

Chapter 14

Discrete Relationships

Three aquifers are sampled to determine whether they differ in their concentrations of copper. In all three, over 40 percent of the samples were below the detection limit. What methods will test whether the distribution of copper is identical in the three aquifers while effectively incorporating data below the detection limit?

Counts of three macroinvertebrate species were measured in three stream locations to determine ecosystem health. The three species cover the range of tolerance to pollution, so that a shift from dominance of one species to another is an indication of likely contamination. Do the three locations differ in their proportions of the three species, or are they identical?

This chapter presents methods to evaluate the relationship between two discrete (also called categorical) variables. The tests are analogous to analysis of variance or t-tests where the response variable is not continuous -- it is recorded only as a discrete number or category (see Figure 4.1). When the response variable is **ordinal** (possible values can be ordered into a logical sequence, such as low, medium and high) the familiar Kruskal-Wallis test can be used. When the response variable is **nominal** (no ordering to the categories, such as with different species of organism), contingency tables can assess association. When both variables are ordinal, Kendall's tau can test for significance in association.

14.1 Recording Categorical Data

Categorical variables are those whose possible values are not along a continuous scale (such as concentration), but may take on only one of a discrete number of values classed into one of several categories. Examples of categorical variables used in water resources studies are: presence or absence of a benthic invertebrate, whether an organic compound is above or below the detection limit, soil type, land use group, and location variables such as aquifer unit, gaging station, etc. To easily inspect the relationship between two categorical variables, the data are recorded as a matrix of counts (Figure 14.1). The matrix is composed of two categorical variables, one assigned to the columns and one to the rows. Both variables will take on several possible values. The entries in the cells of the matrix are the number of observations O_{ij} which fall into the i th row and j th column of the matrix.

<u>Variable 1</u>	<u>Variable 2</u>		
	Group 1	Group 2	Group 3
Response 1	O ₁₁	O ₁₂	O ₁₃
Response 2	O ₂₁	O ₂₂	O ₂₃

Figure 14.1 Structure of a 2-variable matrix

14.2 Contingency Tables (Both Variables Nominal)

Contingency tables measure the association between two nominal categorical variables. Because they are nominal there is no natural ordering of either variable, so that categories may be switched in assignment from the first to the second row, etc. without any loss in meaning. The purpose of contingency table analysis is to determine whether the row classification (variable 1, here arbitrarily assigned to the response variable if there is one) is independent of the column classification (variable 2, here assigned to the location or group-of-origin variable). The null hypothesis H_0 is that the two variables are independent -- that is, the distribution of data among the categories of the first variable is not affected by the classification of the second variable. Evidence may be sufficient to reject H_0 in favor of H_1 : the variables are dependent or related. The statement that one variable causes the observed values for the second variable is not necessarily implied, analogous to correlation. Causation must be determined by knowledge of the relevant processes, not only mathematical association. For example, both variables could be caused by a third underlying variable.

Example 1

Three streams are sampled to determine if they differ in their macrobiological community structure. In particular, the presence or absence of two species are recorded for each stream, one species being pollution tolerant, and one not. If the streams differ in their proportion of pollution-tolerant species, it is inferred that they differ in their pollution sources as well. Test whether the streams are identical in (independent of) the proportion of pollution-tolerant organisms, or whether they differ in this proportion (proportion is dependent on the stream).

H₀: the proportion of one species versus the second is the same for (is independent of) all 3 streams.

H₁: the proportion differs between (is dependent on) the stream.

	Stream 1	Stream 2	Stream 3	
Tolerant	O ₁₁	O ₁₂	O ₁₃	A ₁ = Σ(O ₁₁ +O ₁₂ +O ₁₃)
Intolerant	O ₂₁	O ₂₂	O ₂₃	A ₂ = Σ(O ₂₁ +O ₂₂ +O ₂₃)
	C ₁ =	C ₂ =	C ₃ =	N = (A₁+A₂)
	Σ(O ₁₁ +O ₂₁)	Σ(O ₁₂ +O ₂₂)	Σ(O ₁₃ +O ₂₃)	= (C₁+C₂+C₃)

14.2.1 Performing the Test for Independence

To test for independence, the observed counts O_{ij} (row i and column j) in each cell are summed across rows to produce the row totals A_i, and down columns to produce column totals C_j.

