Chapter 11

Chi-square distributions A family of distributions that take only positive values and are skewed to the right. A particular chi-square distribution is specified by giving its degrees of freedom.

Chi-square goodness-of-fit test Suppose the Random, Large sample size, and Independent conditions are met. To determine if a categorical variable has a specified distribution, expressed as the proportion of individuals falling into each possible category, perform a test of

H_0: The specified distribution of the categorical variable is correct.

H_a: The specified distribution of the categorical variable is not correct.

We can also write these hypotheses symbolically using p_i to represent the proportion of individuals that fall in category *i*:

$H_0: p_1 = \_\_\_, p_2 = \_\_\_, ..., p_k = \_\_\_.$

H_a: At least one of the p_i's is incorrect.

Start by finding the expected count for each category assuming that H_0 is true. Then calculate the chi-square statistic

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

where the sum is over the *k* different categories. The *P*-value is the area to the right of χ^2 under the density curve of the chi-square distribution with $k - 1$ degrees of freedom.

Chi-square statistic A measure of how far the observed counts are from the expected counts. The formula is

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

where the sum is over all possible values of the categorical variable or all cells in the two-way table.

Chi-square test for association/independence Suppose the Random, Large sample size, and Independent conditions are met. You can use the chi-square test for association/independence to test

H_0: There is no association between two categorical variables in the population of interest.

H_a: There is an association between two categorical variables in the population of interest.

Or, alternatively

H_0: Two categorical variables are independent in the population of interest.

H_a: Two categorical variables are not independent in the population of interest.

Start by finding the expected counts. Then calculate the chi-square statistic

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

where the sum is over all cells in the two-way table. If H_0 is true, the χ^2 statistic has approximately a chi-square distribution with degrees of freedom = (number of rows – 1)(number of columns – 1). The *P*-value is the area to the right of χ^2 under the corresponding chi-square density curve.

Chi-square test for homogeneity Suppose the Random, Large sample size, and Independent conditions are met. You can use the chi-square test for homogeneity to test

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.

Start by finding the expected counts. Then calculate the chi-square statistic

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

where the sum is over all cells (not including totals) in the two-way table. If H_0 is true, the χ^2 statistic has approximately a chi-square distribution with degrees of freedom = (number of rows – 1)(number of columns – 1). The *P*-value is the area to the right of χ^2 under the corresponding chi-square density curve.

Components The individual terms $\frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$ that are added together to produce the test statistic χ^2.

Expected counts The expected numbers of individuals in the sample that would fall in each cell of the one-way or two-way table if H_0 were true.

Large sample size condition The rule of thumb that all expected counts must be at least 5.

Multiple comparisons The problem of how to do many comparisons at once with an overall measure of confidence in all our conclusions.

Observed counts The actual numbers of individuals in the sample that fall in each cell of the one-way or two-way table.

One-way table Used to display the distribution of a categorical variable for a sample of individuals.