1-\theta_i). \qquad (7.3)$$

7.2 Estimation for the Multinomial Distribution.

The principle of maximum likelihood applies to multiple dimensional parameters, and it give a simple answer for the case of the multinomial distribution. The likelihood function is

$$L(\theta_1,\theta_2,\cdots,\theta_k) = \frac{n!}{x_1!x_2!\cdots x_k!}\theta_1^{x_1}\theta_2^{x_2}\cdots\theta_k^{x_k} \text{ for } 0 \le \theta_i \le 1, \sum_{i=1}^{k}\theta_i = 1, \qquad (7.4)$$
$$= 0 \text{ otherwise.}$$

Eliminating the redundant $\theta_k = 1 - \theta_1 - \theta_2 - \ldots - \theta_{k-1}$ and taking partial derivatives of the loglikelihood with respect to each remaining θ_i gives

$$\frac{\partial}{\partial\theta_i}\log_e L(\theta_1,\theta_2,\cdots,\theta_k) = \frac{\partial}{\partial\theta_i}x_i\log_e\theta_i + \frac{\partial}{\partial\theta_i}x_k\log_e\left(1-\theta_1-\theta_2-\cdots-\theta_{k-1}\right)$$
$$= \frac{x_i}{\theta_i} - \frac{x_k}{1-\theta_1-\theta_2-\cdots-\theta_{k-1}}$$
$$= \frac{x_i}{\theta_i} - \frac{x_k}{\theta_k}.$$

Setting these equations equal to zero and solving, we see the maximum likelihood estimates are the sample fractions, $\hat{\theta}_i = \frac{X_i}{n}$, for $i = 1,2,\cdots k$.

7.3 Testing Simple Null Hypotheses.

Consider the setting of Example 7.C, where we are interested in testing whether or not the roulette wheel is fairly balanced, based upon a long record of prior results. The problem is, are all the 38 probabilities equal to 1/38, or are they not? There are in effect 38 separate hypotheses, one of them being redundant (if slots #1 through #37 have probability 1/38, so does #38). The null hypothesis is simple – it completely specifies the distribution of the counts as being a particular multinomial distribution – but the alternative of "unfairness" is highly composite. Any attempt to break the problem down into 37 or 38 components runs afoul of the multiple testing phenomenon: the chance of an erroneous conclusion in at least one of 37 tests is certainly not the same as the chance of being wrong in a particular single one, and the relationship of these chances is hard to

determine when the tests are, as they would be here, dependent. The general likelihood ratio test can be applied here, however, and it leads to an interesting result: a test that Karl Pearson introduced in 1900, partially in response to Weldon's question about his dice.

Consider the general problem of testing a simple null hypothesis for a multinomial distribution. That is, we wish to test

H_0: $\theta_1=a_1, \theta_2=a_2, \ldots, \theta_k = a_k$ where the a_i are known pre-specified probabilities, vs.

H_1: "otherwise", meaning at least one of the equalities in H_0 fails to hold.

For the roulette example we would have all $a_i = 1/38$; for Weldon's dice we would have $a_i = \binom{12}{i}\left(\frac{1}{3}\right)^i\left(\frac{2}{3}\right)^{12-i}$, for i = 0, 1, …, 12. The maximum likelihood estimates under H_0 are obvious (there is only one choice!), and the maximum likelihood estimates with no restrictions are the sample fractions as given in Section 7.3. The likelihood ratio test will then reject for small values of

$$\lambda = \frac{L(a_1,a_2,\cdots,a_k)}{L(\hat{\theta}_1,\hat{\theta}_2,\cdots,\hat{\theta}_k)} = \frac{L(a_1,a_2,\cdots,a_k)}{L(\frac{x_1}{n},\frac{x_2}{n},\cdots,\frac{x_k}{n})}$$
$$= \left(\frac{m_1}{x_1}\right)^{x_1}\left(\frac{m_2}{x_2}\right)^{x_2}\cdots\left(\frac{m_k}{x_k}\right)^{x_k},$$

where $m_i = na_i = E(X_i|H_0)$. Now "reject if λ is small" is equivalent to "reject if $-\log_e\lambda$ is large," and the likelihood ratio test can be seen to have the form: Reject if

$$-\log_e \lambda = -\sum_{i=1}^{k}\log_e\left(\frac{m_i}{X_i}\right)^{X_i} = \sum_{i=1}^{k} X_i \log_e\left(\frac{X_i}{m_i}\right) > C. \qquad (7.5)$$

The test can indeed be carried out in this form, as we shall see, but a different form, one that approximates this, has become much more popular. That test is the Chi-squared test; it rejects H_0 if

$$\chi^2 = \sum_{i=1}^{k}\frac{(X_i - m_i)^2}{m_i} > C.$$

The derivation of this approximation to be given here is only sketched, intended to show how two apparently dissimilar expressions can in fact be approximations to one another. The argument uses the Taylor Series expansion of a function f(x) about a value m,

$$f(x) = f(m) + (x-m)f'(m) + \frac{(x-m)^2}{2}f''(m) + \text{Rem}, \qquad (7.6)$$

where the remainder will be small for x near m if f(x) is smooth in that neighborhood. And it uses the identity $\sum_{i=1}^{k}(X_i - m_i) = n - n = 0.$ (7.7)

We apply (7.6) with f(x) = $x\log_e(x/m)$, noting in that case f(m) = 0, f′(m) = 1, and f″(m) = 1/m, and so $f(x) = (x-m) + \frac{1}{2}\frac{(x-m)^2}{m} + \text{Rem}$. We then have

$$
\begin{aligned}
-\log_e \lambda &= \sum_{i=1}^{k} X_i \log_e\left(\frac{X_i}{m_i}\right) \\
&= \sum_{i=1}^{k}(X_i - m_i) + \tfrac{1}{2}\sum_{i=1}^{k}\frac{(X_i - m_i)^2}{m_i} + \sum_{i=1}^{k}\text{Rem(i)}, \quad \text{using (7.6)}, \\
&= 0 + \tfrac{1}{2}\sum_{i=1}^{k}\frac{(X_i - m_i)^2}{m_i} + \text{Rem}_2, \quad \text{using (7.7)}, \\
&= \tfrac{1}{2}\chi^2 + \text{Rem}_2.
\end{aligned}
$$