There are m rows (i=1,m) and k columns (j=1,k). The total sample size N is the sum of all counts in every cell, or N = ΣA_i = ΣC_j = ΣO_{ij}. The **marginal probability** of being in a given row (a_i) or column (c_j), can be computed by dividing the row A_i and column C_i totals by N:

$$a_i = A_i/N \qquad c_j = C_j/N$$

If H₀ is true, the probability of a new sample falling into row 1 (species tolerant of pollution) is best estimated by the marginal probability a₁ regardless of which stream the sample was taken from. Thus the marginal probability for a row ignores any influence of the column variable.

The column variable is important in that the number of available samples may differ among the columns. The probability of being in any column may not be (1/no. columns). Therefore, with H₀ true, the best estimate of the **joint probability** e_{ij} of being in a single cell in the table equals the marginal probability of being in row i times the marginal probability of being in column j

$$e_{ij} = a_i \cdot c_j.$$

Finally, for a sample size of N, the expected number of observations in each cell given H₀ is true can be computed by multiplying each joint probability e_{ij} by N:

$$E_{ij} = N a_i c_j, \quad \text{or}$$

$$E_{ij} = \frac{A_i C_j}{N} \quad [14.1]$$

To test H_0 , a test statistic X_{ct} is computed by directly comparing the observed counts O_{ij} with the counts E_{ij} expected when H_0 is true. This statistic is the sum of the squared differences divided by the expected counts, summed over all $i \cdot j$ cells:

$$X_{ct} = \sum_{i=1}^m \sum_{j=1}^k \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \quad [14.2]$$

If H_0 is not true, the observed counts O_{ij} will be very different from the E_{ij} for at least one cell and X_{ct} will then be large. If H_0 is true, the $O_{ij} \cong E_{ij}$ for all $i \cdot j$ cells, and X_{ct} will be small. To evaluate whether X_{ct} is sufficiently large to reject H_0 , the test statistic is compared to the $(1-\alpha)$ quantile of a chi-square distribution having $(m-1) \cdot (k-1)$ degrees of freedom. Tables of the chi-square distribution are available in most statistics texts.

To understand why there are $(m-1) \cdot (k-1)$ degrees of freedom, when the marginal sums A_{ij} and C_{ij} are known, once $(m-1) \cdot (k-1)$ of the cell counts O_{ij} are specified the remainder can be computed. Therefore, only $(m-1) \cdot (k-1)$ entries can be "freely" specified.

Example 1 cont.

For the table of observed counts O_{ij} below, determine whether the streams differ significantly in their proportion of pollutant-tolerant species.

O_{ij}	Stream 1	Stream 2	Stream 3	
Tolerant	4	8	12	$A_1 = 24$
Intolerant	18	12	6	$A_2 = 36$
	$C_1=22$	$C_2=20$	$C_3=18$	$N = 60$

To determine whether the proportion of pollutant-tolerant species is significantly different for the three streams, a table of expected counts E_{ij} assuming H_0 to be true is computed using equation 14.1:

E_{ij}	Stream 1	Stream 2	Stream 3	
Tolerant	8.8	8.0	7.2	A ₁ = 24
Intolerant	13.2	12.0	10.8	A ₂ = 36
	C ₁ = 22	C ₂ = 20	C ₃ = 18	60

Dividing these expected counts by N results in the table of expected probabilities (e_{ij} = E_{ij} / N):

e_{ij}	Stream 1	Stream 2	Stream 3	
Tolerant	.148	.132	.120	a ₁ = 0.4
Intolerant	.222	.198	.180	a ₂ = 0.6
	c ₁ = 0.37	c ₂ = 0.33	c ₃ = 0.30	1.0

