

Analysis of Contingency Tables

Contingency Tables

Contingency tables, sometimes called cross-classification or crosstab tables, involve two categorical variables. More generally, we will refer to the two variables as each having I or J levels. For simplicity, we will start by assuming two binary variables, forming a 2×2 table, in which $I = 2$ and $J = 2$. The four cells of the design can contain counts, n_{ij} , or proportions, p_{ij} . The first subscript is an index of which level of the first variable is referred to. In the 2×2 case, $i = 1$ for the first row or $i = 2$ for the second row, and $j = 1$ for the first column or $j = 2$ for the second column. For example, the count in the cell for the intersection of the first column and first row will be n_{11} . And the count in the cell in the intersection of the first row and second column is n_{12} and so on. *Marginal* values are referred to with a “+” symbol to designate that either i or j have been combined. So, for example, n_{1+} is used for the first row marginal total count, and n_{+1} is used for the first column marginal total count. Cell counts and proportions with the notation are presented below.

Cell counts/frequencies

n_{11}	n_{12}	n_{1+}
n_{21}	n_{22}	n_{2+}
n_{+1}	n_{+2}	n_{++}

Proportions

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

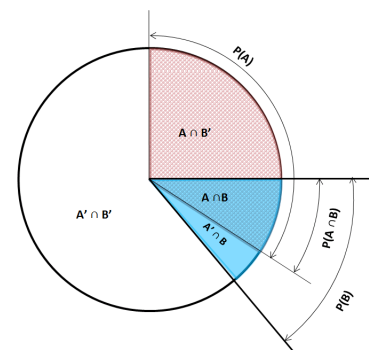
The proportions outside of the cells of each table (e.g., p_{+1} , p_{+2} , p_{1+} , p_{2+}) are marginal proportions that relate to the marginal distribution for that variable. Each marginal proportion involves only the counts for one of the variables and the total sample size, $p_{i+} = n_{i+}/n_{++}$ and $p_{+j} = n_{+j}/n_{++}$. (Note that either n_{++} or n may be used). The proportions inside the cells of the table (p_{11} , p_{12} , p_{21} , p_{22}) are joint proportions that relate to the joint distribution of the two variables. Each joint proportion is the count for that cell divided by the total count, $p_{ij} = n_{ij}/n_{++}$. Another distinction is the *conditional* proportion, which relates to the conditional distribution. Conditional proportions represent estimates of the conditional probability that $P(Y_i = 1)$ given the value of Y_j , written as $P(Y_i = 1|Y_j)$.¹ The conditional probability that event A occurs given event B is the same as the joint probability of A and B occurring relative to the probability of B occurring (or the B sample space).

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Similarly, the sample conditional proportion estimates the conditional probability, such as

$$p_{i|j} = p_{ij} / p_{+j} = n_{ij} / n_{+j}$$

In contingency tables, this is a row proportion (percentage), because it is the proportion of all cases in one row (n_{i+}) that appear in one column, n_{ij} (e.g., proportion of males that say “no”).



Andrew Batishchev

Bayes' Theorem

It is well worth a brief digression to discuss the famous theorem called *Bayes' theorem*, proposed by the eighteenth-century mathematician/clergyman Thomas Bayes, because of its very widespread application to statistics. Bayes' theorem is a simple method of computing the conditional probability of one event given another event if the probability of both events and the other conditional probability is known.

¹ The subscripts for the Y variable are potentially confusing, as Y here generally is assumed to be an individual score, with Y_i representing the individual score for the row variable with a particular value (e.g., $i = 2$) and Y_j representing the individual score on the column variable with a particular value (e.g., $j = 1$).

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

In a contingency analysis, we might be interested in computing the probability that an individual who completed a treatment center program will have a drug relapse within six months. If the probability of a relapse in general is .60, the probability of completing a treatment program is .64, and the probability of having completed a program among those who have relapsed before is .3, then the probability that an individual will relapse if he/she has completed the program is approximately .281.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{(.30)(.60)}{.64} = .281$$

Bayes' theorem is the kernel of Bayesian statistics, based upon the notion that informative prior information can be used in this way to better estimate a probability (or other statistic) than just using general information about the distribution. As with the problem above, its utility depends importantly on the accuracy of the prior information.

