1 Inference for Categorical Data: Chi-Square

1.1 Chi Square Test for Goodness of Fit

A goodness-of-fit test is used to test the hypothesis that an observed frequency distribution fits to some claimed distribution.

To measure the difference betwen the observed and expected counts, and to determine if the difference is significant, we will introduce a new test statistic, called the chi-square statistic:

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

Where O represents each observed count in the distribution and E represents each corresponding expected cout.

- The sampling distribution of the chi-square statistic is not a normal distribution.
- It is a right-skewed distribution that allows only for positive values because the statistic cannot be negative.

When all the counts are at least 5, the sampling distribution of the χ^2 statistic is close to a chi-square distribution with degrees of freedom (df) equal to the number of categories minus 1.

- The chi-square distributions are 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.
- The chi-square goodness-of-fit test uses the chi-square distribution with df = number of categories - 1

The null hypothesis in a chi-square goodness-of-fit test should stake a claim about the distribution of a single categorical variable in the population of interest.

The alternative hypothesis in a chi-square goodness-of-fit test is that the categorical variable does not have the specified distribution, and is easily given in words.

Conditions:

- Random - the data came from a well-designed random sample or randomized experiment
- Independent - When sampling without replacement, the 10% condition is met
- Large counts - All expected counts are at least 5 lets us say that the sampling distribution will follow a chi-squared distributions

When all conditions are met, the chi-squared goodness of fit test can be performed with the hypotheses:

- H_0: the claimed distribution is correct
- H_a: at least one proportion in the claimed distribution is incorrect

1.2 Chi Square Test for Homogeneity

The Chi Square Test for Homogeneity compares the distributions of one categorical variable across two or more populations to see if they are the same or different.

To construct a chi-square test for Homogeneity:

State the hypothesis:

- H_0: There is no difference in the distribution of the categorical variable among two or more groups
- H_A: There is a difference in the distribution of the categorical variable among two or more groups

Check the assumptions and conditions

- Randomness: The individuals whose counts are available for analysis should be a random sample of the population
- 10% condition: The sample size, n, must be no larger than 10% of the population
- Large Counts: We should expect to see at least 5 counts in each category of the categorical variable.

Name the inference method: Chi Square Test for Homogeneity

Calculate the Test Statistics:

- Observed Counts - Actual Frequencies of the variable from your sample
- Expected counts - projected frequencies of the variable if the null hypothesis is true. This is the row total times the column total divided by the table total.

Obtain the P-Value:

2nd-Vars-8:χ^2cdfwith lower being χ^2 and upper being ∞ and df is the number of rows - 1 times the number of columns - 1

In a calculator, enter the observed values into matrix A, and enter the expected values into matrix B, which will population when you run the χ^2 test. Then do stat-tests-c with observed in A and Expected in B.

Make a decision then write the conclusion in context.

1.3 Chi Square Test for Independence

The Chi-Squared Test for Independence compares the distribution of two categorical variables across one population to see if they are independent.

To construct this test:

State the Hypotheses:

- H_0 is that the first variable and the second variable are independent for the population
- H_A is that they are not independent

Check the Assumptions and Conditions (same as above)

Name the Inference Method

Calculate the Test Statistic (same as Homogeneity)

Everything else is the same as independence.

- A small p value is not proof of causation.
- The Chi Square Test for Independence treats the two variables symmetrically, we cannot differentiate the direction of any possible causation even if it existed.
- There is no way to eliminate the possibility that a lurking variable is responsible for the lack of independence.
- Don't say one variable "depends" on the other just because they are not independent.