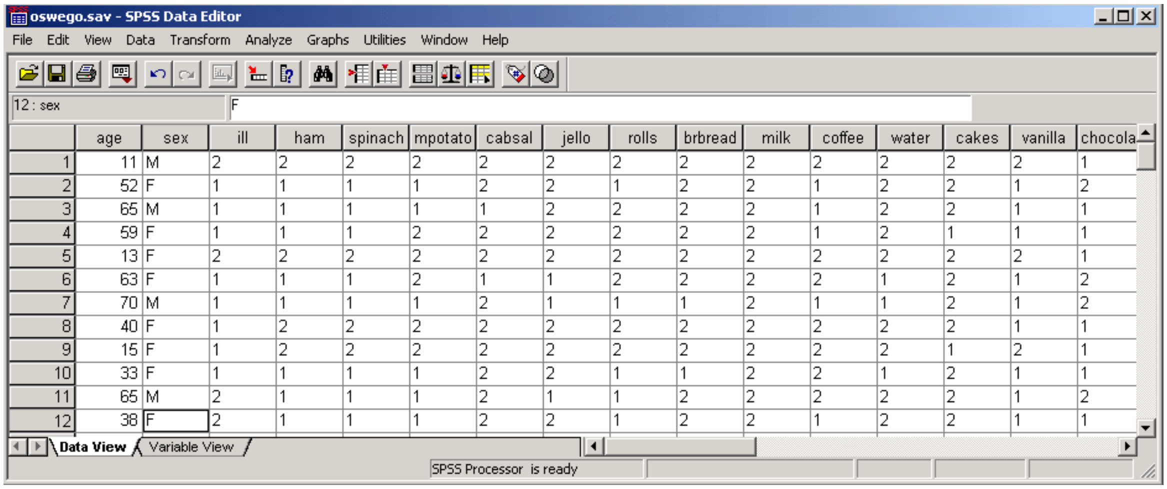
10: Cross-Tabulated Counts and Independent Proportions

Introduction

Data

Consider the analysis of a categorical (nominal) outcome with only two possible values: whether a person became ill or remain well for instance. The outcome variable in such instances is “binary” (“dichotomous,” having only two categories). We start by considering the analysis of a binary outcome in two independent groups.

Illustrative example (`oswego.sav; vanilla*ill`). Seventy-five (75) people attended a church picnic in upstate New York. Forty-size (45) cases of food poisoning occurred following the picnic. Data on many variables were collected as shown in the table in the screen-shot below. We are interested in comparing the incidence of illness (`ill:` 1 = yes and 2 = no) in people who ate and did not eat the vanilla ice cream (`vanilla`: 1 = yes and 2 = no).



12 : sex F

	age	sex	ill	ham	spinach	mpotato	cabsal	jello	rolls	brbread	milk	coffee	water	cakes	vanilla	chocola
1	11	M	2	2	2	2	2	2	2	2	2	2	2	2	2	1
2	52	F	1	1	1	1	2	2	1	2	2	1	2	2	1	2
3	65	M	1	1	1	1	1	2	2	2	2	1	2	2	1	1
4	59	F	1	1	1	2	2	2	2	2	2	1	2	1	1	1
5	13	F	2	2	2	2	2	2	2	2	2	2	2	2	2	1
6	63	F	1	1	1	2	1	1	2	2	2	2	1	2	1	2
7	70	M	1	1	1	1	2	1	1	1	2	1	1	2	1	2
8	40	F	1	2	2	2	2	2	2	2	2	2	2	2	1	1
9	15	F	1	2	2	2	2	2	2	2	2	2	2	1	2	1
10	33	F	1	1	1	1	2	2	1	1	2	2	1	2	1	1
11	65	M	2	1	1	1	2	1	1	2	2	2	2	2	1	2
12	38	F	2	1	1	1	2	2	1	2	2	1	2	2	1	1

