

2

Two-Dimensional Tables

2.1 Two Binomials

We often wish to compare the relative frequency of occurrence of some characteristic for two groups. In a review of the evidence regarding the therapeutic value of ascorbic acid (vitamin C) for treating the common cold, Pauling [1971] describes a 1961 French study involving 279 skiers during two periods of 5–7 days. The study was double-blind with one group of 140 subject receiving a placebo while a second group of 139 received 1 gram of ascorbic acid per day. Of interest is the relative occurrence of colds for the two groups, and Table 2-1 contains Pauling's reconstruction of these data.

If P_1 is the probability of a member of the placebo group contracting a cold and P_2 is the corresponding probability for the ascorbic acid group, then we are interested in testing the hypothesis that $P_1 = P_2$. The observed numbers of colds in the two groups, $x_{11} = 31$ and $x_{21} = 17$ respectively, are observations on independent binomial variates with probabilities of success P_1 and P_2 and sample sizes $n_1 = 140$ and $n_2 = 139$. The difference in observed proportions,

$$\bar{P}_1 - \bar{P}_2 = \frac{x_{11}}{n_1} - \frac{x_{21}}{n_2},$$

has mean $P_1 - P_2$ and variance

$$\frac{P_1(1 - P_1)}{n_1} + \frac{P_2(1 - P_2)}{n_2}.$$

Table 2-1

Incidence of Common Colds in a Double-Blind Study Involving 279 French Skiers (Pauling [1971])

(a) Observed values

		Cold	No Cold	Totals
Treatment	Placebo	31	109	140
	Ascorbic Acid	17	122	139
	Totals	48	231	279

(b) Expected values under independence

		Cold	No Cold	Totals
Treatment	Placebo	24.1	115.9	140
	Ascorbic Acid	23.9	115.1	139
	Totals	48	231	279

If $P_1 = P_2$, then we could estimate the common value by

$$\bar{P} = \frac{\text{total no. of colds}}{n_1 + n_2} \quad (2.1)$$

and the estimated variance of $\bar{P}_1 - \bar{P}_2$ by

$$\bar{P}(1 - \bar{P}) \left(\frac{1}{n_1} + \frac{1}{n_2} \right). \quad (2.2)$$

Assuming that the hypothesis $P_1 = P_2$ is correct, a reasonable test can be based on the approximate normality of the standardized deviate

$$z = \frac{\bar{P}_1 - \bar{P}_2}{\sqrt{\bar{P}(1 - \bar{P}) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}. \quad (2.3)$$

For our example we get

$$z = \frac{\frac{31}{140} - \frac{17}{139}}{\sqrt{\frac{48}{279} \times \frac{231}{279} \times \left(\frac{1}{140} + \frac{1}{139} \right)}} = 2.19,$$

a value that is significant at the 0.05 level. If we take these data at face value, then we would conclude that the proportion of colds in the vitamin C group is smaller than that in the placebo group. This study, however, has a variety of severe shortcomings (e.g., the method of allocation is not specified and the evaluation of symptoms was largely subjective). For a further discussion of these data, and for a general review of the studies examining the efficacy of vitamin C as a treatment for the common cold up to 1974, see Dykes and Meier [1975].

As an alternative to using the normal approximation to the two-sample binomial problem, we could use the Pearson chi-square statistic (see Pearson [1900a]),

$$X^2 = \sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}, \quad (2.4)$$

where the summation is over all four cells in Table 2-1. We obtain the expected values by estimating $P_1 = P_2 = P$ (the null value) as $\bar{P} = 48/279$; that is, we multiply the two sample sizes n_1 and n_2 by \bar{P} , obtaining the expected values for the (1, 1) and (2, 1) cells, and then get the other two expected values by subtraction. Table 2-1b shows these expected values, and on substituting the observed and expected values in expression (2.4) we get $X^2 = 4.81$, a value that may be referred to a χ^2 distribution with 1 d.f. (degree of freedom). A large

value of X^2 corresponds to a value in the right-hand tail of the χ^2 distribution and is indicative of a poor fit. Rather than using the χ^2 table we note that the square root of 4.81 is 2.19, the value of our z -statistic computed earlier. Some elementary algebra shows that, in general, $z^2 = X^2$. If we set $x_{12} = n_1 - x_{11}$ and $x_{22} = n_2 - x_{21}$, then