It can be shown using advanced methods that under the null hypothesis H_0, the remainder Rem_2 is (with high probability) small relative to either of the other terms, and so we have that

$$-2\log_e \lambda \cong \chi^2. \qquad (7.8)$$

The two terms will be numerically nearly the same if both are small, and when one or the other is large the other will be so also, although then the numeric values of the two may be quite different. In any event, the likelihood ratio test (which rejects H_0 when $-\log_e\lambda$ is large, or equivalently when $-2\log_e\lambda > C$), has nearly the same performance characteristics – the same probabilities of error and power – as the Chi-squared test which rejects when $\chi^2 > C$. The Chi-squared test is therefore a close approximation to the likelihood ratio test. It remains to find the cutoff value C.

7.4 The Distribution of the Chi-Squared Test Statistic.

In order to fully specify the Chi-squared test we need C so that $P(\chi^2 > C|H_0) = \alpha$. The Chi-squared distributions were introduced in Chapter 5 (Example 5.G), and at first appearance they would seem unlikely to describe the distribution of this statistic: They are a family of continuous probability distributions, and this statistic, since it is derived from a list of counts, must have a discrete distribution. The distribution of χ^2 under the hypothesis H_0 is indeed discrete, and while in principle it could be evaluated, it would depend upon the hypothesized values for the θ's and the computation would be an enormous headache, one that would require revision with every new application. Fortunately the coincidence of names is no accident: The (discrete) distribution of χ^2 under the null hypothesis is indeed approximately a (continuous) Chi-squared distribution with k–1 degrees of freedom, in the sense that probabilities such as $P(\chi^2 > C|H_0)$ can be approximated closely if n is large by the corresponding area under the Chi-squared density with k–1 degrees of freedom, and that area may be conveniently found from tables or simple computer routines.

The proof that χ^2 has approximately a Chi-squared distribution with k–1 degrees of freedom will only be given here for k = 2. In this case the data are (X_1, X_2) which we may write as (X, n–X), as in Example 7.B. Write the null hypothesis values as $(a_1, a_2) = (a, 1–a)$. Then

$$\begin{aligned}
\chi^2 &= \sum_{i=1}^{2} \frac{(X_i - m_i)^2}{m_i} \\
&= \frac{(X_1 - m_1)^2}{m_1} + \frac{(X_2 - m_2)^2}{m_2} \\
&= \frac{(X - na)^2}{na} + \frac{((n - X) - n(1-a))^2}{n(1-a)} \\
&= \frac{(X - na)^2}{na} + \frac{(X - na)^2}{n(1-a)} \\
&= \frac{(X - na)^2}{na(1-a)} \\
&= \left(\frac{X - na}{\sqrt{na(1-a)}} \right)^2
\end{aligned}$$

Now, the Central Limit Theorem tells us (since X is an aggregate based upon n trials, E(X)=na, and Var(X)=na(1–a)) that the variable in parentheses has approximately a N(0,1) distribution. It follows that χ^2 is the square of an approximately N(0,1) random variable, and thus itself has approximately a Chi-squared distribution with 1 degree of freedom. A proof for larger k can be constructed in a similar manner, by re-expressing χ^2 as the sum of the squares of k–1 random variables that are uncorrelated and each approximately N(0,1). One consequence of that result can be easily verified. The expectation of a Chi-squared distributed random variable is equal to its degrees of freedom, and

$$\begin{aligned}
E(\chi^2) = E\sum_{i=1}^{k} \frac{(X_i - na_i)^2}{na_i} \\
&= \sum_{i=1}^{k} \frac{E(X_i - na_i)^2}{na_i} \\
&= \sum_{i=1}^{k} \frac{Var(X_i)}{na_i} \\
&= \sum_{i=1}^{k} \frac{na_i(1-a_i)}{na_i} \\
&= \sum_{i=1}^{k} (1 - a_i) = k - \sum_{i=1}^{k} a_i = k - 1
\end{aligned}$$

As a consequence of this fact, the tables of the Chi-squared distribution can be used to find the cutoff C for the Chi-squared test. And from the approximation (7.7) the

same cutoff value can be used for the likelihood ratio test in the form $-2\log_e\lambda > C$. In either case,

$$P(\chi^2 > C|H_0) \cong P(-2\log_e\lambda > C|H_0) \cong \alpha,$$

if C is taken as the value such that the probability of a Chi-squared random variable with k–1 degrees of freedom exceeding C is α. The approximation is, like the Central Limit Theorem, the more accurate the larger n is. As a "rule of thumb," statisticians usually regard n "large" if all m_i are at least 5, or even if they exceed 3.5. When one or more m_i are smaller than that, it is customary to group values in performing the test, as illustrated in this example.

Example 7.A (Weldon's Dice, continued). The data Weldon sent to Karl Pearson and others have two low expected counts, in categories 11 and 12. We may combine them with the value in category 10 to get:

Table 7.2. Weldon's dice data with the last three categories grouped together.

No. of Dice X showing 5 or 6	Observed	Theory	Difference
0	185	203	–18
1	1149	1217	–68
2	3265	3345	–80
3	5475	5576	–101
4	6114	6273	–159
5	5194	5018	176
6	3067	2927	140
7	1331	1254	77
8	403	392	11
9	105	87	18
10 – 12	18	14	4
Total	26,306	26,306	0

In this form, k = 11, and $\chi^2 = (-18)^2/203 + (-68)^2/1217 + \ldots + 4^2/14 = 35.9$ (35.5 if full accuracy is carried in computing the m_i). For k–1 = 10 degrees of freedom the chance a Chi-squared distributed random variable exceeds 35.5 is nearly equal to 0.0001, hence for any α .0001 or larger we should reject the hypothesis H_0: Karl Pearson's initial assessment in 1894 that H_0 was "intrinsically incredible" was correct.