The essence of the relation between `ill` and `vanilla` can be distilled by cross-tabulating the data. The following notation is adopted:

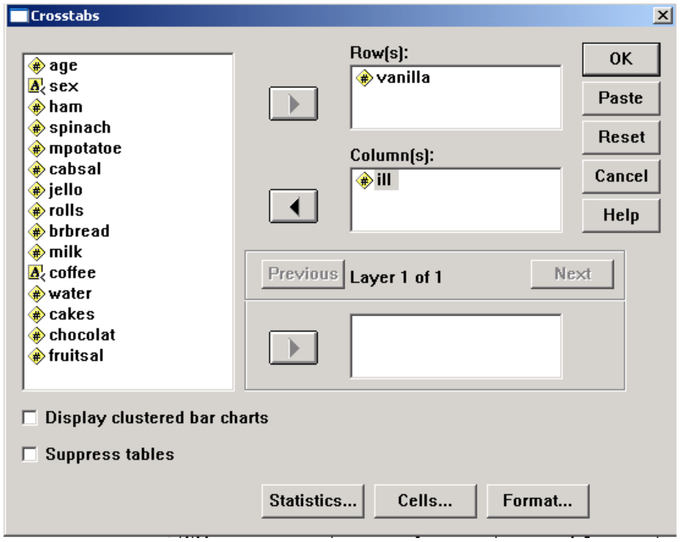
	Disease +	**Disease -**	
Exposure +	A_1	B_1	N_1
Exposure −	A_0	B_0	N_0
	M_1	M_0	N

In this notation, *A* indicates “case” and *B* indicates “noncase.” Subscript $_1$ denotes “exposed” and subscript $_0$ denotes “nonexposed.” For example, A_1 indicates the number of exposed cases, A_0 indicates the number of nonexposed cases, and so on. There are N_1 exposed observations and N_0 nonexposed observations and *N* observations total.

> *Comment:* Cross-tabulated data can be set up with group status across rows or columns. This would *not* materially affect conclusions but *would* require us to change notation and the way proportions are calculated. In this chapter, for the sake of uniformity, we put exposure groups in rows.

Data could be cross-tabulated manually by tossing each observation into its appropriate category and tallying counts. But let's face it, nobody (well almost nobody) does this type of tallying by hand anymore. With a computer, the cross-tabulation can be done with a simple command.

SPSS: Data are cross-tabulated with `Analyze > Descriptive Statistics> Crosstabs.` After making these choices you will be presented with the CrossTabs dialogue box. Select the group (exposure) variable in the Row field and the outcome (disease) variable in the Column field:



Row(s): vanilla

Column(s): ill

□ Display clustered bar charts

□ Suppress tables

After clicking "OK" cross-tabulation reveals:

	ill		
vanilla	Yes	No	Total
Yes	43	11	54
No	3	18	21
Total	46	29	75

Thus, there were 54 exposed observations and 21 non-exposed observations. There were 43 exposed cases and 3 non-exposed cases.

Descriptive Statistics

The next step is to convert counts to relevant proportions. The **sample proportion in group 1** is:

$$\hat{p}_1 = \frac{A_1}{N_1} \qquad \textbf{(10.1)}$$

The **sample proportion in group 0** is

$$\hat{p}_0 = \frac{A_0}{N_0} \qquad \textbf{(10.2)}$$

Illustrative example (`oswego.sav, vanilla*ill`). The (incidence) proportion of illness in the people eating vanilla ice cream is $\hat{p}_1$ = 43 / 54 = .796. The incidence proportion in those not eating vanilla ice cream is $\hat{p}_0$ = 3 / 21 = .143. There was a much higher incidence of food poisoning in the vanilla ice cream eaters.

> Comment: The proportions in the illustrative data represent incidences. "Incidence proportion" is synonymous with "average risk."