$$\begin{aligned} z^2 &= \frac{\left(\frac{x_{11}}{n_1} - \frac{x_{21}}{n_2}\right)^2}{\left(\frac{x_{11} + x_{21}}{n_1 + n_2}\right)\left(\frac{x_{12} + x_{22}}{n_1 + n_2}\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \\ &= \frac{[x_{11}(n_2 - x_{21}) - x_{21}(n_1 - x_{11})]^2(n_1 + n_2)}{(x_{11} + x_{21})(x_{12} + x_{22})n_1n_2} \end{aligned} \quad (2.5)$$

and

$$\begin{aligned} X^2 &= \frac{\left[x_{11} - n_1\left(\frac{x_{11} + x_{21}}{n_1 + n_2}\right)\right]^2}{n_1\left(\frac{x_{11} + x_{21}}{n_1 + n_2}\right)} + \frac{\left[x_{12} - n_1\left(\frac{x_{12} + x_{22}}{n_1 + n_2}\right)\right]^2}{n_1\left(\frac{x_{12} + x_{22}}{n_1 + n_2}\right)} \\ &\quad + \frac{\left[x_{21} - n_2\left(\frac{x_{11} + x_{21}}{n_1 + n_2}\right)\right]^2}{n_2\left(\frac{x_{11} + x_{21}}{n_1 + n_2}\right)} + \frac{\left[x_{22} - n_2\left(\frac{x_{12} + x_{22}}{n_1 + n_2}\right)\right]^2}{n_2\left(\frac{x_{12} + x_{22}}{n_1 + n_2}\right)} \\ &= \frac{[x_{11}(n_2 - x_{21}) - x_{21}(n_1 - x_{11})]^2(n_1 + n_2)}{(x_{11} + x_{21})(x_{12} + x_{22})n_1n_2}. \end{aligned} \quad (2.6)$$

The use of the statistic X^2 is also appropriate for testing for independence in 2×2 tables, as noted in the next section.

Throughout this book we use the Greek quantity χ^2 to refer to the chi-square family of probability distributions, and the Roman quantity X^2 to refer to the Pearson goodness-of-fit test statistic given in general by expression (2.4).

2.2 The Model of Independence

We have just examined a 2×2 table formed by considering the counts generated from two binomial variates. For this table the row totals were fixed

Table 2-2

Counts for Structural Habitat Categories for *sagrei* Adult Male *Anolis* Lizards of Bimini (Schoener [1968])

(a) Observed values

		Perch Diameter (inches)		Totals
		≤ 4.0	> 4.0	
Perch Height (feet)	> 4.75	32	11	43
	≤ 4.75	86	35	121
Totals		118	46	164

(b) Expected values under independence

		Diameter		Totals
		≤ 4.0	> 4.0	
Height	> 4.75	30.9	12.1	43
	≤ 4.75	87.1	33.9	121
Totals		118	46	164

by design. In other circumstances we may wish to consider 2×2 tables where only the total for the table is fixed by design. For example, ecologists studying lizards are often interested in relationships among the variables that can be used to describe the lizard's habitat. Table 2-2a contains data on the habitat of *sagrei* adult male *Anolis* lizards of Bimini, originally reported by Schoener [1968] in a slightly different form. A total of 164 lizards were observed, and for each the perch height (variable 1) and perch diameter (variable 2) were recorded. The two variables were dichotomized partially for convenience and partially because the characteristics of interest for perches are high versus low and wide versus narrow.

Let us denote the observed count for the (i, j) cell by x_{ij} and the totals for the i th row and j th column by x_{i+} and x_{+j} , respectively. Corresponding to this table of observed counts is a table of probabilities,

		Variable 2		Totals
		1	2	
Variable 1	1	p_{11}	p_{12}	p_{1+}
	2	p_{21}	p_{22}	p_{2+}
Totals		p_{+1}	p_{+2}	1

(2.7)

where the probabilities $\{p_{ij}\}$ add to 1, $p_{i+} = p_{i1} + p_{i2}$, and $p_{+j} = p_{1j} + p_{2j}$. In the two-binomial example we wanted to compare two proportions. In the present situation we wish to explore the relationship between the two dichotomous variables corresponding to rows and to columns, that is, to perch height and perch diameter.

