STAT 2120: Notes on Topic 10

Analysis of two-way tables:

- Some relevant summaries of categorical data in two categories:
  - Count or percent of the number of "successes" (as in two-sample setup of inference for proportions).
  - A two-way table, which easily generalizes to multiple categories (*i.e.*, more than just "successes" and "failures") and multiple samples (*i.e.*, more than just two).
- Recall the basic setup of two-way tables:
  - A two-way table involves a row variable and a column variable.
  - Notation uses r to denote the number of categories of the row variable and c to denote the number of categories of the row variable. The table is identified as an $r \times c$ table.
  - A cell is identified with each distinct combination of values among the variables.
  - The table may record the counts or percentages of data falling in each cell. The distribution of individual cells is called a joint distribution.
  - The distributions of the row and column variables appear in the margins of the table, and are called marginal distributions. Given as counts they are called row and column totals.
  - Notation uses n to denote the total number of data points that have been collected. (It is the sum of either a row or column total. In the two-sample setup it is $n = n_1 + n_2$.)
  - A conditional distribution is calculated from the counts of one variable limited to a given category of the other variable. These help to explore relationships between the variables.
- The sampling framework for two-way tables may arise in various ways:
  - Multiple, independent SRSs of categorical data from distinct populations, in which sample labels form one variable and the categorical data-values form the other.
  - A single SRS of paired categorical data, each point of which falls in a cell of a table.
- The question of interest, in "tabular thinking," is whether there is a relationship between the row and column variables.
  - In the two-sample setup, this is equivalent to contemplating $H_0: p_1 = p_2$ *versus* $H_a: p_1 \neq p_2$.
  - Another terminology asks whether there is an association between the variables. Here, the comparison is H_0: no association *versus* H_a: association. The direction of the association under H_a is unspecified.
  - The question may be asked more formally as whether the conditional distributions of one variable do not vary across the categories (given as the conditioning event) of the other variable. That is, one contemplates H_0: no variation among conditional distributions *versus* H_a: variation among conditional distributions.
- The basic approach to inference is to compare the observed cell counts to the expected cell counts, as they would be calculated under the null hypothesis of no association.
  - An expected cell count is calculated as the product of row and column totals corresponding to that combination of categories, divided by n. That is, exp. count = (row total) $\times$ (column total)$/n$.
- The chi-square statistic is a standardized metric that summarizes the distance between observed and expected cell counts, aggregated across all categories.
  - The formula for the chi-square statistic is $X^2 = \sum \frac{(\text{obs.count} - \text{exp.count})^2}{\text{exp.count}}$, where the sum is over all cells of the table.
  - A large value of X^2 provides evidence of a relationship between the variables. Thus, a test of H_0: no association *versus* H_a: association would reject H_0 if X^2 is large.
  - A p-value would be calculated from the sampling distribution of X^2 under H_0.
- The family of chi-square (χ^2) distributions describes the approximate sampling distribution of X^2 under H_0.
  - A specific chi-square distribution is denoted $\chi^2(k)$, or $\chi^2(\text{df})$, where k, or df, is a degree-of-freedom parameter that indexes the family.
  - Every chi-square distribution is right-skewed and takes only positive values.
  - Suppose the random variable V is $\chi^2(k)$, c is a positive number, and α is a number between 0 and 1. In Excel, **chidist**(c, k)$= P(V \geq c)$ and **chiinv**(α, k) is the c for which $P(V \geq c) = \alpha$.

- When H_0 is true (no association), the sampling distribution of the chi-square statistic is approximately $\chi^2(k)$ with $k = (r-1)(c-1)$. Larger (expected) cell counts improve the accuracy of this approximation, especially when for tables larger than 2×2.

- The chi-square test for two-way tables is as follows:
  - Assumptions: A valid sampling framework for two-way tables.
  - The comparison of hypotheses is H_0: no association *versus* H_a: association.
  - The standardized test statistic is $X^2 = \sum \frac{(\text{obs.count} - \text{exp.count})^2}{\text{exp.count}}$.
  - The P-value is calculated as: $P(V \geq X^2)$ for V having a $\chi^2(k)$ distribution with $k = (r-1)(c-1)$.
  - A rule of thumb is that the stated significance level for this test is accurate if: $\frac{n}{rc} = \frac{1}{rc}\sum \text{exp. count} \geq 5$ and every exp. count ≥ 1, when $r > 2$ or $c > 2$; every exp. count ≥ 5 when $r = 2$ or $c = 2$.
- Data in a two-way table may arise from an observational study or a designed experiment. The chi-square test would only establish the causation if the data came from a (designed) comparative, randomized experiment.
- In the 2×2 case, the p-value of the chi-square test is identical to that of the two-sample z test for proportions, in the formulation comparing $H_0: p_1 = p_2$ *versus* $H_a: p_1 \neq p_2$.
- Two-way tables may be useful are a tool of meta-analysis, in which the information of several studies are combined and analyzed together.

<u>Additional comments on the analysis of two-way tables:</u>

- Steps for a generic analysis of data in a two-way table:
  - Explore the data using relevant descriptive statistics such as histograms of the marginal and conditional distributions (usually after converting to relevant percentages).
  - Calculate expected cell counts and with these calculate the chi-square statistic.
  - Complete the test for an association by calculating a P-value (and checking the relevant rule of thumb for validity).
  - Draw conclusions about association based on the outcome of the test.
- Comments on exploratory analysis of conditional distributions:
  - As discussed before, examination of conditional distributions may help to explore relationships between categorical variables.
  - To explore a one-way relationship, one would typically examine the conditional distributions of the response variable across the categories of the explanatory variable.
  - Percentages may be used to check calculations, since row and column percentages must add to one hundred.
- Expected cell counts should add to the same row and column total as the observed cell counts.
- Comments about the chi-square statistic and the associated P-value:
  - Exploratory provides visualization if interesting, possibly complex patterns in the data. The chi-square test assesses the strength of evidence in those patterns of a relationship between the variables.
  - When the chi-square statistic is so large as to conclude a rejection of H_0, examination of individual terms of the chi-square statistic (those appearing mainly responsible for the statistic's large value) may indicate which observed cell counts are unusual (in the sense of being very different from the corresponding expected cell count).
- The sampling framework that leads to a two-way table reflects the design and purpose of a study.
  - The multi-sample framework matches the setup in which one variable is the response and the other is the exploratory variable, which records the labels of the populations. The objective is to explore the dependency of the response on the exploratory variable. (This is not really regression, though, since we're not fitting straight lines.)
  - The translation of "association" in the multi-sample setup is that the distribution of the response variable changes from population to population.
  - In the one-sample, paired-data framework, each individual in the sample yields two categorical variables. The translation of "association" is that the variables are not independent (as random variables).
- Distinctions between designs may be identified by asking whether the row and column totals are random variables subject to variability.
  - In the multi-sample case, the total counts of explanatory variable are fixed sample sizes (and those of the response are random).
  - In the one-sample, paired-data case, both row and column totals are random.
- Analysis of categorical data in a multi-way table is possible (using a straightforward generalization of the chi-square test), but not discussed here.