7.5 Testing Composite Hypotheses.

In Sections 7.3 and 7.4 we considered the likelihood ratio and Chi-squared tests for testing simple null hypotheses, null hypotheses that completely specify the distribution of the list of counts. Most applications of these techniques require one or more parameters to be determined under the null hypothesis, to be estimated using the same data as are at hand to test the hypothesis.

Example 7.A (Weldon's Dice, Continued). When Weldon was confronted with the results of Pearson's analysis, he suggested that perhaps the dice may in fact be slightly weighted towards 5 and 6; they may be independently tossed with the same

chance θ of a 5 or 6, but that chance is not exactly 1/3. Weldon was in effect suggesting that Pearson should test the null hypothesis that the table represents count from a Binomial (n, θ) distribution, rather than a Binomial (n, 1/3) distribution, where θ would be determined from the data. Weldon found from the data of Table 7.1 that the total number of 5's and 6's in the entire 12x26,306 = 315,672 dice tossed is 0x185+1x1149+2x3265+ … +11x4+12x0 = 106,602. Then if we consider the entire data set as a gigantic experiment where 315,672 dice produce a count of 106,602 showing 5's or 6's, the maximum likelihood estimate of θ is 106,602/315,672 = 0.33769862, slightly larger than 1/3. Table 7.2 shows Weldon's data with the "Theory" column recomputed according to $m_i = 26{,}306 \times \binom{12}{i} (.33769862)^i (.66230138)^{12-i}$ for i = 0, 1, 2, …, 12.

Table 7.2. Weldon's Dice Data. The Theory column has been recomputed using the maximum likelihood estimate of the probability of a 5 or 6, namely 0.33769862.

No. of Dice X showing 5 or 6	Observed	Theory	Difference
0	185	187.4	-2.4
1	1149	1146.5	2.5
2	3265	3215.2	49.8
3	5475	5464.7	10.3
4	6114	6269.3	-155.3
5	5194	5114.7	79.3
6	3067	3042.5	24.5
7	1331	1329.7	1.3
8	403	423.8	-20.8
9	105	96.0	9.0
10	14	14.7	-0.7
11	4	1.4	2.6
12	0	0.1	-0.1
Total	26,306	26,306	0.0

If χ^2 is recomputed from these numbers, grouping the last three categories together, we find $\chi^2 = 8.2$, much smaller than the value 35.5 found before. Is this procedure legitimate? Can we allow the data to play a role in choosing the null hypothesis? And if we do, is the comparison with a Chi-squared distribution with k–1 degrees of freedom still valid? The answers turn out to be "yes", "yes", and "not quite".

The testing problem we are faced with here can be described formally as follows: The data consist of observed counts (X_1, X_2, X_3, … X_k) modeled by a multinomial (n; θ_1, θ_2, …, θ_k) distribution. We wish to test a composite null hypothesis vs. a composite alternative. In particular, we wish to test

H_0: $\theta_1=a_1(\theta)$, $\theta_2=a_2(\theta)$, … , $\theta_k = a_k(\theta)$ where the $a_i(\theta)$ are known functions of an unknown parameter θ, which may be one or more dimensional, vs.

H_1: "otherwise", meaning at least one of the equalities in H_0 fails to hold.

We emphasize that this is testing a different hypothesis than before; we no longer insist the dice be perfectly balanced and now only test if the data are characteristic of some binomial distribution, not necessarily one with $\theta = 1/3$. We allow the data to select the value of θ by maximum likelihood, as provided for by the general likelihood ratio test. We then compute the Chi-squared statistic from

$$\chi^2 = \sum_{i=1}^{k} \frac{(X_i - m_i(\hat{\theta}))^2}{m_i(\hat{\theta})},$$

where $m_i(\hat{\theta}) = na_i(\hat{\theta})$ are the maximum likelihood estimates of the expected counts. The theory of these tests is more complicated than for the simple case (indeed, Karl Pearson got the wrong answer in his initial publication in 1900), but the correct answer is simple nonetheless: If the maximum likelihood estimate of θ is calculated as assumed for Fisher's Approximation Theorem in Chapter 5 (i.e. from differentiating the likelihood function or loglikelihood function), then under the null hypothesis χ^2 has approximately a Chi-squared distribution as before, but with only $k - 1 - m$ degrees of freedom, where m is the number of parameters estimated under the null hypothesis.

In the case of Weldon's dice data, $m = 1$ and after grouping $k = 11$, so the value $\chi^2 = 8.2$ should be compared with a Chi-squared distribution with $11 - 1 - 1 = 9$ degrees of freedom, rather than 10 as was the case for testing a completely specified hypothesis. The value 8.2 is very nearly the median of the Chi-squared distribution with 9 degrees of freedom; Weldon's second hypothesis cannot be rejected.

Strictly speaking, to use this approximation the estimate $\hat{\theta}$ should be found from the same data used to compute χ^2. With Weldon's data we found $\hat{\theta}$ from the original data, but computed χ^2 for the grouped data. As long as the grouping is not severe the effect is minimal, however.

The intuitive rationale for this adjustment for an estimated θ should be clear: As long as the data are allowed to choose the closest from among a class of null hypotheses, we should expect the fit to be closer. In the case of Weldon's data, choosing 0.33768862 instead of 1/3 as the value for θ reduced χ^2 from 35.5 to 8.2. In that case there does appear to be a slight bias to the dice, but even if the dice had been absolutely perfectly balanced there would have been a reduction in χ^2 unless exactly 1/3 of the 315,672 dice showed 5 or 6. To allow for this anticipated improvement, we compensate by reducing the standard for comparison, by reducing the degrees of freedom. In the example, before reduction the expected value of the comparison distribution is 10; after reduction it is 9.

