STAT1005 Week 10 Cheat Sheet

Matched Pair Design

- When you want to compare two populations, but we design an experiment where the pairs are matched
  - Eg. case-control clinical trials
- This allows us to reduce the two-sample t test to a 1 sample
- This results in a more powerful (significantly more sensitive) test
- ***Pearson Correlation Coefficient***
  - Is a covariance test for matched pair data
  - Measures the linear relationship
  - Is quite strongly effected by outliers
  - Pearson correlation coefficient denoted by r_{xy} is found by substituting the sample quantities into the theoretical definition of correlation:
    - $r_{xy} = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$
    - Where:
      - $\bar{x}$ is equivalent to μ_x,
        - $\bar{x} = n^{-1} \sum_{i=1}^{i=n} x_i$
      - $\bar{y}$ is equivalent to μ_y,
        - $\bar{y} = n^{-1} \sum_{i=1}^{i=n} y_i$
      - S_{xy} is equivalent to Cov(X, Y) = E[(X - μ_x)(Y - μ_y)]
        - $S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$
        - Which in practicality is:
          - $S_{xy} = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}$
      - S_{xx} is equivalent to σ_x^2,
        - $S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2$
        - Which in practicality is:
          - $S_{xx} = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$
      - S_{yy} is equivalent to σ_y^2,
        - $S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2$
        - Which in practicality is:
          - $S_{yy} = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2$
- ***Spearman's Correlation Coefficient***
  - This method is based purely on the ranks of the data
    - From smallest to largest
  - This is helpful for:
    - Ordinal data
      - Collected as numbers and arranged in order
    - Very good for if there are outliers as the result are not effected much by outliers
  - The Spearman's correlation coefficient is denoted by r_s and is calculated with:
    - $r_s = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2 - 1)}$
    - Where:
      - d_i^2 represents the squared difference between the ranks of each category per variable

Goodness-of-Fit Testing

- A goodness-of-fit test assesses whether a particular probability model is an appropriate measure for the data collected
  - Or assessing if the distribution of the data collected is consistent with the distribution of a population for the probability model
- This is done by assessing if the data collected is "close enough" to what we would expect to observe if the probability model is a good description
- ***Chi-Squared***
  - The chi-squared (χ^2) distribution is a special case of the gamma distribution
    - If the degrees of freedom is ν, the gamma parameters are:
      - $\alpha = \frac{\nu}{2}$
      - $\beta = 2$
  - Chi-squared is a goodness-of-fit model with an approximately chi-squared distribution given by:
    - $\chi^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i} = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}$
      - Where:
        - o_i is the observed frequency based on the quantity
        - e_i is the expected frequency based on the quantity
        - $n = \sum_{i=1}^{k} n_i$
      - It has k - 1 degrees of freedom
        - Provided that $np_i > 5$ in each of the i = 1, 2, 3, ..., k cells
        - This is because degrees of freedom represent how many variables are independent, and as there is a set number of variables for this test, once you know how many variables fit into each category, you know how many will be in the final category
    - The size of the discrepancy indicates how good a fit the model is
      - A small discrepancy indicates it might be a good fit
      - A large discrepancy indicates that the model is inadequate
      - *To assess whether a value is "large", we can*:
        - 1. Calculate a p-value
          - Which is the probability, that if the data does come from the distribution p(x), of observing a χ_i^2 value as large or larger than we observed
        - 2. Compare the calculated value of χ^2_{k-1} with a critical value of χ^2_{α} which is determined by the significance level α, (significance level is the same type of significance level we look at for usual p-values).
          - Calculating p- and critical values
            - Using R
              - pchisq(χ^2_{k-1}, k - 1)
              - qchisq(α, k - 1)
    - Additional considerations:
      - Previously for p_i, i = 1, 2, 3, ..., k, however it can be made up of by smaller number of parameters θ_1, ...,θ_m, (m < k). So a specific hypothesis involving the θ_i will give you the values of the p_i to be used to find χ_i^2 .
      - The χ_i^2 test can be used to test if a sample matches a specific continuous probability distribution
        - To do this the cell properties of H will be give by:
          - $p_i = P(a_{i-1} \leq X < a_i) = \int_{a_{i-1}}^{a_i} f(x)\,dx$

Two-Way Contingency Tables

- This is used when there are multiple rows of information per column of information
  - Previously we looked at data where multiple columns of information were tested for a single row of information
  - Now we are looking at multiple rows of information per column of information
    - There are I rows ($I \geq 2$)
    - And J columns
    - Therefore there are IJ cells
- When this might occur:
  - 1. There are multiple populations being tested across multiple categories
    - When this is the case we want to test if the populations are *homogenous* with respect to these categories
  - 2. There is a single population being tested for two different factors, each of which have multiple categories
    - When this is the case we want to investigate whether the categories of the two factors occur *independently* in the population
      - We will focus on this case, and will use a chi-squared test again
- ***Testing for independence***
  - Each individual in the population is assumed to belong in exactly one of the I categories and exactly one of the J categories
  - The null hypothesis would be that which category I an individual is apart of will have no impact on which category of J they are apart of, and vice versa.
    - To test this we compare observed cell counts with expected cell counts under H_0.
      - expected cell count = $\frac{\text{row total} \times \text{column total}}{n}$
      - $X^2 = \sum \frac{(\text{observed count} - \text{expected count})^2}{\text{expected count}} \sim \chi^2_{(I-1)(J-1)}$
  - How to do this in R
    - 1. Set up the data as a table
      - variablename = as.table(rbind(c(x, x, x,), c(x, x, x)))
    - 2. Create category descriptors for the table
      - dimnames(variablename) <- list(column labels = c("x", "x", "x"), row labels = c("x", "x", "x"))
    - 3. Run the test
      - chisq.test(variablename)
        - *This is a pre-made function in R to run Pearson's Chi-Squared test
        - This will give you X-squared = x, df = x, p-value = x