Estimation of the Risk Difference

Sample proportions $\hat{p}_1$ and $\hat{p}_0$ are statistical estimates of p_1 and p_0, respectively. Presumed effects of the exposure can be summarized in the form of a ratio ("risk ratio") or difference ("risk difference"). Let us consider the risk difference. (The risk ratio is considered in HS267.)

The risk difference (RD) parameter ($p_1 - p_0$) is an absolute measure of the effect of being in group 1 (the exposed group). This **risk difference estimator** is:

$$\hat{p}_1 - \hat{p}_0 \qquad \textbf{(10.3)}$$

Illustrative example (`oswego.sav; vanilla*ill`). The risk difference in the illustrative data = .796 – .143 = .653. Therefore, eating the vanilla ice cream increased risk by .653 (about 65%).

OPTIONAL: A 95% confidence interval for the risk difference is given by

$$\hat{p}_1 - \hat{p}_0 \pm (1.96)(se_{\hat{p}_1 - \hat{p}_0}) \qquad \textbf{(10.4)}$$

where $se_{p1\text{-}p0}$ represents the standard error of the proportion difference:

$$se_{\hat{p}_1 - \hat{p}_0} = \sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_0\hat{q}_0}{n_0}} \qquad \textbf{(10.5)}$$

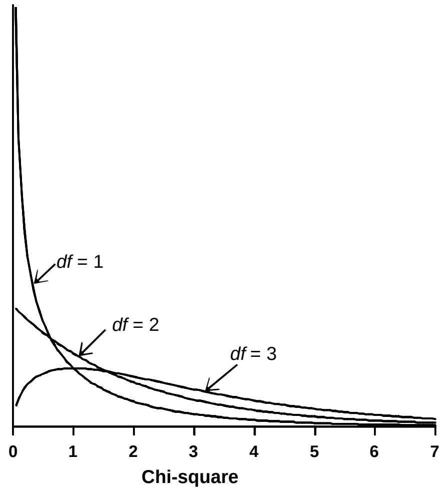
Illustrative example (`oswego.sav`). The standard error of the proportion difference, $se_{p1\text{-}p0}$ = $\sqrt{\frac{(.796)(1-.796)}{54} + \frac{(.143)(1-.143)}{21}}$ = .094. The 95% confidence interval for $p_1 - p_0$ = (.796 –.143) ± (1.96)(.094) = .653 ± .184 = (.469, .837). As with all confidence intervals, this gives us a much better idea of the long-run result, or "what might be."

Hypothesis Test

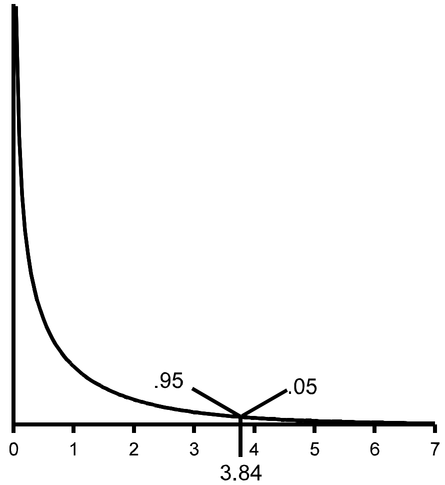
Chi-square distribution

There are several ways to test data for statistical significance. Because of its utility in a variety of situations, a chi-square test is used in this chapter.

Chi-square tests are based on chi-square probability distributions. Chi-square probability are are asymmetrical with have long right-tails. They have degrees of freedom, much like *t* distributions. Chi-square distributions with 1, 2, and 3 degrees of freedom look like this:



Let $\chi^2_{df,p}$ represent the p^{th} percentile on a chi-square distribution with *df* degrees of freedom. Such percentiles are looked-up in chi-square tables such as the one in the back of this book. As an example, the 95th percentile on a chi-square distribution with 1 degree of freedom is 3.84. Graphically, this looks like:



Chi-Square Test

(A) Hypotheses. We imagine a big population. This big population is sampled to uncover sample proportions $\hat{p}_1$ and $\hat{p}_0$. Based on data in a given sample, we want to know whether the population proportions differ. Under the null hypothesis there is no difference in population proportions and under the alternative there is. Thus, we test:

$$H_0\text{: } p_1 = p_0 \text{ versus } H_1\text{:“}H_0 \text{ is false”}$$

(B) Alpha threshold. Alpha levels are needed for fixed-level testing and are optional for if conducting flexible significance testing.