7.6 Testing Independence in Contingency Tables.

Galton's data on the heights of married couples was introduced in Chapter 6; they are given again here, with the addition of the marginal totals.

Table 7.3. Galton's data on the relative heights of husbands and wives for 205 married couples (Galton, 1889, p. 206)

	Wife:	Tall	Medium	Short	Totals
	Tall	18	28	14	60
Husband:	Medium	20	51	28	99
	Short	12	25	9	46
	Totals	50	104	51	205

These data are an example of a contingency table, a cross-classification of a number n (=205, here) of "individual cases" (couples, here) according to two different classificatory criteria. These data are then a list of counts, but because of their relationship to the classifying factors they are a "rectangular list" that is best described with double indices. Accordingly, let us denote the count in the (i,j)th "cell" – the entry in row i and column j – by X_{ij}, and so if we have r rows and c columns, there will be rxc counts in the list, and we will have $\sum_{i=1}^{r}\sum_{j=1}^{c} X_{ij} = n$. If the two criteria for classification are A (at levels $A_1, A_2, \ldots, A_r$) and B (at levels $B_1, B_2, \ldots, B_c$), and we let the marginal totals be given by $X_{i+} = \sum_{j=1}^{c} X_{ij}$ and $X_{+j} = \sum_{i=1}^{r} X_{ij}$, we have the general two-way rxc table,

		Factor B:					
		B_1	B_2	…	…	B_c	Totals
	A_1	X_{11}	X_{12}	…	…	X_{1c}	X_{1+}
	A_2	X_{21}	X_{22}	…	…	X_{2c}	X_{2+}
Factor A:	…	…	…			…	…
	…	…	…			…	…
	A_r	X_{r1}	X_{r2}	…	…	X_{rc}	X_{r+}
	Totals	X_{+1}	X_{+2}	…	…	X_{+c}	$n=X_{++}$

Under some circumstances tables such as these can be modeled by a multinomial distribution; for example, if the n individual cases are distributed at random among the rxc table cells according to the same probability distribution. Essentially, it as if the table represents the outcomes from n "spins" of a rectangular roulette wheel. Let $\theta_{ij} = P(A_i \cap B_j)$ be the probability that an individual case is assigned to the (i,j)th cell. Then $\sum_{i=1}^{r}\sum_{j=1}^{c} \theta_{ij} = 1$ and we can arrange the probabilities in the same pattern as the data:

		B_1	B_2	…	…	B_c	Totals
	A_1	θ_{11}	θ_{12}	…	…	θ_{1c}	θ_{1+}
	A_2	θ_{21}	θ_{22}	…	…	θ_{2c}	θ_{2+}
Factor A:	…	…	…			…	…
	…	…	…			…	…
	A_r	θ_{r1}	θ_{r2}	…	…	θ_{rc}	θ_{r+}
	Totals	θ_{+1}	θ_{+2}	…	…	θ_{+c}	$1=\theta_{++}$

We may then test the hypothesis that the categories are independent of one another; that is, that $P(A_i \cap B_j) = P(A_i)P(B_j)$ for all i and j. In terms of our notation this becomes

H_0: $\theta_{ij} = a_{ij}(\theta_{1+},\theta_{2+},\ldots\theta_{r+},\theta_{+1},\theta_{+2\ldots},\theta_{+c}) = \theta_{i+},\times\theta_{+j}$ for all i and j, vs.

H_1: "otherwise".

This formulation places the problem exactly in the class considered in Section 7.5, with the known functions a_{ij} depending upon the marginal probabilities $\theta_{1+},\theta_{2+},\ldots\theta_{r+},\theta_{+1},\theta_{+2\ldots},\theta_{+c}$, which must then be estimated, based upon the data using the method of maximum likelihood.

The estimation of the marginal probabilities is a simple matter, since, for example, the row marginal total counts $X_{1+},X_{2+},\ldots X_{r+}$ have a multinomial (n; $\theta_{1+},\theta_{2+},\ldots\theta_{r+}$) distribution, we have from Section 7.2 that the maximum likelihood estimates are the fractions X_{1+}/n, $X_{2+}/n,\ldots, X_{r+}/n$. Similarly, the maximum likelihood estimates of the column marginal probabilities are X_{+1}/n, $X_{+2}/n,\ldots, X_{+c}/n$. The under the null hypothesis H_0 the maximum likelihood estimates of the cell probabilities θ_{ij} are the products $(X_{i+}/n)(X_{+j}/n)$, and the maximum likelihood estimates of the expected counts are

$$\hat{m}_{ij} = n\left(\frac{X_{i+}}{n}\right)\left(\frac{X_{+j}}{n}\right) = \frac{X_{i+}X_{+j}}{n}.$$

The Chi-squared statistic then becomes

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{\left(X_{ij} - \left(\frac{X_{i+}X_{+j}}{n}\right)\right)^2}{\left(\frac{X_{i+}X_{+j}}{n}\right)},$$

While at first it may appear that r + c parameters are being estimated under the null hypothesis, in fact since the row margins and the column margins each sum to 1.0, in each case one parameter is redundant (e.g. the first r–1 row marginal probabilities determine the rth). And so there are but (r–1) + (c–1) parameters being estimated. The

degrees of freedom for this Chi-squared statistic is therefore rc – 1 – [(r–1) + (c–1)] = (r–1)(c–1).