If perch height is independent of perch diameter, then

$$\begin{aligned} p_{ij} &= \Pr\{\text{row category} = i \text{ and column category} = j\} \\ &= \Pr\{\text{row category} = i\} \Pr\{\text{column category} = j\} \\ &= p_{i+} p_{+j} \end{aligned} \quad (2.8)$$

for $i = 1, 2$ and $j = 1, 2$. Since only the total sample size N is fixed, $\{x_{ij}\}$ is an observation from a multinomial distribution with sample size N and cell probabilities $\{p_{ij}\}$. The expected value of x_{ij} (viewed as a random variable) is $m_{ij} = Np_{ij}$, and under the model of independence $m_{ij} = Np_{i+}p_{+j}$. Finally, if we substitute the observed row proportion x_{i+}/N for p_{i+} and the observed column proportion x_{+j}/N for p_{+j} , we get the well-known formula for the estimated expected value in the (i, j) cell:

$$\hat{m}_{ij} = x_{i+}x_{+j}/N. \quad (2.9)$$

Table 2-2b displays the estimated expected values for the lizard data (assuming that perch height is independent of perch diameter). We can then test this hypothesis of independence using the Pearson chi-square statistic of expression (2.4):

$$\begin{aligned} X^2 &= \sum_{i=1}^2 \sum_{j=1}^2 \frac{(x_{ij} - x_{i+}x_{+j}/N)^2}{x_{i+}x_{+j}/N} \\ &= \frac{N(x_{11}x_{22} - x_{12}x_{21})^2}{x_{1+}x_{2+}x_{+1}x_{+2}}. \end{aligned} \quad (2.10)$$

For the data in Table 2-2 $X^2 = 0.18$, and comparing this value with a table for the χ^2 distribution with 1 d.f., such as in Appendix III, we conclude that the model of independence of perch height and perch diameter fits the data quite well.

Table 2-3a presents a set of data similar to that in Table 2-2a, but for a different species of *Anolis* lizards, *distichus*, and a different size, adults and subadults. Were perch height and perch diameter independent here, we would expect the entries in the table to be as in Table 2-3b. The model of independence, if applied to the data in Table 2-3a, yields $X^2 = 1.83$, and again the model fits the data quite well.

Table 2-3

Counts for Structural Habitat Categories for *distichus* Adult and Subadult *Anolis* Lizards of Bimini (Schoener [1968])

(a) Observed values

		Perch Diameter (inches)		Totals
		≤ 4.0	> 4.0	
Perch Height (feet)	> 4.75	61	41	102
	≤ 4.75	73	70	143
	Totals	134	111	245

(b) Expected values under independence

		Diameter		Totals
		≤ 4.0	> 4.0	
Height	> 4.75	55.8	46.2	102
	≤ 4.75	78.2	64.8	143
	Totals	134	111	245

We note that X^2 as given by expression (2.10) is the same as \dot{X}^2 in expression (2.6), provided we equate the totals in the corresponding tables, that is, provided we set

$$n_1 = x_{1+}, \quad n_2 = x_{2+}, \quad \text{and} \quad N = n_1 + n_2.$$

2.3 The Loglinear Model

In the two-dimensional tables just examined, the estimated expected values were of the form

$$\hat{m}_{ij} = \frac{x_{i+}x_{+j}}{N}, \quad i = 1, 2, \quad j = 1, 2, \quad (2.11)$$

both under the model of independence of row and column variables and under the model of equality of binomial proportions (more generally referred to as the model of *homogeneity of proportions*). The hat in expression (2.11) is a reminder that m_{ij} is a parameter being estimated by \hat{m}_{ij} .

Taking the natural logarithm of both sides of equation (2.11),

$$\log \hat{m}_{ij} = \log x_{i+} + \log x_{+j} - \log N, \quad (2.12)$$

and thinking in terms of an $I \times J$ table with I rows and J columns reveals a

close similarity to analysis-of-variance notation. Indeed, the additive form suggests that the parameter m_{ij} be expressed in the form

$$\log m_{ij} = u + u_{1(i)} + u_{2(j)}, \quad (2.13)$$

where u is the grand mean of the logarithms of the expected counts,

$$u = \frac{1}{IJ} \sum_{i=1}^I \sum_{j=1}^J \log m_{ij}, \quad (2.14)$$

$u + u_{1(i)}$ is the mean of the logarithms of the expected counts in the J cells at level i of the first variable,

$$u + u_{1(i)} = \frac{1}{J} \sum_{j=1}^J \log m_{ij}, \quad (2.15)$$

and, similarly,

$$u + u_{2(j)} = \frac{1}{I} \sum_{i=1}^I \log m_{ij}. \quad (2.16)$$

Because $u_{1(i)}$ and $u_{2(j)}$ represent deviations from the grand mean u ,

$$\sum_i u_{1(i)} = \sum_j u_{2(j)} = 0. \quad (2.17)$$

In the case of the model of homogeneity of proportions, $n_i = x_{i+}$ is fixed, and thus

$$\log \hat{m}_{ij} = -\log(n_1 + n_2) + \log n_i + \log x_{+j}. \quad (2.18)$$

The term $u_{1(i)}$ in the corresponding model of the form (2.13) is fixed by the sample design and is not otherwise directly interpretable.