(C) Test statistic. The chi-square test statistic is

$$c^{2}_{\text{stat}} = \sum \frac{(O_i - E_i)^2}{E_i} \tag{10.6}$$

where O_i represents the observed frequency in table cell i and
E_i represent the expected frequency calculated as :

$$E_i = \frac{\textbf{row\_ total} \times \textbf{column\_ total}}{\textbf{total}} \tag{10.7}$$

The test statistic has df = $(R-1)(C-1)$ where R represents the number of rows in the cross-table and C represents the number of columns. In testing a 2-by-2 cross-tabulation, df = $(2-1)(2-1) = 1$.

(D) Conclusion. An approximate p value is determined by placing the chi-square statistic on its proper chi-square distribution and determining the area under the curve to the right of the χ^2_{stat}. We use percentile landmarks from the χ^2 table for this purpose .

A more precise p value can be found with computer programs. We can also derive a precise p values for a χ^2_{stat} with 1 degree of freedom by taking the square root of the χ^2_{stat} (the “χ_{stat}”) and using a χ table to look up the p value. There is a χ table in the back of this book.

The p value is compared to the alpha level (fixed-level testing) or used in a flexible way to test the significance of results.

Illustrative example (`oswego.sav; vanilla*ill`). (A) The population is hypothetical in this instance. We imagine a “super-population” in which an infinite number of people are exposed to the vanilla ice cream in question. We then imagine what would have happened in this population if it were non-exposed. Under the null hypothesis the incidence risks in the two population would not differ. Thus, H_0: $p_1 = p_0$ versus H_1: *“the null hypothesis is wrong.”*

(B) Let us conduct a flexible significance test.

(C) We calculate expected frequencies:

	`ill`		
`vanilla`	1	2	Total
1	(54)(46)/(75) = 33.12	(54)(29)/(75) = 20.88	54
2	(21)(46)/(75) = 12.88	(21)(29)/(75) = 8.12	21
	46	29	75

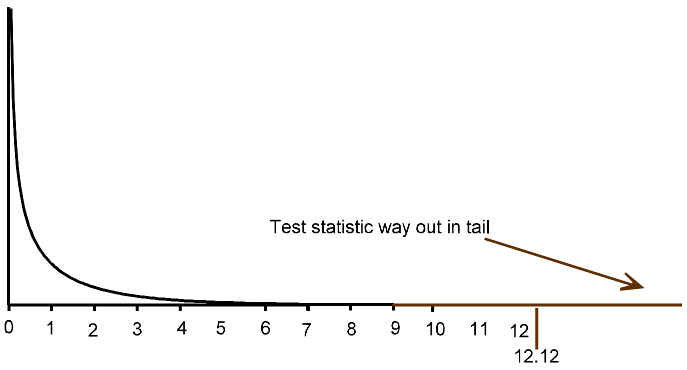
The chi-square test statistic is:

$$c^2_{stat} = \frac{(43-33.12)^2}{33.12} + \frac{(11-20.88)^2}{20.88} + \frac{(3-12.88)^2}{12.88} + \frac{(18-8.12)^2}{8.12}$$