Example 7.D (Galton's data, continued). Here we find that

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{\left(X_{ij} - \left(\frac{X_{i+}X_{+j}}{n}\right)\right)^2}{\left(\frac{X_{i+}X_{+j}}{n}\right)} = \frac{\left(18 - \left(\frac{60\cdot 50}{205}\right)\right)^2}{\left(\frac{60\cdot 50}{205}\right)} + \cdots,$$

$$= 2.91$$

with (r–1)(c–1) = 2x2 = 4 degrees of freedom. Thus χ^2 is even smaller than its expected value of 4, and this test supplies no grounds for rejecting the hypothesis that in Galton's society marriage partners were selected without regard to height.

7.7 Tests of Homogeneity.

Any statistical analysis involves an assumed model. Sometimes, as with scientifically randomized surveys, the appropriateness of the model can be guaranteed by design. More often the model will represent a plausible reference frame and the strength of the conclusions will vary according to the degree of model plausibility and the sensitivity of the inferences to departures from the model. The contingency test analysis of the previous section was based upon the multinomial model: the n individual cases are modeled as being distributed among the rxc cells of the table independently according to rxc probabilities, which may or may not correspond to the null hypothesis of independence. This is referred to as Full Multinomial Sampling.

If one set of marginal totals is fixed, either by design or after conditioning, the Product Multinomial Sampling model may be appropriate. If both marginal totals are fixed, the Hypergeometric Sampling model may be a reasonable basis for analysis. We shall introduce these models through examples. Both of these models are intimately related mathematically to the Full Multinomial Sampling model (and all are related to another model, the Poisson model, to be introduced later). This mathematical relationship will permit us to perform the same formal analysis in all cases, although the terms we employ in describing the hypotheses and the interpretations we make are subtly different.

Example 7.E (The Draft Lottery). In 1970, the US conducted a draft lottery to determine the order of induction of males aged 19 – 26. The 366 possible birthdates (including February 29) were put in capsules and drawn "at random" from a bowl; the order in which they were selected was their "drawing number." The following table summarizes the results (from S. E. Fienberg, "Randomization and Social Affairs: The 1970 Draft Lottery," Science (1971), Vol. 171, pp. 255-261). The question at issue is, was the lottery fairly conducted?

Table 7.4.

Drawing numbers	Months												
	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Totals
1-122	9	7	5	8	9	11	12	13	10	9	12	17	122
123-244	12	12	10	8	7	7	7	7	15	15	12	10	122
245-366	10	10	16	14	15	12	12	11	5	7	6	4	122
Totals	31	29	31	30	31	30	31	31	30	31	30	31	366

Example 7.F (Right-Handedness). To what degree is the propensity to be right-handed socially determined? Is it the same in different cultures? In different historical epochs? Two psychologists addressed this question by examining works of art that portrayed activities that could be judged as being done right- or left-handedly. (Stanley Coren and Clare Porac, "Fifty Centuries of Right-Handedness: The Historical Record" Science (1977), Vol. 198, pp. 631-632.) The following tables summarize their findings, looking at the data in two different ways.

Table 7.5. Counts of 1180 art works showing activity that can be categorized as left- or right-handed, (a) by geographical area, and (b) by historical epoch.

(a)

	Right	Left	Total	% Right
Central Europe	312	23	335	93%
Medit. Europe	300	17	317	95%
Middle East	85	4	89	96%
Africa	105	12	117	90%
Central Asia	93	8	101	92%
Far East	126	13	139	91%
Americas	72	10	82	88%
Total	1093	87	1180	92.6%

(b)

Pre 3000 BC	35	4	39	90%
2000 BC	44	7	51	86%
1000 BC	89	10	99	90%
500 BC	134	8	142	94%
~0 BC	130	4	134	97%
AD 500	39	3	42	93%
AD 1000	57	7	64	89%
AD 1200	40	1	41	98%
AD 1400	44	6	50	88%
AD 1500	63	5	68	93%
AD 1600	68	4	72	94%
AD 1700	66	5	71	93%
AD 1800	95	6	101	94%
AD 1850	38	1	39	97%
AD 1900	71	6	77	92%
AD 1950	80	10	90	89%
Total	1093	87	1180	92.6%

Example 7.G. (First Word Usage). In a study of literary characteristics, samples of prose from the British economists James Mill and his son John Stuart Mill were examined, and a count was made of the choices of the words they used to start sentences. Do the two have the same propensity to choose initial words for sentences?

Table 7.6 (O'Brien and Darnell, 1982, p. 116)

First Word:	But	Where	This/It/ Thus/And	A/By	All others	Totals
James Mill	39	26	339	33	638	1075
John Stuart Mill	38	16	112	11	274	451
Totals	77	42	451	44	912	1526

Example 7.H (Severity of Smallpox and its Relationship to Vaccination) Here is a data set published by Karl Pearson in 1910 (in Biometrika, Vol. 7, p. 257), that gives data collected by Dr. John Brownlee on the relationship between the severity of an attack of smallpox and the length of time since vaccination. Is there evidence here for a beneficial effect of vaccination? There are five degrees of severity, from most severe (Haemorrhagic) to least (very sparse).

Table 7.7

Years since Vaccination	Severity of Attack: Hæmorrhagic	Confluent	Abundant	Sparse	Very Sparse	Totals
0 – 10	0	1	6	11	12	30
10 – 25	5	37	114	165	136	457
25 – 45	29	155	299	268	181	932
Over 45	11	35	48	33	28	155
Unvaccinated	4	61	41	7	2	115
Totals	49	289	508	484	359	1689

7.8 Sampling Models for the Examples.

In the case of Galton's table (Table 7.3), the data were gathered as a part of a larger study of human heredity. He published an account book (the "Record of Family Faculties") that people might use for recording their family history, and he offered £500 in prizes (a very large sum then) for the best and most complete records he received. He obtained 205 families' data in this way, and those data included the married couple's heights. Was this a truly random sample from a population of married couples? If we consider that it effectively was, and that Galton classified their heights according to a pre-specified definition of the categories "Tall", "Medium", and "Short" (the definition was different for the two sexes), the assumption of full multinomial sampling would be amply justified. The actual sample probably missed this ideal, but not by much, unless willingness in Victorian society to record personal data were related to height. The categories may have been pre-specified, but they more likely were simply a convenient grouping of data he collected: the marginal totals would then be approximately fixed by definition. In that case there is hardly a random distribution of heights among categories, even if choices of mates were made completely without regard to height. Since rigorous

fulfillment of the full multinomial model is rare, and knowledge of the exact conditions under which data are gathered even rarer, it is useful to ask how such departures might influence the conclusions.