It is rarely the case that the model of independence fits as well as it did in the two examples in the preceding section, and this result is all the more surprising in that the frequency of wide perches tends to decrease with increasing perch height in many natural habitats. If we think in terms of perch height and perch diameter interacting, we can add an “interaction term” to the independence model, yielding

$$\log m_{ij} = u + u_{1(i)} + u_{2(j)} + u_{12(ij)}, \quad (2.19)$$

where, in addition to (2.17), we have

$$\sum_{i=1}^I u_{12(ij)} = \sum_{j=1}^J u_{12(ij)} = 0. \quad (2.20)$$

For the example on vitamin C and the common cold, a term $u_{12(ij)}$ would represent the fact that the two binomial proportions are not equal. The model given by expressions (2.19), (2.17), and (2.20) is the most general one for the two-dimensional table.

2.4 Sampling Models

There are three commonly encountered sampling models that are used for the collection of counted cross-classified data:

- (1) *Poisson*: We observe a set of Poisson processes, one for each cell in the cross-classification, over a fixed period of time, with no a priori knowledge regarding the total number of observations to be taken. Each process yields a count for the corresponding cell (see Feller [1968]). The use of this model for contingency tables was first suggested by Fisher [1950].
- (2) *Multinomial*: We take a fixed sample of size N and cross-classify each member of the sample according to its values for the underlying variables. This was the model assumed for the lizard examples of Section 2.3.
- (3) *Product-Multinomial*: For each category of the row variable we take a (multinomial) sample of size x_{i+} and classify each member of the sample according to its category for the column variable (the roles of rows and columns can be interchanged here). This was the sampling model used in the example on vitamin C and the common cold.

A more technical discussion of these sampling models will be given after multidimensional tables have been considered. The basic result of interest here is that the three sampling schemes lead to the same estimated expected cell values and the same goodness-of-fit statistics. (For a more general statement of this result, see the theoretical summary in Appendix II.)

Some stress has been placed on the sampling model in this section, but it is equally important to distinguish between response variables and explanatory variables. For the example of Table 2-1, which has a product-binomial sampling model, treatment (placebo or vitamin C) is an explanatory variable and the occurrence of colds is the response variable. For the two lizard examples, both perch height and perch diameter can be viewed as response variables, and the sampling model is approximately Poisson. It is also possible to have a Poisson or a multinomial sampling model with one response and one explanatory variable. For example, suppose we take a random sample of husband-wife pairs and cross-classify them by voting behavior (e.g., liberal or conservative) in the last election, with the husband's voting behavior as the row variable and the wife's voting behavior as the column variable. If we wish to assess the effect of the husband's behavior

Table 2-4

Piano Choice of Soloists Scheduled for the 1973–1974 Concert Season, for Selected Major American Orchestras

Orchestra	Piano Choice		Totals
	Steinway	Other	
Boston Symphony	4	2	6
Chicago	13	1	14
Cleveland	11	2	13
Minnesota	2	2	4
New York Philharmonic	9	2	11
Philadelphia	6	0	6
Totals	45	9	54

on the wife, for example, the wife's behavior is a response variable and the husband's behavior an explanatory variable. If we condition on the husband's behavior, we go from a multinomial sampling model to a situation based on a product-multinomial model.

It must be noted that not all two-dimensional tables are necessarily generated by one of the three sampling models listed above. Table 2-4 presents data on the piano choice of soloists scheduled during the 1973–1974 concert season, for selected major American orchestras. For these data the basic sampling unit is the soloist; a given soloist may appear with several orchestras during a concert season, however, and in all appearances he or she will use the same brand of piano. Indeed, the total of 9 for “other” can probably be attributed to two or three pianists who use Baldwin pianos.

2.5 The Cross-Product Ratio and 2×2 Tables

In the examples above $I = J = 2$, and, because of the constraints,

$$u_{12(11)} = -u_{12(12)} = -u_{12(21)} = u_{12(22)}. \quad (2.21)$$

Thus there is, in effect, one parameter to measure interaction. When this parameter is set to zero, we get the one degree of freedom associated with the chi-square tests for independence or homogeneity of proportions. For 2×2 tables we can show that

$$u_{12(11)} = \frac{1}{4} \log \alpha, \quad (2.22)$$

where

$$\alpha = \frac{m_{11}m_{22}}{m_{12}m_{21}}. \quad (2.23)$$

We can also write

$$\alpha = \frac{p_{11}p_{22}}{p_{12}p_{21}}, \quad (2.24)$$

since the expected value for the (i, j) cell is just the probability associated with that cell times the sample size:

$$m_{ij} = Np_{ij}. \quad (2.25)$$

This relationship between expected values and cell probabilities follows from results related to the sampling models considered above.