To perform the significance test:

$$\begin{aligned}
 X_{ct} &= \frac{(4.0-8.8)^2}{8.8} + \frac{(8-8.0)^2}{8.0} + \frac{(12-7.2)^2}{7.2} + \\
 &\quad \frac{(18-13.2)^2}{13.2} + \frac{(12-12)^2}{12} + \frac{(6-10.8)^2}{10.8} \\
 &= 9.70
 \end{aligned}$$

H_O should be rejected if X_{ct} exceeds the (1-α) quantile of the chi-square distribution with 1•2 = 2 degrees of freedom. For α = 0.05, χ²_(.95, 2) = 5.99. Therefore, H_O is rejected. The proportion of pollutant-tolerant species is not the same in all three streams. Thus the overall marginal probability of 0.4 is not an adequate estimate of the probability of pollution-tolerant species for all three streams.

14.2.2 Conditions Necessary for the Test

The chi-square distribution is a good approximation to the true distribution of X_{ct} as long as

- all E_{ij} > 1 and
- at least 80% of cells have E_{ij} ≥ 5 (Conover, 1980).

If either condition is not met,

- combine two or more rows or columns and recompute, or
- enumerate the exact distribution of X_{ct}. See Conover (1980) for details.

A contingency table test is not capable of extracting the information contained in any natural ordering of rows or columns. Contingency tables are designed to operate on nominal data without this ordering. The columns or rows can be rearranged without changing the expected values E_{ij}, and therefore without altering the test statistic. When the response variable or both variables have a natural scale of ordering, the test statistic should change as the ordinal variable is rearranged. Methods more powerful than contingency tables should be used when one or

both variables are ordinal. When only the response variable is ordinal, the Kruskal-Wallis test of the next section will have more power to see differences between groups than will contingency tables. When both variables are ordinal, Kendall's tau can measure the relationship as shown in section 14.4.

14.2.3 Location Of the Differences

When a contingency table finds an association between the two variables, it is usually of interest to know how the two are related. Which cells are higher or lower in proportion than would be expected had H_0 been true? A guide to this are the individual cell chi-square statistics.

Cells having large values of $\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$ are the cells contributing most to the rejection of the null hypothesis. The sign of the difference between O_{ij} and E_{ij} indicates the direction of the departure. For example, the individual cell chi-square statistics for the species data of example 1 are as follows:

χ^2	Stream 1	Stream 2	Stream 3
Tolerant	$\frac{(4.0-8.8)^2}{8.8} = 2.6$	$\frac{(8-8.0)^2}{8.0} = 0$	$\frac{(12-7.2)^2}{7.2} = 3.2$
Intolerant	$\frac{(18-13.2)^2}{13.2} = 1.7$	$\frac{(12-12)^2}{12} = 0$	$\frac{(6-10.8)^2}{10.8} = 2.1$

Stream 3 has many more counts of the pollution-tolerant species than the number expected if all three streams were alike, and stream 1 has many less. Therefore stream 1 appears to be the least affected by pollution, stream 2 in-between, and stream 3 the most affected.

14.3 Kruskal-Wallis Test for Ordered Categorical Responses

In Chapter 5 the Kruskal-Wallis test was introduced as a nonparametric test for differences in medians among 3 or more groups. The response variable in that case was continuous. In Chapter 13 the test was applied to response data whose lower end of a continuous scale was below a reporting limit. All censored data were treated as ties. Now the test will be applied to data which are ordinal -- the response variable can only be recorded as belonging to one of several ordered categories. All observations in the same response category (row) are tied with each other. The test takes on its most general form in this situation, as a test for whether a shift in the distribution has occurred, rather than as a test for differences in the median of continuous data. The test may be stated as:

- H₀: the proportion of data in each response category (row) is the same for each group (column).
- H₁: the proportion differs among (is dependent on) the groups.