Conceptualizing Contingency Table Analysis

Pearson's chi-squared test (Pearson, 1900) is easily generalizable to analysis of contingency tables. There are three different but equivalent ways of conceptualizing the test. One is as a test of homogeneity, which refers to whether two groups are different on a binary response (e.g., whether males and females differ in the choice between two candidates). This conceptualization assumes one variable is an explanatory variable and one is a response, but the chi-square does not require such a designation. The goodness-of-fit conceptualization concerns the degree to which the observed frequencies fit the frequencies that would be expected do to chance. The same equation we discussed in the single variable case is also used for contingency analysis, but uses a different set of expected values that take the marginal frequencies of each of the two variables into account.

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}, \quad \text{where each cell's expected values is } E_{ij} = \frac{n_{i+}n_{+j}}{n_{++}}$$

Notice that the expected frequency for each cell is a joint frequency that is expected from the marginal frequencies for each of its respective rows and columns. The larger the discrepancies between the observed and expected frequencies are the greater lack of fit from what is expected due to chance. The third conceptualization is that chi-squared is a test of independence of two binary variables, or, rather, the test of whether they are dependent or correlated. As with correlation, we do not need to designate one variable as an explanatory variable and one as a response under this conceptualization. Although the equation most clearly reflects the goodness-of-fit aspect of the test, it is just as valid to consider it a test of questions about homogeneity or independence.

So, these different interpretations or rationales of chi-square mean that the statistic is useful to test many different types of hypotheses, given a data situation that involves all categorical variables. These three interpretations of chi-square also highlight the fact that when we are testing to see if two groups are different, we are also testing the hypothesis about whether a grouping variable (i.e., the dichotomous independent variable) is correlated with the dependent variable.

Computation

Let's use the Quinnipiac University poll data to examine the extent to which independents (non-party affiliated voters) support Biden and Trump.² Here are the frequencies:

² These results are based on a national Quinnipiac University poll from Oct 4-7, 2019, <https://poll.qu.edu/national/release-detail?ReleaseID=3643>. Methodological details are here https://poll.qu.edu/images/polling/us/us10082019_demos_uljv62.pdf/.

	Trump	Biden	
Party affiliated	338	363	701
Independent	125	156	281
	463	519	982

To answer the question whether Biden or Trump have a higher proportion of independent voters, we are making a comparison of the proportion of Biden supporters who are independents, $p_{2|2} = n_{22}/n_{+2} = 156/519 = .30$, or 30.0%, to the proportion of Trump supporters who are independents, $p_{2|1} = n_{21}/n_{+1} = 125/463 = .27$, or 27.0%. So, the table appears to suggest that Biden's supporters are more likely to be independents than Trump's supporters. Notice that this is a comparison of the conditional proportions, which correspond to column percentages in cross-tabulation output.³

First, we need to compute the expected frequencies for each cell. R_1 is the frequency for row 1, C_1 is the frequency for row 2, and N is the total sample size. The first cell is:

$$E_{11} = \frac{R_1 C_1}{N} = \frac{(701)(463)}{982} = 330.51$$

Filling in the rest of the cells in the same way for each expected value, E_{ij} , using the same equation but by using frequencies from the corresponding row R_i and column C_j for each cell, I obtained the following expected values:

$E_{11} = \frac{(701)(463)}{982} = 330.51$	$E_{12} = \frac{(701)(519)}{982} = 370.49$
$E_{21} = \frac{(281)(463)}{982} = 132.49$	$E_{22} = \frac{(281)(519)}{982} = 148.51$

$$\begin{aligned} \chi^2 &= \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \frac{(338 - 330.51)^2}{330.51} + \frac{(363 - 370.49)^2}{370.49} + \frac{(125 - 132.49)^2}{132.49} + \frac{(156 - 148.51)^2}{148.51} \\ &= 0.170 + 0.151 + 0.423 + 0.378 \\ &= 1.12 \end{aligned}$$

The result of the chi-square is compared to the tabled critical value based on $df = (I - 1)(J - 1)$, where I and J represent the number of rows and the number of columns, respectively. So, with $df = (I - 1)(J - 1) = 1$, the critical value is 3.84, and the computed value is not significant.