$$= 2.95 + 4.68 + 7.58 + 12.02$$
$$= 27.23$$

df = (2-1)(2-1) = 1

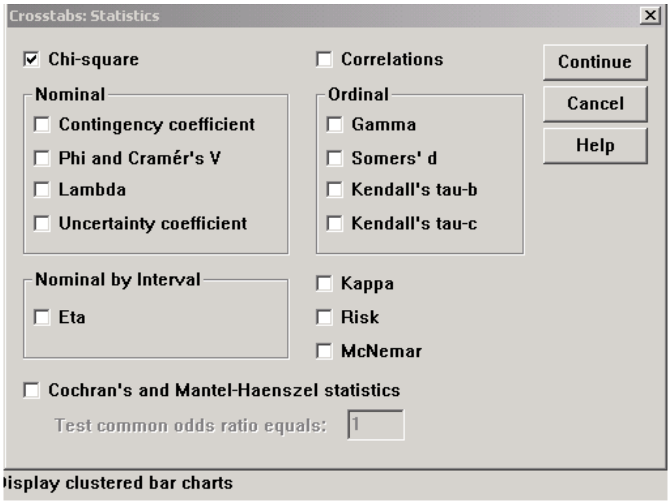
(D) The χ^2_{stat} is placed on the chi-square distribution. We go to the chi-square table and find the largest χ^2 percentile (landmark) is 12.12 with a right-tail region of .0005. The χ^2_{stat} is beyond this landmark indicating that $p < .0005$. (In general, the bigger the χ^2_{stat} the smaller the p value.) The observed difference is *not* likely to be random; the difference is “significant.”



In addition, $\chi_{stat} = \sqrt{(27.23)} = 5.21$. The largest χ value in our table is 3.99, corresponding with p = .00007. Since the current χ_{stat} is greater than 3.99, we know $p < .00007$.

SPSS

To get a chi-square statistic in SPSS, click `Analyze > Descriptive Statistics> Crosstabs`. Then select variables for the row and column of the table. Click the Statistics button and check the Chi-square box:



Crosstabs: Statistics

☑ Chi-square

☐ Correlations

Nominal
- ☐ Contingency coefficient
- ☐ Phi and Cramér's V
- ☐ Lambda
- ☐ Uncertainty coefficient

Ordinal
- ☐ Gamma
- ☐ Somers' d
- ☐ Kendall's tau-b
- ☐ Kendall's tau-c

Nominal by Interval
- ☐ Eta

☐ Kappa

☐ Risk

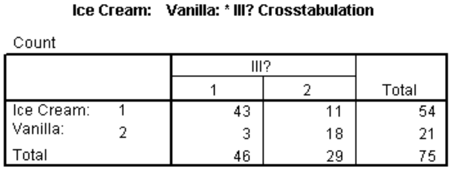
☐ McNemar

☐ Cochran's and Mantel-Haenszel statistics

Test common odds ratio equals: 1

)isplay clustered bar charts

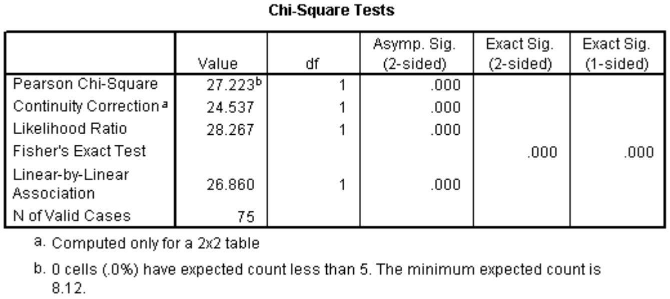
SPSS output looks like this:



Ice Cream: Vanilla: * Ill? Crosstabulation

Count

		Ill?		
		1	2	Total
Ice Cream: Vanilla:	1	43	11	54
	2	3	18	21
Total		46	29	75



Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	27.223[b]	1	.000		
Continuity Correction[a]	24.537	1	.000		
Likelihood Ratio	28.267	1	.000		
Fisher's Exact Test				.000	.000
Linear-by-Linear Association	26.860	1	.000		
N of Valid Cases	75				