14.3.1 Computing the Test

The data are organized in a matrix identical to that for a contingency table, but the computations at the margins differ (Figure 14.2). Once the row sums A_i are computed, ranks R_i are assigned to each observation in the table in accord with levels of the response variable. Ranks for all observations in the category with the lowest responses (response row 1 in Figure 14.2) will be tied at the average rank for that row, or $\bar{R}_1 = (A_1 + 1)/2$. All observations in the row having the next highest response are also assigned ranks tied at the average of ranks within that row, and so on up to the highest row of responses. For response 2 in the Figure 14.2 the average rank equals $\bar{R}_2 = A_1 + (A_2 + 1)/2$. For any row x of a total of m rows, the average rank will equal

$$\bar{R}_x = \sum_{i=1}^{x-1} A_i + (A_x + 1)/2. \tag{14.3}$$

To determine whether the distribution of proportions differs among the k groups (the k columns), the average column ranks \bar{D}_j are computed as

$$\bar{D}_j = \frac{\sum_{i=1}^m O_{ij} \bar{R}_i}{C_j} \quad \text{where } C_j = \sum_{i=1}^m O_{ij}. \tag{14.4}$$

	Group 1	Group 2	Group 3	
response 1	O ₁₁	O ₁₂	O ₁₃	A ₁ = Σ(O ₁₁ +O ₁₂ +O ₁₃) A ₂ = Σ(O ₂₁ +O ₂₂ +O ₂₃) N
response 2	O ₂₁	O ₂₂	O ₂₃	
	\bar{D}_1	\bar{D}_2	\bar{D}_3	

where

$$\bar{D}_1 = \frac{(O_{11} \bar{R}_1 + O_{21} \bar{R}_2)}{O_{11} + O_{21}} \quad \bar{D}_2 = \frac{(O_{12} \bar{R}_1 + O_{22} \bar{R}_2)}{O_{12} + O_{22}} \quad \bar{D}_3 = \frac{(O_{13} \bar{R}_1 + O_{23} \bar{R}_2)}{O_{13} + O_{23}}$$

Figure 14.2 2x3 matrix for Kruskal-Wallis analysis of an ordered response variable

The Kruskal-Wallis test statistic is then computed from these average group ranks. If H_0 is true, the average ranks \bar{D}_j will all be about the same, and similar to the overall average rank of $(N+1)/2$. If H_0 is not true, the average rank for at least one of the columns will differ. The Kruskal-Wallis test statistic is computed using equation 14.5:

$$K = (N-1) \frac{\sum_{j=1}^k (C_j \bar{D}_j^2) - N \left[\frac{N+1}{2} \right]^2}{\sum_{i=1}^m (A_i \bar{R}_i^2) - N \left[\frac{N+1}{2} \right]^2} \quad [14.5]$$

where C_j is the number of observations in column j ,
 \bar{D}_j is the average rank of observations in column j ,
 A_i is the number of observations in row i , and
 \bar{R}_i is the average rank of observations in row i .

To evaluate its significance, K is compared to a table of the chi-square distribution with $k-1$ degrees of freedom.

Example 2

An organic chemical is measured in 60 wells screened in one of three aquifers. The concentration is recorded only as being either above or below the reporting limit (rl). Does the distribution of the chemical differ among the three aquifers?

First, ranks are assigned to the response variable. There are 36 observations in the lower category (below rl), each with a rank equal to the mean rank of that group. The mean of numbers 1 to 36 is $(36+1)/2 = 18.5$. The higher category contains 24 observations with ranks 37 to 60, so that their mean rank is $36 + (24+1)/2$, or 48.5.