The quantity α defined by (2.23) or (2.24) is usually referred to as the *cross-product ratio* (see Mosteller [1968]) or *odds-ratio*, and it is a basic measure of association in 2×2 tables, in part because it is appropriate for all three types of sampling models described in Section 2.4. Surprisingly, this measure appears only rarely in social science research literature, although it is widely used in chemical, genetic, and medical contexts (see, e.g., Anderson and Davidovits [1975], Kimura [1965], and Cornfield [1956]). The cross-product ratio α has several desirable properties:

- (1) It is invariant under the interchange of rows or columns (except for its “sign”: if we interchange rows or columns but not both, the sign of $\log \alpha$ changes).
- (2) It is invariant under row and column multiplications: Suppose we multiply the probabilities in row 1 by $r_1 > 0$, row 2 by $r_2 > 0$, column 1 by $c_1 > 0$, and column 2 by $c_2 > 0$, and then renormalize these values so that they once again add to 1. The normalizing constant cancels out and we get

$$\alpha' = \frac{(r_1 c_1 p_{11})(r_2 c_2 p_{22})}{(r_1 c_2 p_{12})(r_2 c_1 p_{21})} = \frac{p_{11}p_{22}}{p_{12}p_{21}} = \alpha. \quad (2.26)$$

- (3) Clear interpretation: If we think of row totals as fixed, then p_{11}/p_{12} is the odds of being in the first column given that one is in the first row, and p_{21}/p_{22} is the corresponding odds for the second row. The relative odds for the two rows, or the odds-ratio, is then

$$\frac{p_{11}/p_{12}}{p_{21}/p_{22}} = \frac{p_{11}p_{22}}{p_{12}p_{21}} = \alpha. \quad (2.27)$$

- (4) It can be used in $I \times J$ tables (and multidimensional tables) either through a series of 2×2 partitionings or by looking at several 2×2 subtables.

The quantity α runs from 0 to ∞ and is symmetric in the sense that two values of the cross-product ratio α_1 and α_2 such that $\log \alpha_1 = -\log \alpha_2$ represent the same degree of association, although in opposite directions.

If $\alpha = 1$, the variables corresponding to rows and columns are independent; if $\alpha \neq 1$, they are dependent or associated.

The observed cross-product ratio,

$$\hat{\alpha} = \frac{x_{11}x_{22}}{x_{12}x_{21}}, \quad (2.28)$$

is the maximum-likelihood estimate of α for all three types of sampling models considered above.

For the data in Tables 2-2a and 2-3a we estimate α as $\hat{\alpha}_1 = 1.18$ and $\hat{\alpha}_2 = 1.42$, respectively. The estimate of the large-sample standard deviation of $\log \hat{\alpha}$ is

$$s_{\alpha} = \sqrt{\frac{1}{x_{11}} + \frac{1}{x_{12}} + \frac{1}{x_{21}} + \frac{1}{x_{22}}}, \quad (2.29)$$

and we can use $\log \hat{\alpha}$ and s_{α} to get confidence intervals for $\log \alpha$, since for large samples $\log \hat{\alpha}$ is normally distributed with mean $\log \alpha$. Also, we can use the statistic

$$X^2 = \frac{(\log \hat{\alpha})^2}{s_{\alpha}^2} = \frac{(\log x_{11} + \log x_{22} - \log x_{12} - \log x_{21})^2}{\left(\frac{1}{x_{11}} + \frac{1}{x_{12}} + \frac{1}{x_{21}} + \frac{1}{x_{22}}\right)} \quad (2.30)$$

as an alternative to the usual chi-square statistic (2.10) to test for independence in a 2×2 table. Several authors have suggested alternative ways of getting confidence intervals and tests of significance for α or its logarithm. For an excellent review and discussion of these methods, see Gart [1971].

