Chapter 9
Analysis of Two-Way Tables

9.1 Inference for Two-Way Tables
9.2 Formulas and Models for Two-Way Tables
9.3 Goodness of Fit
9.1 Inference for Two-Way Tables

- Two-Way Tables
- Expected Cell Counts
- The Chi-Square Statistic
- The Chi-Square Distributions
- The Chi-Square Test

Two-Way Tables

The two-sample z procedures of Chapter 8 allow us to compare the proportions of successes in two populations or for two treatments. What if we want to compare more than two samples or groups? More generally, what if we want to compare the distributions of a single categorical variable across several populations or treatments? We need a new statistical test. The new test starts by presenting the data in a two-way table.

Two-way tables of counts have more general uses than comparing distributions of a single categorical variable. They can be used to describe relationships between any two categorical variables.
Expected Cell Counts

Two-way tables sort the data according to two categorical variables. We want to test the hypothesis that there is no relationship between these two categorical variables ($H_0$).

To test this hypothesis, we compare actual counts from the sample data with expected counts, given the null hypothesis of no relationship.

The expected count in any cell of a two-way table when $H_0$ is true is:

\[
\text{expected count} = \frac{\text{row total} \times \text{column total}}{n}
\]

The Problem of Multiple Comparisons

To perform a test of

- $H_0$: there is no difference in the distribution of a categorical variable for several populations or treatments
- $H_a$: there is a difference in the distribution of a categorical variable for several populations or treatments

we compare the observed counts in a two-way table with the counts we would expect if $H_0$ were true.

The problem of how to do many comparisons at once with an overall measure of confidence in all our conclusions is common in statistics. This is the problem of multiple comparisons. Statistical methods for dealing with multiple comparisons usually have two parts:

1. An overall test to see if there is good evidence of any differences among the parameters that we want to compare.
2. A detailed follow-up analysis to decide which of the parameters differ and to estimate how large the differences are.

The overall test uses the chi-square statistic and distributions.
The Chi-Square Statistic

To see if the data give convincing evidence against the null hypothesis, we compare the observed counts from our sample with the expected counts assuming $H_0$ is true.

The test statistic that makes the comparison is the \textbf{chi-square statistic}.

The \textbf{chi-square statistic} is a measure of how far the observed counts are from the expected counts. The formula for the statistic is:

$$
\chi^2 = \sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}
$$

where "observed" represents an observed cell count, "expected" represents the expected count for the same cell, and the sum is over all $r \times c$ cells in the table.

The Chi-Square Distributions

If the expected counts are large and the observed counts are very different, a large value of $\chi^2$ will result, providing evidence against the null hypothesis. The $P$-value for a $\chi^2$ test comes from comparing the value of the $\chi^2$ statistic with critical values for a \textbf{chi-square distribution}.

The \textbf{chi-square distributions} are a family of distributions that take only positive values and are skewed to the right. A particular $\chi^2$ distribution is specified by giving its \textbf{degrees of freedom}.

The $\chi^2$ test for a two-way table with $r$ rows and $c$ columns uses critical values from the $\chi^2$ distribution with $(r-1)(c-1)$ degrees of freedom. The $P$-value is the area under the density curve of this $\chi^2$ distribution to the right of the value of the test statistic.
Cell Counts Required for the Chi-Square Test

The chi-square test is an approximate method that becomes more accurate as the counts in the cells of the table get larger. We must therefore check that the counts are large enough to allow us to trust the $P$-value. Fortunately, the chi-square approximation is accurate for quite modest counts.

**Cell Counts Required for the Chi-Square Test**

You can safely use the chi-square test with critical values from the chi-square distribution when the average of the expected counts is 5 or more and all individual expected counts are 1 or greater. In particular, all four expected counts in a $2 \times 2$ table should be 5 or greater.

The Chi-Square Test

The chi-square test is an overall test for detecting relationships between two categorical variables. If the test is significant, it is important to look at the data to learn the nature of the relationship. We have three ways to look at the data:

1) **Compare selected percents**: which cells occur in quite different percents of all cells?

2) **Compare observed and expected cell counts**: which cells have more or less observations than we would expect if $H_0$ were true?

3) **Look at the terms of the chi-square statistic**: which cells contribute the most to the value of $\chi^2$?
The Chi-Square Test

One of the most useful properties of the chi-square test is that it tests the null hypothesis “the row and column variables are not related to each other” whenever this hypothesis makes sense for a two-way table.

**Uses of the Chi-Square Test**

Use the chi-square test to test the null hypothesis

\[ H_0: \text{there is no relationship between two categorical variables} \]

when you have a two-way table from one of these situations:

- Independent SRSs from two or more populations, with each individual classified according to one categorical variable
- A single SRS, with each individual classified according to both of two categorical variables