Comparing Proportions

An equivalent test, though rarely used, is to compare two proportions.

$$z = \frac{p_1 - p_2}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Where p without the subscript is a weighted (or "pooled") average for the two proportions, p_1 and p_2 . This version of the standard error for the denominator is a score test and yields the same result as the

³ There are other questions we might ask, of course. Asking whether the proportion of independents who support Biden is higher than the proportion of non-independents (affiliated) who support Biden is an equivalent question to the one above (comparison of conditional row proportions rather than conditional column proportions). We also might ask whether independents are more likely to support Biden than Trump, which is a simple two-cell comparison among independents, which would be made by simply selecting out independent voters and using the z -proportions or chi-square test previously discussed.

Pearson chi-square (Agresti, 2003, section 3.2.4). There are a variety of other proposed approaches to the standard error/confidence intervals including a Wald and exact intervals (see Newcombe, 1998).⁴

Residuals

Residuals, representing the discrepancy between the observed and expected frequencies are sometimes discussed or used in computations of other statistics, and called “Pearson residuals.” The sum of the squared residuals is equal to the Pearson chi-square. Typically, they are standardized by dividing by the standard deviation of the expected proportions, as below.

$$r_{ij} = \frac{O_{ij} - E_{ij}}{\sqrt{E_{ij}(1 - p_{i+})(1 - p_{+j})}}$$

For the standardized Pearson residuals, each cell can be evaluated for difference from what is expected under the null hypothesis using the z distribution or the chi-squared distribution, because $r_{ij} = z_{ij}$ and $r_{ij}^2 = \chi_{ij}^2$.

Minimum Expected Frequencies and Fisher’s Exact Test

Fisher’s exact test, proposed by R.A. Fisher (Fisher, 1935) and sometimes called the “Fisher-Irwin” test, is often printed along with the Pearson χ^2 . It is not so much a modification of the chi-square test as an alternative approach to testing the association between two binary variables for significance. The test has been suggested for use with small samples in which the expected frequencies in some cells are low. The concept is to use the hypergeometric distribution to compute the exact probability of the particular configuration of obtained frequencies. If the marginal frequencies (n_{+1} and n_{1+}) that related to one of the cells, say n_{11} , remain the same, the Fisher’s exact test asks how likely the obtained frequency is compared to all of the other more or less extreme outcomes. Only one cell is needed because the other cell counts are determined given a certain count in n_{11} as long as the marginal frequencies are the same. The hypergeometric distribution function bears some resemblance to the binomial distribution function which was based on combinatorials for probability of all the possible outcomes. The probability of occurrence of any given obtained frequency for n_{11} given its marginal frequencies is $P(n_{11})$.

$$P(n_{11}) = \frac{\binom{n_{11} + n_{12}}{n_{11}} \binom{n_{21} + n_{22}}{n_{21}}}{\binom{n_{++}}{n_{11} + n_{21}}} = \frac{(n_{11} + n_{12})!(n_{21} + n_{22})!(n_{11} + n_{21})!(n_{21} + n_{22})!}{n_{11}!n_{12}!n_{21}!n_{22}!n_{++}!}$$