Example 7.E (The Draft Lottery) presents one extreme case of a departure from Full Multinomial Sampling. The row totals are fixed by definition (each row consists of a third of the data), and while the column totals vary slightly, the lengths of months are historically and astronomically pre-determined. A skeptic might ask, why divide the drawing numbers into thirds and the year into months, why not sixths and quarters? But if we assume (as seems reasonable) that the grouping was not dependent upon the data, and was not done to mask or accentuate features of the association of the two factors, then this example represents an extreme case in that both margins are fixed by design. This is called the Hypergeometric Sampling model because the probability distribution of the list of counts under the null hypothesis is called a hypergeometric distribution; the name is derived from the use of the term in the branch of mathematics called "special functions", where the null hypothesis probabilities appear. The null hypothesis here is clear: all possible assignments of the numbers 1 to 366 to birthdates are equally likely. The counting arguments of Chapter 1 can be used to find the probabilities of these assignments; for example, the probability under H_0 that the counts would arise exactly as given in Table 7.4 is

$$P(Table) = P(Jan)P(Feb \mid Jan)P(Mar \mid Jan, Feb)\cdots P(Dec \mid Jan-Nov)$$

$$= \frac{\binom{122}{9}\binom{122}{12}\binom{122}{10}}{\binom{366}{31}} \frac{\binom{113}{7}\binom{110}{12}\binom{112}{10}}{\binom{335}{29}} \cdots \frac{\binom{17}{17}\binom{10}{10}\binom{4}{4}}{\binom{31}{31}}$$

$$= \frac{\left(\frac{122!122!122!}{9!12!10!7!12!10!\cdots 17!10!4!}\right)}{\left(\frac{366!}{31!29!\cdots 31!}\right)}$$

The analysis of this table will be discussed later.

On the other hand, Examples 7.F, 7.G, and 7.H might be viewed as belonging to an important class of problems where only one of the table's set of marginal totals is reasonably treated as fixed by sample design, and where it is convenient to do the analysis conditionally on the marginal totals. The artists' choices of "models" might plausibly be considered to be effectively random with regard to handedness, so each row in Tables 7.5 (a) and (b) would then be a multinomial experiment (with k=2); the question then would be, is P(Right-handed) the same in every row? The choice by a writer of the first word of a sentence might be considered essentially random (according to that writer's propensities) and the question there is, are the propensities the same for the two writers? Given the date and fact of vaccination, the severity of a subsequent smallpox attack might follow a multinomial distribution (with k=5), but the probabilities of the degrees of severity might differ with age of vaccination. In all three cases the row counts might be considered to be multinomial distributed, and the independence of the rows could be defended, but the distribution of counts between rows is another matter.

The choices for the row totals may even have been predetermined (say if the art works resulted from a stratified sample that guaranteed a certain representation for each area, epoch). But even if not pre-determined, we cannot in any case consider the counts to have arisen from randomly allocating individual artists, sentences, or vaccinees to different rows; we cannot reasonably say the individuals were randomly row-assigned, and an analysis that treats the row margins as given has appeal.

Let us then consider the Product Multinomial Sampling model, where a table of counts $\{X_{ij}; i=1, 2, \ldots, r, j=1, 2, \ldots, c\}$ satisfies

(1) Each row i, $(X_{i1}, X_{i2}, \ldots, X_{ic})$, has a multinomial $(c; \theta_{i1}, \theta_{i2}, \ldots, \theta_{ic})$ distribution, and

(2) The rows are mutually independent; note that now $\theta_{i1} + \theta_{i2} + \ldots + \theta_{ic} = 1$ for each row.

Categories B:		B_1	B_2	…	…	B_c	Totals
	A_1	θ_{11}	θ_{12}	…	…	θ_{1c}	1
	A_2	θ_{21}	θ_{22}	…	…	θ_{2c}	1
Rows A:	…	…	…			…	…
	…	…	…			…	…
	A_r	θ_{r1}	θ_{r2}	…	…	θ_{rc}	1
	Averages	θ_1	θ_2	…	…	θ_c	1

The hypotheses to be tested are

H_0: $(\theta_{i1}, \theta_{i2}, \ldots, \theta_{ic}) = (\theta_1, \theta_2, \ldots, \theta_c)$ for all i, vs. H_1: "otherwise".

That is, we test if the probabilities are the same for each row. For example, that
P(Right-handed|Pre-3000 BC) = P(Right-handed|AD 1800), or
P("Where"|J. Mill) = P("Where"|J. S. Mill), or
P("Confluent"|0-10 years) = P("Confluent"|unvaccinated).

