Cheat Sheet: Chi-Square Statistics

Essential Concepts

- A chi-square (χ^2, pronounced “kai-square”) statistic is a test that measures how a model compares to actual observed data.
- The χ^2 test statistic measures the overall distance between observed and expected counts. The greater the chi-square test statistic, the further the observed counts are from what we expected.
- The formula for the chi-square test statistic:

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

- A goodness of fit hypothesis test determines whether or not the distribution of a categorical variable in a sample fits a claimed distribution in the population.
- Hypotheses
    - The null hypothesis states a specific distribution of proportions for each category of the variable in the population.
    - The alternative hypothesis says that the distribution is different from that stated in the null hypothesis.
- A chi-square model is a good fit for the distribution of the chi-square test statistic only if the following conditions are met:
    - Random: Observed counts must come from a random sample (to ensure our conclusions are free from sampling bias).
    - 10%: The sample size must be less than a tenth of the population size (to satisfy independence assumptions).
    - Large Sample: The sample size must be large enough such that the expected count in each cell is at least 5. (to ensure our sampling distribution resembles a chi-square distribution).

- A chi-square test of homogeneity determines if two or more populations (or subgroups of a population) have the same distribution of a single categorical variable. We use the test of homogeneity if the response variable has two or more categories and we wish to compare two or more populations (or subgroups.)

- Hypotheses

    - The null hypothesis is a statement of no difference or no change, so for a chi-square test of homogeneity, the null hypothesis is always that the distribution of the categorical variable is the same among all the populations.
    - The alternative hypothesis is that the distributions are not the same among all the populations, so this test looks for evidence that there are differences among the samples that are larger than those you would expect to see from just sampling variation if there really is no difference in the distributions for the populations.

- Conditions for a Chi-Squared Test of Homogeneity

    - Appropriate Data and Variables: This is not an official condition, but it is important to make sure that we are dealing with data that give the counts for each value of a categorical variable. That categorical variable should be measured for a sample from each population of interest.
    - Independence/Randomness Condition: The samples from our populations should be independent, random samples or independent samples that can be considered representative of the respective populations.
    - Large Sample Sizes Condition: The sample sizes need to be large enough so that the expected count in each cell is at least five.

- Residuals are calculated using the formula: Residual = Observed – Expected

- Standardized residuals are values that standardize the residuals so that if the null hypothesis is assumed to be true, they can be interpreted as normal z-scores.

- In the test of independence, we select individuals at random from a population and record data for two categorical variables. The null hypothesis says that the variables are independent.

- The null and alternative hypotheses for chi-square test of independence are the following:

    - H_0: The two variables of interest are independent.
    - H_A: The two variables of interest are not independent.

- Conditions for chi-square Test of Independence:

    - The data represent the counts for two categorical variables measured for individuals in one sample from one population.
    - Independence/Randomness Condition: The sample from our population should be an independent, random sample or independent sample that can be considered representative of the population.

    - Large Sample Size Condition: The sample size must be large enough so that the expected count in each cell is at least five.

- It could be that there is a third variable not included in our study that impacts the values of both of the variables we are considering. Such a variable is called a lurking variable.
- Fisher's Exact Test (also known as Fisher's Exact Test of Independence) is a statistical significance test used in the analysis of a 2×2 contingency table. It is used to determine whether or not there is a significant association between two categorical variables.

Key Equations

chi-square test statistic

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

degrees of freedom

$$df = (number\ of\ categories - 1)$$

degrees of freedom (homogeneity and independence tests)

$$df = (r - 1)(c - 1)$$

residual

$$\text{Residual} = \text{Observed} - \text{Expected}$$

standardized residual

$$\text{Standardized Residual} = \frac{\text{Observed} - \text{Expected}}{\sqrt{\text{Expected}}}$$

Glossary

lurking variable

a third variable not included in a study that impacts the values of both of the variables considered

marginal distribution

the distribution of one of the variables with no regard to the other variable whatsoever

standardized residuals

values that standardize the residuals so that if the null hypothesis is assumed to be true, they can be interpreted as normal z-scores