a. Computed only for a 2x2 table

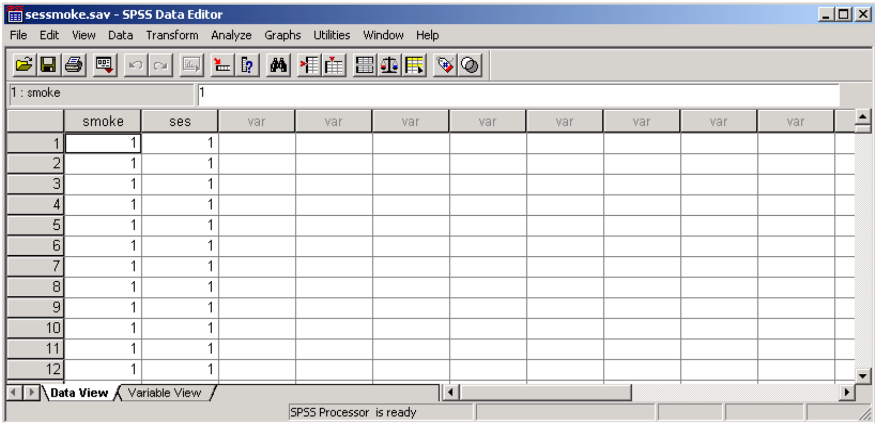
b. 0 cells (.0%) have expected count less than 5. The minimum expected count is 8.12.

Only the cross-tabulation and Pearson's chi-square statistic have been covered thus far. The continuity correction chi-square and Fisher's exact test is covered in this course.

R-by-C Table

The chi-square test can be used to test cross-tabulated counts from any size frequency table. Let *R* denote the number of rows in the table and *C* denote the number of columns. We speak of *R*-by-*C* tables.

Illustrative example (`sessmoke.sav`). A cross-sectional survey is conducted to explore the relation between socioeconomic class (`SES`) and smoking (`SMOKE`). `SES` data are coded into 5 ordinal categories, with 1 indicating low SES and 5 indicating high SES. Smoking is coded 1 = current smoker, 2 = not a current smoker. A screen shot from the SPSS data set looks like this:


sessmoke.sav - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
Data View
Variable View
SPSS Processor is ready

1 : smoke 1

	smoke	ses	var	var	var	var	var	var	var	var
1	1	1								
2	1	1								
3	1	1								
4	1	1								
5	1	1								
6	1	1								
7	1	1								
8	1	1								
9	1	1								
10	1	1								
11	1	1								
12	1	1								

Cross-tabulation reveals:

	SMOKE		
SES	1	2	Total
1	17	40	57
2	76	195	271
3	34	88	122
4	32	53	85
5	20	30	50
Total	179	406	585

When data are set up with the explanatory variable along the rows and outcome variable along the columns (at they are above), we calculate relevant proportions as $\hat{p}_i = \dfrac{\text{no. positive\_for\_attribute}}{\text{row\_total}}$.

For the illustrative data, $\hat{p}_1$ = 17 / 57 = .298, $\hat{p}_2$ = 76 / 271 = .280, $\hat{p}_3$ = 34 / 122 = .279, $\hat{p}_4$ = 32 / 85, = .377, $\hat{p}_5$ = 20 / 50 = .400, indicating slightly higher smoking proportions with high SES.

Hypothesis Test for R-by-C Data

(A) Null and Alternative Hypotheses. The null and alternative hypotheses may be stated:

H_0: no association between the row and column variable
H_1: "association"

(B) Alpha levels. Alpha levels are needed for fixed-level testing but not for flexible significance testing.

(C) Test statistic. Formulas 10.6 and 10.7 apply. The chi-square statistic for *R*-by-*C* data have $(R-1)(C-1)$ degrees of freedom.