The name of this model comes from the form of the likelihood function:

Under H_0:

$$L(\theta_1,\theta_2,\cdots,\theta_c) = \prod_{i=1}^{r}\left(\frac{X_{i+}!}{\prod_{j=1}^{c} X_{ij}!}\theta_1^{X_{i1}}\theta_2^{X_{i2}}\cdots\theta_c^{X_{ic}}\right)$$

$$= \prod_{i=1}^{r}\left(\frac{X_{i+}!}{\prod_{j=1}^{c} X_{ij}!}\right)\theta_1^{X_{+1}}\theta_2^{X_{+2}}\cdots\theta_c^{X_{+c}}$$

With no restrictions:

$$L(\theta_{11},\theta_{12},\cdots,\theta_{rc}) = \prod_{i=1}^{r}\left(\frac{X_{i+}!}{\prod_{j=1}^{c} X_{ij}!}\right)\prod_{i=1}^{r}\left(\theta_{i1}^{X_{i1}}\theta_{i2}^{X_{i2}}\cdots\theta_{ic}^{X_{ic}}\right)$$

It follows from Section 7.2 that the respective maximum likelihood estimates are, under H_0. $\hat{\theta}_j = \frac{X_{+j}}{X_{++}}$ and $\hat{m}_{ij} = \frac{X_{i+}X_{+j}}{X_{++}}$; under no restrictions, $\hat{\theta}_{ij} = \frac{X_{ij}}{X_{i+}}$ and the MLE of the mean is $X_{i+}\hat{\theta}_{ij} = X_{ij}$. The likelihood ratio becomes

$$\lambda = \prod_{i=1}^{r}\prod_{j=1}^{c}\left(\frac{\hat{m}_{ij}}{X_{ij}}\right)^{X_{ij}}.$$

This is the same as for the Full Multinomial Sampling model, and so it will lead to the same test, the Chi-squared test! With row totals fixed there are c–1 "freely varying" cells in each of r rows, for a total of (c–1)r, and there are (c–1) estimated parameters, leading to (c–1)r – (c–1) = (r–1)(c–1) degrees of freedom, just as in the case of full multinomial sampling. This means that both models use exactly the same test!

For the Hypergeometric Sampling model the Chi-squared test statistic can be shown under the null hypothesis to also have approximately a Chi-squared distribution with (r–1)(c–1) degrees of freedom; that is, the same test can be used for all three models and in this important sense we do not have to worry as to which of the three models is appropriate. Whether Galton fixed the numbers of heights for wives, for husbands, for both, or for neither, his hypothesis that mates are selected without regard to height cannot be rejected.

While the tests are the same in the three cases, the hypotheses are different. With product multinomial sampling we test the hypothesis of homogeneity, that is, the row probabilities are all the same – the rows are homogeneous in that they all reflect samples from the same distribution. With the row classification being assigned by design, it would simply not make sense to talk about the probability of the being in the ith row, and so the hypothesis of independence does not apply here. With hypergeometric sampling neither row nor column probabilities apply; we test the hypothesis that all of the possible arrangements of the n individuals among the tables cells are equally likely, among those consistent with the marginal totals. Sometimes this is called the hypothesis of no association between row and column factors.

The mathematical relationship among the models is simple: The conditional distribution of a full multinomial table of counts, given a set of marginal totals, is product multinomial. The conditional distribution of a product multinomial table of counts, given the remaining set of marginal totals, is hypergeometric. These claims are easy to verify and left for an exercise.

In summary, we have a general class of tests for counted data, all derived from the likelihood ratio test. These Chi-squared tests are all of the form

$$\chi^2 = \sum_{\text{all categories}} \frac{(\text{observed count} - \text{expected})^2}{\text{expected}}$$

where "expected" = E(observed count|H_0), or, when this is incompletely specified, the maximum likelihood estimate of E(observed count|H_0).

The test is accomplished by comparing this statistic with a percentage point of a Chi-squared distribution with the appropriate number of degrees of freedom. The situations we have encountered are

1) Test of goodness of fit, completely specified H_0, k categories, and k–1 df. (Example: Roulette)

2) Test of goodness of fit, incompletely specified H_0, k categories, and k–1 – (#parameters estimated) df. (Example: Testing for a binomial distribution)

3) Testing independence with an rxc contingency table with full multinomial sampling, and (r–1)(c–1) df.

4) Testing homogeneity with an rxc contingency table with product multinomial sampling, and (r–1)(c–1) df.

5) Testing no association with an rxc contingency table and hypergeometric sampling, and (r–1)(c–1) df.

It is important to bear in mind that while the likelihood ratio idea can be applied for any size sample, the use of a Chi-squared distribution for reference is an approximation to a discrete distribution and will only give serviceable answers if the size n of the sample is at least moderately large. As a rule of thumb we look for "expected"≥ 3.5 in all cells, but the larger n is, the better the approximation. The Chi-squared statistics are derived from the likelihood ratio test and the test can as well be applied in the form, reject if $-2\log_e\lambda > K$, K taken from the same Chi-squared reference distribution. The tests will give similar but not necessarily identical results; there is no clear preference between them on the basis of power. The study of which is more powerful for which alternatives is an ongoing research problem. These tests are sometimes described as "omnibus" tests, in that they test a fairly specific null hypothesis against a very broad class of alternatives, and thereby guard against a sort of "data dredging" where one searches out a largest deviation (there always will be one!) and tests as if it were the only one that was ever of interest.

7.9 P- Values.

The approach to testing that has been outlined is the so-called classical approach, where a particular level α is specified, often at 5% or 1%, and on that basis a simple "accept or reject" decision is made. This approach was developed at a time when the use of tabled values was a practical necessity, but it has the unattractive feature of elevating unimportant differences to a decisive level. For example, for a Chi-squared distribution with 6 df, a table will tell you that the upper 5% point is 12.5916. Taken literally, this

would mean we should reject H_0 with $\chi^2 = 12.60$ but accept H_0 with 12.50. This would be unpalatable even if we were not employing the Chi-squared distribution only as an approximation. Not only is "5%" only a social convention, but the evidence against the null hypothesis is about the same in both cases, and a way of expressing this as the slight difference it is would be desirable. With the ability to compute an exact percentage point easily the practice of quoting "P-values" has grown. For a given test statistic (say χ^2) and form of test (say "reject if $\chi^2 \geq K$"), the P-value of the observed test statistic is the smallest level α at which the null hypothesis would be rejected. In the above example, $\chi^2 = 12.60$ would have $P = 0.0498$, while $\chi^2 = 12.50$ would have $P = 0.0516$, making the slight difference visible. When tables are used, P is often reported through inequalities: as with "P> .05", or "P <.05", or "P<<.01" (meaning P is much less than .01).