	Aquifer 1	Aquifer 2	Aquifer 3	A_i	\bar{R}_i
below rl	18	12	6	36	18.5
above rl	4	8	12	24	48.5
	$\bar{D}_1 = 527/22$ = 24	$\bar{D}_2 = 610/20$ = 30.5	$\bar{D}_3 = 693/18$ = 38.5		

$$K = (59) \frac{\sum (22 \cdot 24^2 + 20 \cdot 30.5^2 + 18 \cdot 38.5^2) - 60 \left[\frac{61}{2} \right]^2}{\sum (24 \cdot 48.5^2 + 36 \cdot 18.5^2) - 60 \left[\frac{61}{2} \right]^2}$$

$$= 9.75$$

The chi-square statistic $\chi^2_{(.95, 2)} = 5.99$. Thus H_0 is rejected, and the groups are found to have differing percentages of data above the reporting limit.

14.3.2 Multiple Comparisons

Once differences between the groups (columns) have been found, it is usually of interest to determine which groups differ from others. This is done with multiple comparison tests as stated in section 7.4. Briefly, multiple Kruskal-Wallis tests are performed between pairs of columns. After a significant KW test occurs for k groups, place the groups in order of ascending average rank. Perform new KW tests for the two possible comparisons between groups which are $p = (k-1)$ columns apart (the first versus the next-to-last column, and the second versus the last). Note that the observations must be re-ranked for each test. If significant results occur for one or both of these tests, continue attempting to find differences between smaller subsets of any groups found to be significantly different. In order to control the overall error rate, set the individual error rates for each KW test at α_p , below:

$$\begin{aligned}\alpha_p &= 1 - (1-\alpha)^{p/k} && \text{for } p < (k-1) \\ &= \alpha && \text{for } p \geq (k-1)\end{aligned}$$

14.4 Kendall's Tau for Categorical Data (Both Variables Ordinal)

When both row and column variables are ordinal, a contingency table would test for differences in distribution of the row categories among the columns, but would ignore the correlation structure of the data -- do increases in the column variable coincide with increases or decreases in the row variable? This additional information contained in the correlation structure of ordinal variables can be evaluated with a rank correlation test such as Kendall's tau.

14.4.1 Kendall's τ_b for Tied Data

Kendall's correlation coefficient tau (τ) must be modified in the presence of ties. In Chapter 8 a tie modification was given for ties in the response variable only. Now there are many more ties, the ties between all data found in the same row and column of a contingency table. Kendall (1975) called this tie modification τ_b (tau-b).

$$\tau_b = \frac{S}{\frac{1}{2} \sqrt{(N^2 - SS_a)(N^2 - SS_c)}} \quad [14.6]$$

The numerator S for τ_b is $P-M$, just as in Chapter 8, the number of pluses minus the number of minuses. Consider a contingency table structure with the lowest values on the upper left (the

rows are ordered from lowest value on the top to the highest value on the bottom, and the columns from lowest on the left to highest on the right -- see Figure 14.3). With this format, the number of pluses are the number of comparisons with data in cells to the right and below each cell (Figure 14.4). The number of minuses are the number of comparisons with data in cells to the left and below (Figure 14.5). Data in cells of the same row or column do not contribute to S . Therefore, summing over each cell of row x and column y ,

$$S = P - M = \sum_{xy} O_{xy} (\sum O_{\text{southeast}} - \sum O_{\text{southwest}}), \text{ or}$$

$$S = \sum_{\text{all } x \ y} \sum_{i>x} \sum_{j>y} O_{xy} \cdot O_{ij} - \sum_{i<x} \sum_{j<y} O_{xy} \cdot O_{ij} \quad [14.7]$$

The denominator for τ_b is not $(n \cdot n - 1)/2$ as it was for τ , equal to the total number of comparisons. Instead S is divided by the total number of untied comparisons. To compute this efficiently with a contingency table, SS_a and SS_c (the sums of squares of the row and column marginal totals, respectively) are computed as in equation 14.8, and then used in equation 14.6 to compute τ_b .