Goodman and Kruskal [1954, 1959, 1963, 1972] discuss various measures of association for two-dimensional tables, and in the 2×2 case a large number of these are simply monotone functions of the cross-product ratio. For example, Yule's [1900] Q can be written as

$$\begin{aligned} \hat{Q} &= \frac{x_{11}x_{22} - x_{12}x_{21}}{x_{11}x_{22} + x_{12}x_{21}} \\ &= \frac{\hat{\alpha} - 1}{\hat{\alpha} + 1}. \end{aligned} \quad (2.31)$$

Pielou [1969] and others have been critical of such measures because they preclude the distinction between complete and absolute association. For example, Table 2-5 contains two cases in which $Q = 1$ (and $\alpha = \infty$). Pielou's claim is that any ecologist would assert that the association in Table 2-5b

Table 2-5
Examples of Complete and Absolute Association

(a) Complete association

		Species B		Totals
		Present	Absent	
Species A	Present	60	20	80
	Absent	0	20	20
	Totals	60	40	100

(b) Absolute association

		Species B		Totals
		Present	Absent	
Species A	Present	80	0	80
	Absent	0	20	20
	Totals	80	20	100

is greater by far than the association in Table 2-5a, and thus Q (or α) is not the most desirable measure in ecological contexts. Of course, measures that are not monotonic functions of α are tied into the relative values of the marginal totals, and this too is undesirable. For a further discussion of this point, see the papers by Goodman and Kruskal, or Bishop, Fienberg, and Holland [1975].

Returning to model (2.19) for the 2×2 table, it is of interest that the remaining two parameters, $u_{1(1)}$ ($= -u_{1(2)}$) and $u_{2(1)}$ ($= -u_{2(2)}$), are not simply functions of the row and column marginal totals, as one might have expected. In fact,

$$u_{1(1)} = \frac{1}{4} \log \frac{m_{11}m_{12}}{m_{21}m_{22}} \quad (2.32)$$

and

$$u_{2(1)} = \frac{1}{4} \log \frac{m_{11}m_{21}}{m_{12}m_{22}}. \quad (2.33)$$

All the subscripted parameters in the general loglinear model for two-dimensional tables can thus be expressed as functions of various cross-product-ratio-like terms. Nevertheless, α and the marginal totals $\{m_{i+}\}$ and $\{m_{+j}\}$ completely determine the 2×2 table.

2.6 Interrelated Two-Dimensional Tables

As we mentioned in Chapter 1, a typical analysis of a multidimensional cross-classification often consists of the analysis of two-dimensional marginal totals. Moreover, we noted that such analyses have severe shortcomings. To illustrate this point we now consider the analysis of a three-dimensional

Table 2-6
Two-Dimensional Marginal Totals of Three-Dimensional Cross-Classification of 4353 Individuals (See Table 3-6)

(a) Occupational group vs. educational level

	(low) E1	E2	E3	(high) E4	Totals
O1	239	309	233	53	834
O2	6	11	70	199	286
O3	1	7	12	215	235
O4	794	781	922	501	2998
Totals	1040	1108	1237	968	4353

- O1 = self-employed, business
- O2 = self-employed, professional
- O3 = teacher
- O4 = salaried, employed

(b) Aptitude vs. occupational level

	O1	O2	O3	O4	Totals
(low) A1	122	30	20	472	644
A2	226	51	66	704	1047
A3	306	115	96	1072	1589
A4	130	59	38	501	728
(high) A5	50	31	15	249	345
Totals	834	286	235	2998	4353

(c) Aptitude vs. educational level

	E1	E2	E3	E4	Totals
A1	215	208	138	83	644
A2	281	285	284	197	1047
A3	372	386	446	385	1589
A4	128	176	238	186	728
A5	44	53	131	117	345
Totals	1040	1108	1237	968	4353

example solely in terms of its two-dimensional marginal totals. Later we shall reconsider this example and contrast the conclusions drawn from the two analyses.

Table 2-6a contains data on the cross-classification of 4353 individuals into four occupational groups (O) and four educational levels (E). These data have been extracted from a larger study encompassing many more occupational groups. The Pearson chi-square test for independence of occupation and education yields $X^2 = 1254.1$ with 9 d.f., so that occupation and education are clearly related.

Tables 2-6b and 2-6c present two additional two-dimensional cross-classifications of the same 4353 individuals, the first by aptitude (as measured at an earlier date by a scholastic aptitude test) and occupation, the second by aptitude and education. Testing for independence in these tables yields, for Table 2-6b, $X^2 = 35.8$ with 12 d.f., and for Table 2-6c, $X^2 = 178.6$ also with 12 d.f. Both of these values are highly significant when referred to the corresponding chi-square distribution, and we are forced to conclude that occupation, aptitude, and education are pairwise related in all possible ways.

Now, in what ways is this analysis unsatisfactory? Everything we have done appears to be correct. We have just not done enough, and we have not looked at the data in the best possible way. Before we can complete the discussion of this example, however, we must consider details of the analysis of three-dimensional tables.