The estimated probability is then compared to probabilities for other configurations of possible cell frequencies given the same marginal frequencies, but with more extreme results. If n_{+1} and n_{1+} are both 12, the possible values between the null and most extreme alternative hypothesis range from 6 to 12. Fisher used the approach in the famous test of his colleague Dr. Muriel Bristol’s assertion that she could taste the difference between cups of tea in which the milk was added first compared with cups in which the milk was added last.⁵ The test was originally a one-tailed test and Irwin (1935) proposed a two-tailed approach. The problem with Fisher’s exact test is that it can be overly conservative and its use is often recommended when not necessary. Some software packages print a warning when 20% of the cells have an expected frequency below 5 (known as Cochran’s rule). First thing to notice, however, is that it is the expected frequency that is of concern and not the observed frequency. Secondly, simulation studies (e.g., Camilli & Hopkins, 1978) suggest that Pearson’s χ^2 has nominal alpha values with

⁴ The Wald SE is $\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$. See Caffo (2007) <http://ocw.jhsph.edu/courses/MethodsInBiostatisticsII/PDFs/lecture18.pdf>.

⁵ Incidentally, if you are a complete stat geek like me, you will enjoy David Salsburg’s book *The Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century*.

expected values as low as 1 as long as the total sample size is 20 or larger. So, the upshot is that Fisher's exact test is not needed in very many circumstances.

Yates' Continuity Correction

Yates suggested a correction to the Pearson's χ^2 based on the notion that a test of discrete variables which should follow a discrete distribution are tested using a normal approximation, the chi-squared distribution. The Yates' correction for continuity is a simple modification of the chi-squared test formula by subtracting $\frac{1}{2}$ or .5 from the frequency difference.

$$\chi^2 = \sum \frac{(|O_i - E_{ij}| - .5)^2}{E_{ij}}$$

There is good evidence and fairly wide consensus that the results with the Yates correction are too conservative (e.g., Grizzle, 1967; Camilli & Hopkins, 1978).

Partitioning

The chi-squared values for the set of all possible *orthogonal* chi-squares add up to the chi-square for the whole design (or the omnibus test). The likelihood ratio test, G^2 , however, cannot be partitioned in the same way. Planned follow-up analyses to an omnibus Pearson χ^2 in complex contingency tables are simply chi-square analyses based on chi-squared tests for two-cell comparisons or smaller contingency tables (e.g., a 2×2 from a 5×3 design). Such tests may involve marginal proportions or cell proportions as well. With many post hoc tests, the researcher may wish to use a familywise error adjustment. The straight Bonferroni, although the most commonly applied, is overly conservative. I recommend other alternatives instead (e.g., Hochberg's sequential or Sidák-Bonferroni). General procedures for any set of significance tests in SAS (PROC MULTEST) and R (p.adjust) can provide some of these tests.

Comparing Chi-squared Values

It is also possible to compare two independent (between subjects not repeated measures) chi-squared values. The formula below (D'Agostino & Rosman, 1971) assumes the same dimensional tables and, therefore, the same *df*.

$$z = \frac{\sqrt{\chi_1^2} - \sqrt{\chi_2^2}}{\sqrt{1 - \left(\frac{1}{4}\right)df}}$$

Three or More Dimensions

Although 2×2 contingency table looks like a 2×2 factorial table, they are not analogous. The homogeneity conceptualization of chi-squared tests involves a two-group comparison of a binary outcome, which is analogous to a *t*-test in the continuous case. Because one of the columns (or rows) is for the dependent variable, it is really the three-way table that is analogous to the factorial design in ANOVA, which requires an analysis of a three-way contingency table ($2 \times 2 \times 2$) in the binary outcome case. I will cover these tests in the next section.

Software Examples

Below I illustrate SPSS, R, and SAS procedures for the Pearson's χ^2 test to independents on their favorability of the two candidates using the Reuters polling data.

SPSS

The crosstabs table has row (% within ind), column (% within response), and total (% of total) percents, although I usually only request percent row (or column) in practice.