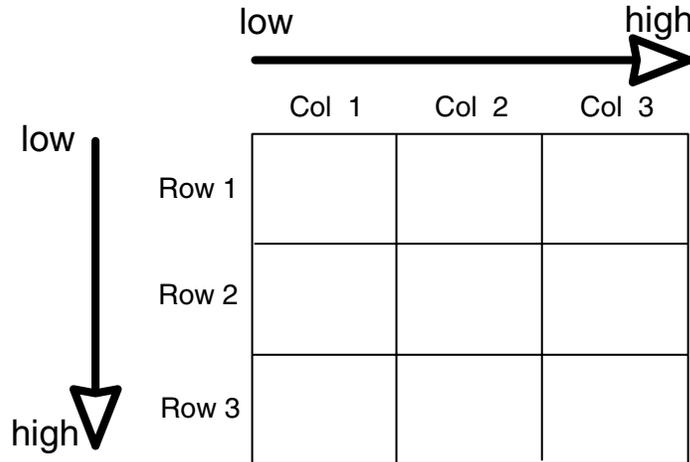


Figure 14.3 Suggested ordering of rows and columns for computing τ_b .

$$SS_a = \sum_{i=1}^m A_i^2 \qquad SS_c = \sum_{j=1}^k C_j^2 \quad [14.8]$$

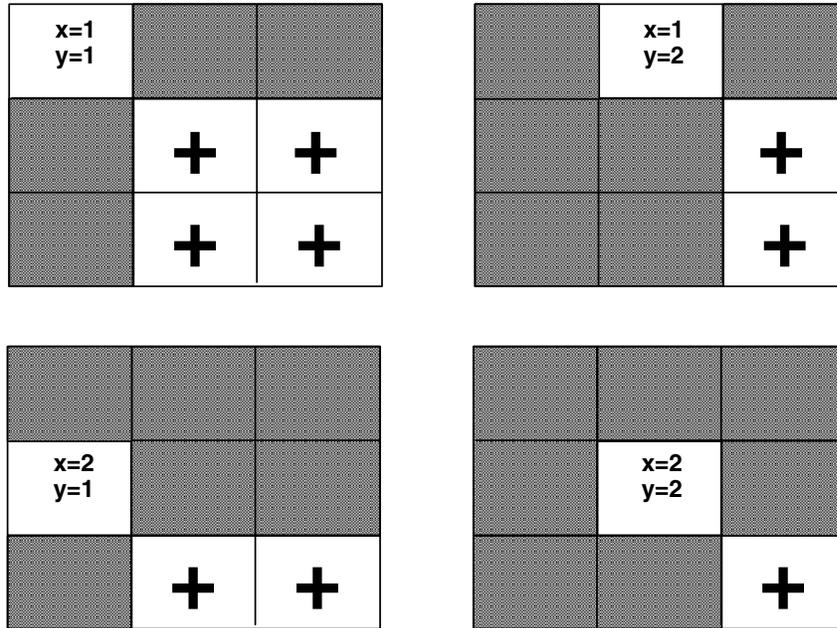


Figure 14.4 3x3 matrix cells contributing to P ($i > x$ and $j > y$).