7.10 A Further Discussion of the Examples.

Most applications of a statistical method encounter peculiarities. Either aspects of a study's design do not quite fit with statistical theory, or the scientific questions are not quite answered by the analysis provided by that theory. In the case of Galton's data, we have already discussed potential problems with the sampling method and category definition, but neither of these seems like a decisive challenge to the analysis. One other point is worth remarking on, however. The Chi-squared test is an omnibus test of a null hypothesis against a bewildering variety of alternatives. This can be seen as a virtue; it is also a limitation. The statistic and the test are completely insensitive to the ordering of the categories: The test is unaffected if "Tall" is placed between "Short" and "Medium". A consequence is that the test, by protecting against all possibilities, may have lower power than others for some alternatives. We have seen one example of this phenomenon in Chapter 6, with the test for a normal distribution mean of $H_0: \mu=\mu_0$ vs. $H_1: \mu\neq\mu_0$. That test achieved its goal at the expense of lower power against $H_1: \mu>\mu_0$, lower than a test designed for that specific set of alternatives. In the case of contingency tables with ordinal categories, such as Galton's, directed tests are available that can have greater power for ordered alternatives. (See Agresti, 1984, for example.) In Galton's case such an analysis does not change the conclusion. Interestingly, Galton himself tried an ordinal analysis. He focussed upon the Table's corners, noting that a total of 32 couples were "Short" married to "Tall", and there were 27 couples with both husband and wife in the extreme groups. These numbers were so close as to be statistically indistinguishable (if a coin produced 32 heads and 27 tails no one would question its fairness). Galton's conclusion was "There are undoubtedly sexual preferences for moderate contrast in height, but marriage choice is guided by so many and more important considerations that questions of stature appear to exert no perceptible influence upon it." (Galton, 1886, p. 251) The effect, if present, is sufficiently small that it would require much more data to reliably detect it.

Low drawing numbers in the Draft Lottery put young men at serious risk in 1970, and the lottery was carefully scrutinized for signs of unfairness. The data appeared to show lower numbers assigned to later birthdates (e.g December) than to earlier (e.g.

January). The hypothesis of hypergeometric sampling does seem like the reasonable one to entertain here. The Chi-squared statistic for Table 7.4 is χ^2 32.16, with 2x11 = 22 df, corresponding to a P-value of 2%. Thus the lottery does not seem to have achieved the desired level of randomness. This is discussed and confirmed by other analyses by Fienberg (1971).

The data on right-handedness can at best be considered as an essentially random sample of the way artists choose to depict people, and that choice itself could be socially influenced. In addition, the propensity of artists to copy one another could introduce a dependence among the trials, making even the product multinomial sampling model suspect. If we accept (as seems plausible) that such dependence is weak, then the values $\chi^2 = 8.14$ (6 df) and $\chi^2 = 17.04$ (15 df) for Table 7.5 suggest homogeneity in the right-handedness propensity both geographically and historically. An analysis of Table 7.5(b) that incorporated time order would not lead to a different conclusion. It would be interesting to examine the three-way classification for which Tables 7.5(a) and (b) are marginal totals, but those data were not published.

On the face of it, the data of Table 7.6 show a strong indication of very different styles for the two Mills: $\chi^2 = 21.34$ (4 df) and $P<<.005$. To first appearances, either they have different styles or a very rare fluke has occurred in these passages. A closer investigation of the sampling and presentation of these data shows this conclusion unfounded. One possible criticism would be that sentences are not independent, that each Mill would avoid too many similar sentence beginnings in close proximity. But that effect (which might in the aggregate be expected to be weak) would tend to spread out the distribution in rows and make the styles appear more similar. The real problem with the analysis lies elsewhere. The data were collected to help decide which Mill wrote an anonymous book review of the time. The question would have been hopeless if the two had exhibited the same style. To accentuate differences, the investigators sought out a way of grouping first words that made the two men look quite different, with the unfortunate result that the validity of the analysis was destroyed. Indeed, if you take any two extensive prose selections by the same author (even alternate paragraphs from the same book) and experiment with ways of grouping "first words" into five groups, you can invariably produce the appearance of two quite dissimilar styles! This is discussed further in a review of the source of Table 7.6, (S. M. Stigler, Review of Authorship Puzzles in the History of Economics: A Statistical Approach, by D. P. O'Brien and A. C. Darnell, in Journal of Economic History, June 1983, Vol. 43, pp. 547-550).

The data of Table 7.7 show a strong beneficial influence of vaccination; $\chi^2 = 214.06$ (16 df) and $P\approx 0$. No doubt the conclusion that vaccination is beneficial is correct, but there is potentially serious "confounding" here. These data are confined to people who have experienced some form of attack, and the ages of the people are not given. Given the time of the study we would expect a strong relationship between age and years-since-vaccination (no young people could have been vaccinated 45 years before the attack, and it is conceivable that older people are over-represented in the unvaccinated group). Modern studies would include age in the analysis and would try to use longitudinal data – data on the same people followed through time.

References

Coren, Stanley, and Clare Porac (1977). Fifty Centuries of Right-Handedness: The Historical Record. Science 198: 631-632.

Fienberg, Stephen E. (1971). Randomization and Social Affairs: The 1970 Draft Lottery. Science 171: 255-261.

Galton, Francis (1886). Regression towards Mediocrity in Hereditary Stature. Journal of the Anthropological Institute 15:246-263.

Galton, Francis (1889). Natural Inheritance. London: Macmillan.

Pearson, Karl (1910). On a New Method of Determining Correlation, When One Variable is Given by Alternative and the Other by Multiple Correlation. Biometrika 7: 248-257.