SPSS

```
crosstabs /tables=ind by response
/cells=count row column expected
/statistics=chisq phi.
```

ind ind vs affil * response intended vote Crosstabulation

		response intended vote			
		0 Trump	1 Biden	Total	
ind ind vs affil	.00 affiliate	Count	338	363	701
		Expected Count	330.5	370.5	701.0
		% within ind ind vs affil	48.2%	51.8%	100.0%
		% within response intended vote	73.0%	69.9%	71.4%
		% of Total	34.4%	37.0%	71.4%
	1.00 independent	Count	125	156	281
		Expected Count	132.5	148.5	281.0
		% within ind ind vs affil	44.5%	55.5%	100.0%
		% within response intended vote	27.0%	30.1%	28.6%
		% of Total	12.7%	15.9%	28.6%
Total	Count	463	519	982	
	Expected Count	463.0	519.0	982.0	
	% within ind ind vs affil	47.1%	52.9%	100.0%	
	% within response intended vote	100.0%	100.0%	100.0%	
	% of Total	47.1%	52.9%	100.0%	

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	1.122 ^a	1	.290		
Continuity Correction ^b	.977	1	.323		
Likelihood Ratio	1.123	1	.289		
Fisher's Exact Test				.322	.161
Linear-by-Linear Association	1.121	1	.290		
N of Valid Cases	982				

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 132.49.
b. Computed only for a 2x2 table

Symmetric Measures

		Value	Approximate Significance
Nominal by Nominal	Phi	.034	.290
	Cramer's V	.034	.290
N of Valid Cases		982	

R

```
> #this lessR BarChart function produces a chi-square test by default6
> BarChart (response, by=ind, horiz = FALSE, stat = "proportion", beside = TRUE)
```

Joint and Marginal Frequencies

```
response
ind    0    1 Sum
0    338 363 701
1    125 156 281
Sum  463 519 982
```

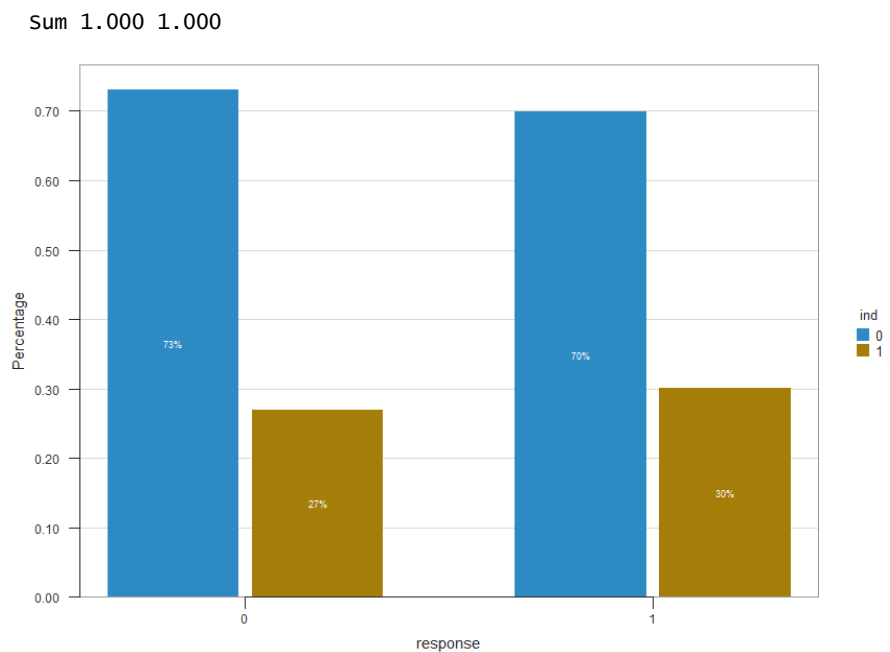
Cramer's v (phi): 0.034

Chi-square Test: Chisq = 1.122, df = 1, p-value = 0.290

Cell Proportions within Each Column

```
response
ind    0    1
0    0.730 0.699
1    0.270 0.301
```

⁶ This simpler syntax, BarChart (response, by=ind) provides the same chi-square test, but the extra statements I give above produces a better figure with the conditional proportions.