(D) Conclusion. Computer programs will calculate precise *p* values for chi-square tests. When a computer is unavailable, you must look up the *p* value using a χ^2 table. The *p* value corresponds to the area under the curve in the tail of the proper distribution. You will *not* be able to find the precise *p* value in the table. However, you will be able to find the approximate value of the *p* value as follows:

(1) Draw the χ^2 curve. (Recall that the χ^2 distribution will be asymmetrical with a long right tail.
(2) Place the χ^2_{stat} on the curve in its approximate location.
(3) Shade the tail to the right of the χ^2_{stat} . This represents the *p* value for the problem.
(4) Use the proper df row, find the χ^2 landmark that is just to the left of the χ^2_{stat}.
(5) Report the *p* value as an inequality.

Illustrative example (`sessmoke.sav`). We want to test whether there is an association between the row variable (SES) and column variable (SMOKE).

(A) Hypotheses . The null and alternative are H_0: no association between the SES and smoking in the population versus H_1: association between SES and smoking in the population.

(B) Alpha level. Let us conduct a fixed-level test at $\alpha = .05$.

(C) Test statistic. Expected frequencies are calculated according to formula 10.7:

	SMOKE		
SES	1	2	Total
1	17.4[†]	39.6	57
2	82.9	188.1	271
3	37.3	84.7	122
4	26.0	59.0	85
5	15.3	34.7	50
Total	179	406	585

[†] Example: Expected value in this cell= (57)(179)/585 = 17.4.

The chi-square test statistic is calculated according to formula 10.6:

$$
\begin{array}{lllll}
\chi^2_{\text{stat}} & = & (17 - 17.4)^2 / (17.4) & + & (40 - 39.6)^2 / (39.6) & + \\
 & & (76 - 82.9)^2 / (82.9) & + & (195 - 188.1)^2 / (188.1) & + \\
 & & (34 - 37.3)^2 / (37.3) & + & (88 - 84.7)^2 / (84.7) & + \\
 & & (32 - 26.0)^2 / (26.0) & + & (53 - 59.0)^2 / (59.0) & + \\
 & & (20 - 15.3)^2 / (15.3) & + & (30 - 34.7)^2 / (34.7) & +
\end{array}
$$

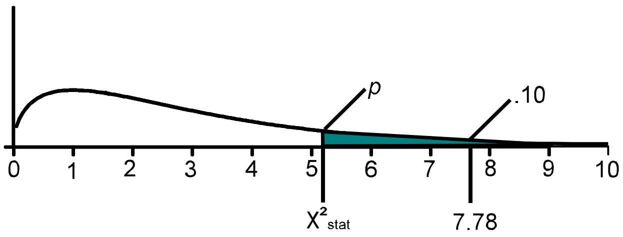
$$
\begin{array}{llllll}
= & 0.01 & + & 0.00 & + \\
 & 0.57 & + & 0.25 & + \\
 & 0.29 & + & 0.13 & + \\
 & 1.38 & + & 0.61 & + \\
 & 1.44 & + & 0.64 &
\end{array}
$$

$$= 5.32$$

These data have $df = (R-1)(C-1) = (5-1)(2-1) = 4$.

(D) Conclusion. The p value is the area under the curve beyond the test statistic under in a χ^2 with 4 degrees of freedom. The precise p value (determined by computer) is .25. Therefore, H_0 is retained; the association is not significant.

Had a computer been unavailable, we would have look up the p value using a χ^2 table. We note that the 90th percentile on a χ^2 distribution with 4 degrees of freedom is 7.78:



Therefore, $p > .10$.

Assumptions necessary for chi-square tests.

1. *Expectations exceed 5.* Chi-square tests can be used only when expected frequencies are 5 or greater. When expected frequencies are less than 5, exact calculation methods are necessary (see next chapter for details).

2. *Sampling independence.* The above chi-square test can be used only when observations in the sample are independent. A different method is necessary when data are paired or matched.

3. *Freedom from systematic error.* Data must be free from systematic and other non-sampling errors (such as confounding). Since this is at best an approximation of reality, p values may be viewed to represent a *minimal* level of uncertainty (Tukey, 1986, p. 75).