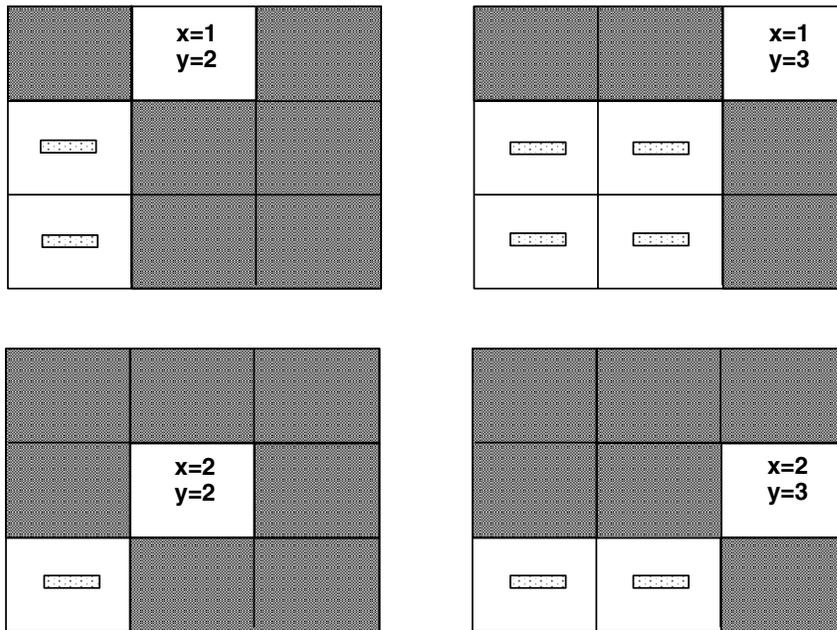


Figure 14.5 3x3 matrix cells contributing to M ($i < x$ and $j < y$).

Example 3

Pesticide concentrations in shallow aquifers were investigated to test whether their distribution was the same for wells located in three soil classes, or whether concentrations differed with increasing soil drainage. The laboratory reported concentrations for the pesticide when levels

were above the reporting limit. The compound was reported only as "present" when concentrations were between the reporting limit and the instrument detection limit (dl), and as "<dl" if concentrations were below the detection limit. Compute Kendall's tau for this data.

Concentration	Soil Drainage			A _i	a _i
	Poor	Moderate	High		
< dl	18	12	7	37	0.47
"present"	5	10	8	23	0.29
> rl	2	6	11	19	0.24
C _j	25	28	26	79	
c _j	0.32	0.35	0.33		1.0

The number of pluses P = 18(10+8+6+11) + 12(8+11) + 5(6+11) + 10(11) = 1053

The number of minuses M = 12(5+2) + 7(5+10+2+6) + 10(2) + 8(2+6) = 329

So S = 1053 - 329 = 724.

To compute the denominator of τ_b ,

$$SS_a = 37^2 + 23^2 + 19^2 = 2259.$$

$$SS_c = 25^2 + 28^2 + 26^2 = 2085.$$

$$\text{and } \tau_b = \frac{724}{\frac{\sqrt{(79^2 - 2259)(79^2 - 2085)}}{2}} = \frac{724}{2034} = 0.36.$$

14.4.2 Test Of Significance for τ_b

To determine whether τ_b is significantly different from zero, S must be divided by its standard error σ_S and compared to a table of the normal distribution, just as in Chapter 8. Agresti (1984) provides the following approximation to σ_S which is valid for P and M larger than 100:

$$\sigma_S \cong \sqrt{\frac{1}{9} \cdot \left(1 - \sum_{i=1}^m a_i^3\right) \left(1 - \sum_{j=1}^k c_j^3\right) \cdot N^3} \quad [14.9]$$

where a_i and c_j are the marginal probabilities of each row and column.

The exact formula for σ_S (Kendall, 1975) is much more complicated. It is the square root of equation 14.10:

$$\sigma_S^2 = \frac{\left(n(n-1)(2n+5) - \sum_{i=1}^m A_i(A_i-1)(2A_i+5) - \sum_{j=1}^k C_j(C_j-1)(2C_j+5) \right)}{18} + \frac{\left(\sum_{i=1}^m A_i(A_i-1)(A_i-2) \right) \left(\sum_{j=1}^k C_j(C_j-1)(C_j-2) \right)}{9 \cdot N(N-1)(N-2)} + \frac{\left(\sum_{i=1}^m A_i(A_i-1) \right) \left(\sum_{j=1}^k C_j(C_j-1) \right)}{2 \cdot N(N-1)} \quad [14.10]$$

If one variable were continuous and contained no ties, equation 14.10 would simplify to the square of equation 8.4.

To test for significance of τ_b , the test statistic Z_S is computed as in Chapter 8:

$Z_S =$	$\begin{cases} \frac{S-1}{\sigma_s} & \text{if } S > 0 \\ 0 & \text{if } S = 0 \\ \frac{S+1}{\sigma_s} & \text{if } S < 0 \end{cases}$	[14.11]
---------	--	---------

Z_S is compared to the $\alpha/2$ quantile of the normal distribution to obtain the two-sided p-value for the test of significance on τ_b .