To get marginal frequencies and proportions in R

```
> BarChart(response, by=ind, horiz = FALSE, stat = "proportion", beside = TRUE)
```

Joint and Marginal Frequencies

```
response
ind      0      1 Sum
0      338  363 701
1      125  156 281
Sum    463  519 982
```

Cramer's v (phi): 0.034

Chi-square Test: Chisq = 1.122, df = 1, p-value = 0.290

Cell Proportions within Each Column

```
response
ind      0      1
0      0.730 0.699
1      0.270 0.301
Sum    1.000 1.000
```

```
> # the MASS package also can be used
```

```
> library(MASS)
```

```
> tbl = table(d$ind, d$response)
```

```
> tbl
```

```
      0      1
0 338 363
1 125 156
```

```
> margin.table(tbl, 2) # Frequencies summed over ind
```

```
      0      1
463 519
```

```
> chisq.test(tbl, correct = FALSE) #correct = FALSE turns off Yates continuity correction
```

Pearson's Chi-squared test

```
data:  tbl
```

```
X-squared = 1.1217, df = 1, p-value = 0.2896
```

SAS

```
*contingency chi-square;
proc freq data=one;
    tables ind*response /chisq;
run;
```

ind(ind vs affil)
 response(intended vote)

Frequency Percent Row Pct Col Pct	Trump	Biden	Total
affiliate	338 34.42 48.22 73.00	363 36.97 51.78 69.94	701 71.38
independent	125 12.73 44.48 27.00	156 15.89 55.52 30.06	281 28.62
Total	463 47.15	519 52.85	982 100.00

Frequency Missing = 6

Statistics for Table of ind by response

Statistic	DF	Value	Prob
Chi-Square	1	1.1217	0.2896
Likelihood Ratio Chi-Square	1	1.1235	0.2892
Continuity Adj. Chi-Square	1	0.9769	0.3230
Mantel-Haenszel Chi-Square	1	1.1205	0.2898
Phi Coefficient		0.0338	
Contingency Coefficient		0.0338	
Cramer's V		0.0338	

Fisher's Exact Test

Cell (1,1) Frequency (F)	338
Left-sided Pr <= F	0.8708
Right-sided Pr >= F	0.1615
Table Probability (P)	0.0323
Two-sided Pr <= P	0.3221

Sample Size = 982
 Frequency Missing = 6

Example write-up. A chi-square test was used to determine whether there was a significant difference between the proportion of Biden and Trump's supporters who are independent. Results indicated that 30.1% of Biden's supporters were independents, whereas 27.0% of Trumps supporters were independents. This difference was not significant, $\chi^2(1) = 1.12$, $p = .29$ The phi coefficient, $\phi = .03$, suggested a small effect.

Alternatively, and equivalently, the report could focus on the row proportions rather than the column proportions. In that case, the question would be stated in terms of the proportion of party affiliates vs. the proportion of independents who supported one candidate. So, the first and second sentence from above might read be stated in the following way (the chi-square is the same). A chi-square test was used to determine whether there was a significant difference between the proportion of party affiliates and independents who supported Biden. Results indicated that 51.8% of party affiliates supported Biden, whereas 55.5% of independents supported Biden.

References and Further Reading

- Camilli, G., & Hopkins, K. D. (1978). Applicability of chi-square to 2× 2 contingency tables with small expected cell frequencies. *Psychological Bulletin*, 85(1), 163-167.
- D'Agostino, R. B.: & Rosman, B. (1971) A normal approximation for testing the equality of two independent chi-square values. *Psychometrika*, 36, 251- 252.
- Fisher, R. A. (1935). *The design of experiments*. Edinburgh, UK: Oliver and Boyd.
- Grizzle, J. E. (1967). The Teacher's Corner: Continuity Correction in the χ^2 -test for 2× 2 Tables. *The American Statistician*, 21(4), 28-32.
- Irwin, J. O. (1935). Tests of significance for differences between percentages based on small numbers. *Metron*, 12(2), 84-94.
- Newcombe, R. G. (1998). Two-sided confidence intervals for the single proportion: comparison of seven methods. *Statistics in Medicine*, 17(8), 857-872.
- Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine Series 5*, 50 (302), 157–175.