2.7 Correction for Continuity

The practice of comparing the Pearson chi-square statistic given by expression (2.4) to the tail values of the χ^2 distribution with the appropriate degrees of freedom is an approximation that is appropriate only when the overall sample size N is large. Yates [1934] suggested that a correction be applied to X^2 , the *correction for continuity*, for 2×2 tables to make the tail areas correspond to those of the hypergeometric distribution (used when both row and column margins are fixed). Thus, instead of using the chi-square formula (2.10), many introductory texts suggest using the “corrected” statistic:

$$\begin{aligned}
 X_c^2 &= \sum_{i=1}^2 \sum_{j=1}^2 \frac{[|x_{ij} - (x_{i+}x_{+j}/N)| - 1/2]^2}{x_{i+}x_{+j}/N} \\
 &= \frac{N(|x_{11}x_{22} - x_{21}x_{12}| - N/2)^2}{x_{1+}x_{2+}x_{+1}x_{+2}}
 \end{aligned} \tag{2.34}$$

Cox [1970b] provides an elementary analytical derivation of the continuity correction that is applicable in this case. Rao [1973, p. 414] provides a slightly different method for correcting X^2 .

If, however, our aim is to correct the statistic X^2 so that it more closely adheres to the large-sample χ^2 distribution, rather than to the hypergeometric distribution, then the use of the corrected chi-square statistic (2.34) may not necessarily be appropriate. In fact, Plackett [1964], Grizzle [1967], and Conover [1974] have shown that using X_c^2 in place of X^2 results in an overly conservative test, one that rejects the null hypothesis too rarely relative to the nominal level of significance. For example, in 500 randomly generated multinomial data sets examined by Grizzle, with $N = 40$ and cell probabilities

0.56	0.24
0.14	0.06

the statistic X^2 exceeded the 0.05 level of significance (i.e., 3.84) 26 times (5.2%) whereas the corrected statistic X_c^2 exceeded the 0.05 level about 6 times (1.2%).

Because of this empirical evidence, and because of our use of the χ^2 distribution as the reference distribution for the chi-square statistic X^2 , we make no use of continuity corrections in this book.

Much attention has been given in the statistical literature to the use of Fisher's exact test for 2×2 tables, which is based on the hypergeometric distribution referred to above. Because this book focuses primarily on multi-dimensional tables, and because the extension of the exact test even to hypotheses regarding models for $2 \times 2 \times 2$ tables is complicated at best, we have chosen not to consider exact tests at all. For a recent debate on the appropriateness of the exact test for 2×2 tables, the interested reader is referred to Berkson [1978] and Kempthorne [1979]. Haberman [1974a] presents a detailed discussion of exact tests in several dimensions, under the heading of conditional Poisson models (see also Plackett [1974, pp. 83–87]).

2.8 Other Scales for Analyzing Two-Dimensional Tables

The discussion in this monograph is focused on models for cell probabilities or expected values which are linear in the logarithmic scale. Other possible scales of interest are the linear scale and those based on the angular (arc sine) or integrated normal (probit) transforms. Another possibility is a model that is linear in the logistic scale; that is, if p is a probability, its logistic transform

is $\log(p/(1-p))$. In Chapter 6 the relationship between loglinear and linear logistic models is discussed in detail.

Cox [1970a, pp. 26–29] examines the linearizing transformations from $p(x)$, the probability of success at level x , to the following scales:

$$(1) \text{ logistic: } \log \left(\frac{p(x)}{1-p(x)} \right), \quad (2.35)$$

$$(2) \text{ linear: } p(x), \quad (2.36)$$

$$(3) \text{ integrated normal: } \phi^{-1}(p(x)), \text{ where} \quad (2.37)$$

$$\phi(u) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^u e^{-y^2/2} dy;$$

$$(4) \text{ arc sine: } \sin^{-1}(\sqrt{p(x)}). \quad (2.38)$$

Cox notes that all four transformations are in reasonable agreement when the probability of success is in the range 0.1–0.9, and that in the middle of this range analyses in terms of each of the four are likely to give virtually equivalent results.

For further details the interested reader is referred to Cox [1970a] and to the discussion in Bishop, Fienberg, and Holland [1975, pp. 368–369], which includes an example of a $2 \times 2 \times 2 \times 2$ table analyzed using both a loglinear or logistic model and an ANOVA model following an arc sine transformation. Hewlett and Plackett [1979] present an elementary analysis of biological data involving quantal responses using both the probit transformation of (2.37) and the logistic transformation of (2.35). Haberman [1978, pp. 344–346] also provides a helpful discussion of computations for probit models.

The remainder of this book focuses on models and analyses based on the logarithmic or logistic scale.

Problems

2.1 (Armitage [1955]). In a study of children aged 0 to 15, concern was focused on the presence or absence of the carrier for *Streptococcus pyogenes*, and the relationship between the presence of the carrier and tonsil size (Table 2-7).