Example 3, cont.

Is the value of $\tau_b = 0.36$ significantly different from zero? From equation 14.9 the approximate value of σ_S is

$$\begin{aligned} \sigma_S &\cong \sqrt{\frac{1}{9} \cdot (1 - (0.47^3 + 0.29^3 + 0.24^3)) \cdot (1 - (0.32^3 + 0.35^3 + 0.33^3)) \cdot 79^3} \\ &\cong \sqrt{\frac{(0.86) \cdot (0.89) \cdot 79^3}{9}} = \sqrt{42329} = 205.74 \\ Z_S &\cong \frac{724 - 1}{205.74} = 3.51 \end{aligned}$$

and from a table of the normal distribution the one-sided p-value is $p = 0.0002$. Therefore $H_0: \tau_b = 0$ is rejected, which means that pesticide concentrations increase (the distribution shifts to a greater proportion of higher classes) as soil drainage increases.

14.5 Other Methods for Analysis of Categorical Data

One other method is prominent in the statistical literature for analysis of all three situations discussed in this chapter -- loglinear models (Agresti, 1984). Loglinear models transform the expected cell probabilities $e_{ij} = a_i \cdot c_j$ by taking logarithms to produce a linear equation $\ln(e_{ij}) = \mu + \ln(a_i) + \ln(c_j)$, where μ is a constant. Models may be formulated for the completely nominal case, as well as for one or more ordinal variables. Detailed contrasts of the probability of being in column 2 versus column 1, column 3 versus 2, etc. are possible using the loglinear model. Tests for higher dimensioned matrices (such as a 3-variable 3x2x4 matrix) can be formulated. Interactions between the variables may be formulated and tested analogous to an analysis of variance on continuous variables. Though the computation of such models is not discussed here, Agresti (1984) provides ample detail.

Exercises

14.1 Samples of water quality collected at USGS National Stream Quality Accounting Network (NASQAN) stations from 1974 to 1981 show more frequent increases in chloride ion than decreases. 245 stations were classified by Smith et al. (1987) by their trend analysis results at $\alpha = 0.1$. One reasonable cause for observed trends is road salt applications. Estimates of tons of road salt applied to the 245 basins in 1975 and 1980 are used to place the stations into into 3 groups: decreases (1980 is more than 20% less than 1975), increases (1980 is more than 20% greater than 1975), and little or no change. The two variables are then summarized by this 3x3 table:

Trend in Cl⁻ (1974-81, $\alpha=0.1$)

<u>Δ road salt appl.</u>	Down	No trend	Up	Totals
Down	5	32	9	46
No change	14	44	25	83
Up	10	51	55	116
Totals	29	127	89	245

Test H_0 : a basin's trend in chloride ion is independent of its change in road salt application, versus the alternative that they are related.

- a) using a contingency table. Interpret the test result.
- b) using Kendall's tau. Interpret the test result.
- c) which test is more appropriate, and why?

14.2 Fusillo et al. (1985) sampled 294 wells in New Jersey for volatile organic compounds. The wells were classified by whether they were located in an outcrop location near the surface, or whether they were further downdip and somewhat more protected from direct contamination. Determine whether the probability of finding detectable levels of volatile compounds differs based on the location of the well.

<u>Location</u>	Non-detects	Detect VOC	Totals
Downdip	106	9	115
Outcrop	129	50	179
Totals	235	59	294

- 14.3 Regulation of organo-tin antifouling paints for boats was announced in 1988 in Switzerland. Concentrations of tributyltin (TBT, in ng/L) in unfiltered water samples from Swiss marinas were measured in 1988 to 1990 (Fent and Hunn, 1991). Is there evidence of a decrease in TBT concentrations in marina waters as these paints were being taken off the market?

<u>Year</u>	<u>Number of samples</u>		Totals
	TBT ≤ 200	TBT > 200	
1988	2	7	
1989	9	13	
1990	10	10	
Totals			51