Table 2-7

	Tonsils present, not enlarged	Tonsils enlarged + + +	
Carriers	19	29	24
Total	516	589	293
Proportion of carriers	0.0368	0.0492	0.0819

- Test the hypothesis of homogeneity of proportions for the carriers and noncarriers.
- Compute $\hat{\alpha}$ for columns 1 and 2, and for columns 2 and 3.
- What alternative to the equal-proportions hypothesis is suggested by the data?
- What other possible explanatory variables might be relevant to the interpretation of the data?

2.2 (Reiss [1980]). The data in Table 2-8, from the National Crime Survey, represent repeat-victimization information, i.e., successive pairs of victimizations for households in the survey. If, over the course of the period in question, a household is victimized m times, then these m victimizations will lead to $m - 1$ entries in the table, one for each successive pair of victimizations. (Here A is assault; B, burglary; HL, household larceny; MV, motor vehicle theft; PL, personal larceny; PP/PS, pocket picking and purse snatching; Ra, rape; Ro, robbery.)

Table 2-8

		Second Victimization in Pair								Totals
		Ra	A	Ro	PP/PS	PL	B	HL	MV	
First Victimi- zation in Pair	Ra	26	50	11	6	82	39	48	11	273
	A	65	2997	238	85	2553	1083	1349	216	8586
	Ro	12	279	197	36	459	197	221	47	1448
	PP/PS	3	102	40	61	243	115	101	38	703
	PL	75	2628	413	229	12,137	2658	3689	687	22,516
	B	52	1117	191	102	2649	3210	1973	301	9595
	HL	42	1251	206	117	3757	1962	4646	391	12,372
	MV	3	221	51	24	678	301	367	269	1914
Totals		278	8645	1347	660	22,558	9565	12,394	1960	

- (a) Test these data for homogeneity of row proportions.
- (b) Examine a table of the square roots of the contribution from each cell to X^2 , i.e., $(\text{Observed} - \text{Expected})/\sqrt{\text{Expected}}$. Which cells stand out?
- (c) Provide an interpretation for your findings in part (b).
- (d) Why might you not have bothered to do the full calculation of X^2 after you found the contribution from the (1, 1) cell?

2.3. The 1970 draft lottery was intended to provide a sequence of birthdates which was to be used to select males between the ages of 19 and 26 for induction into the armed forces of the United States. Each day of the year (including February 29) was typed on a slip of paper and inserted into a capsule. The capsules were mixed and were assigned a “drawing number” according to their position in the sequence of capsules picked from a bowl. The data in Table 2-9, from Fienberg [1971], summarize the results of the drawing. The dates are grouped by month, and the drawing numbers by thirds. Is there evidence in the table to cast suspicion on the randomness of the sequence of drawing numbers?

Table 2-9

Drawing numbers	Months												Totals
	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.	
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

2.4 (Fienberg [1980]). In a study of citation practices in the field of operations research, 336 articles were examined from the 1969 and 1970 issues of two journals, *Management Science* and *Operations Research*. The 336 articles have been cross-classified in Table 2-10 according to the journal in which the article appeared and the location of other articles referenced.

Table 2-10

		Cited Journal			
		Neither	MS only	OR only	Both
Citing Journal	MS	61	63	20	59
	OR	42	3	47	41

- (a) Which variables in this table are explanatory and which can be thought of as response variables?
- (b) Viewing these data as if they were a sample from a hypothetically infinite population, test the hypothesis of homogeneity of proportions corresponding to the structure identified in (a). Do citation practices differ for the two journals?
- (c) Interpret your findings.

2.5 (Yule [1900]). The data in Table 2-11 for 205 married persons, reported initially by Galton, give the number of cases in which a tall, medium, or short man was mated with a tall, medium, or short woman.

Table 2-11

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

- (a) Test the hypothesis that the heights of husbands and wives are independent.
- (b) Form two 2×2 tables from this 3×3 table contrasting tall and medium versus short, and tall versus medium and short. Test the hypothesis of independence for each of these 2×2 tables.
- (c) Compute Yule's Q for each of the 2×2 tables in part (b).
- (d) Assign "scores" of 1, 2, and 3 to the categories short, medium, and tall. Then compute the observed correlation between the heights of husbands and wives based on these scores.
- (e) By interpreting the quantities computed in parts (b), (c), and (d), or otherwise, discuss how the heights of husbands and wives are associated.

The Analysis of Cross-Classified Categorical Data

Fienberg, S.E.

2007, XIV, 198 p., Softcover

ISBN: 978-0-387-72824-7