AMS 394: Statistical Laboratory

Project Description

Due December 2, 2022, 11:59pm

Kai Li

Fall 2022

1 Description

The purpose of this class is to translate realistic research problems into a statistical context and perform the analysis using R. We have covered a wide range of basic and intermediate level statistical procedures and their applications in R. Based on the realistic data at hand, it is still likely that none of the statistical techniques covered in this course is the best statistical procedure for the data set. In this project, we consider an experimental design based on independent samples and (complete) matched pairs designs, called partially matched pairs design.

For example, in clinical studies, a medical researcher may be interested in comparing two methods, call them methods A and B, for measuring cardiac output. If no missing value is present in the samples, the samples are called (complete) matched samples. Regular hypothesis testing procedures for (complete) matched pairs design, such as the paired t test method, can be applied to test the null hypothesis of the mean difference between two sets of paired observations. However, data has missing values for whatever reason in practice. Appropriate statistical procedures are required for partially matched samples.

Each student in the class will write a function implementing appropriate statistical procedures that we have covered, including Kim et al.'s modified t-statistic and Looney and Jones's corrected Z-test that is going to be introduced, for partially matched samples in R.

2 Background

We have studied independent samples and (complete) matched samples throughout the semester. We can choose the correct statistical method to make inferences on the population(s) by following the below steps:

- One population: one-sample test
    1. Mean: one-sample t test and Wilcoxon signed-rank test
        (a) Normal: one-sample t test
        (b) Non-normal: Wilcoxon signed-rank test
    2. Proportion: one-sample proportion test
        (a) Two categories/cells: one-sample proportion z test and Binomial exact test
            i. Large sample size: one-sample proportion z test

ii. Small sample size: Binomial exact test

(b) $k > 2$ categories/cells: the chi-squared test

- Two populations: two-sample test

1. Mean: two-sample t test and two-sample Wilcoxon test
   (a) Independent samples: independent (unpaired) two-sample test
      i. Normal: two-sample t test
         A. Equal variances: the pooled two-sample t test
         B. Unequal variances: Welch's two-sample t test
      ii. Non-normal: two-sample Wilcoxon test
   (b) Matched pairs: paired (dependent) two-sample test
      i. Normal: the paired t test
      ii. Non-normal: the matched-pairs Wilcoxon test
2. Proportion: two-sample proportion test
   (a) Two categories/cells: two-sample proportion z test and Fisher's exact test
      i. Large sample size: two-sample proportion z test
      ii. Small sample size: Fisher's exact test
   (b) $k > 2$ categories/cells ($r \times c$ tables): the chi-squared test

- $k > 2$ populations: k-sample test

1. Mean: ANOVA and the Kruskal-Wallis test
   (a) Independent samples: ANOVA and the Kruskal-Wallis test
      i. Normal: ANOVA
         A. Equal variances: the pooled ANOVA
         B. Unequal variances: Welch's ANOVA
      ii. Non-normal: the Kruskal-Wallis test
   (b) Matched pairs: repeated measures ANOVA
2. Proportion ($r \times c$ tables): the chi-squared test

For simplicity, let us consider inference on two populations. Recall that the data from an independent samples design are of the following form:

Sample 1	x_1	x_2	$\cdots$	x_{n_1}
Sample 2	y_1	y_2	$\cdots$	y_{n_2}

The data from a (complete) matched pairs design are of the following form:

Pair	1	2	$\cdots$	n
Sample 1	x_1	x_2	$\cdots$	x_n
Sample 2	y_1	y_2	$\cdots$	y_n

3 Partially Matched Samples Data Structure

Mathematically speaking, suppose that a random sample of size $2n$ of paired cardiac output, i.e., n pairs of matched samples of cardiac output with missing values, are drawn from a subject population. Suppose that n_1 pairs of the sample are entirely matched. That is, the n_1 subject pairs do not have any missing cardiac output values. Furthermore, suppose that n_2 and n_3 subject pairs have only missing cardiac output values for methods A and B, respectively. Then, the data for measuring cardiac output is an example of partially matched data. It is also possible that n_4 pairs of the cardiac outputs are all missing. This scenario will not change the inherent characteristics of partially matched samples because the n_4 pairs of missing values can be omitted directly.

The data from a partially matched pairs design are of the following form. Note that pairs are typically not ordered like the following:

Pair	1	2	$\cdots$	n_1	n_1+1	$\cdots$	n_1+n_2	n_1+n_2+1	$\cdots$	$n_1+n_2+n_3$	$n_1+n_2+n_3+1$	$\cdots$	n
Sample 1	x_1	x_2	$\cdots$	x_{n_1}	x_{n_1+1}	$\cdots$	$x_{n_1+n_2}$	NA	$\cdots$	NA	NA	$\cdots$	NA
Sample 2	y_1	y_2	$\cdots$	y_{n_1}	NA	$\cdots$	NA	$y_{n_1+n_2+1}$	$\cdots$	$y_{n_1+n_2+n_3}$	NA	$\cdots$	NA

Pair n is, in fact, Pair $n_1+n_2+n_3+n_4$. Method A and Method B correspond to Sample 1 and Sample 2, respectively, in the cardiac output example. Denote the subset of data `data1` with the n_1 fully matched pairs, the subset of data `data2` with the n_2 Sample 1 observations paired with the missing values in Sample 2, the subset of data `data3` with the n_3 Sample 2 observations paired with the missing values in Sample 1, and the subset of data `data4` with the n_4 matched NA observations.

Recall from Section 1 that partially matched samples are based on independent samples and (complete) matched samples. Specifically, the partially matched pairs design can be considered as a combination of the following two experimental designs:

- n_1 matched pairs or repeated measures (`data1`)
- independent groups with n_2 and n_3 per group (Sample 1 from `data2` and Sample 2 from `data3`), where both group's experimental designs intend to estimate the same parameter (e.g., the difference of means of the two groups is 0)

Again, `data4` with n_4 pairs of missing values is discarded from our analysis since it is only meaningful for the completeness of inspecting the data structure.

Consider the most general case of partially matched samples, where we have $n_1 > 0$, $n_2 > 0$, $n_3 > 0$ in the given data. There are two options if we only consider the basic data analysis procedures to analyze partially matched samples:

- Analyze the n_1 matched samples `data1` using paired-sample methods.
- Treat n_1+n_2 samples (the first samples of the combined data of `data1` and `data2`) and n_1+n_3 samples (the second samples of the combined data of `data1` and `data3`) as two independent samples.

Kuan and Huang [2] claim that the above approaches are not the best choices because paired-sample methods for a subset of data do not use all the information given in the original data, whereas considering the partially matched samples as two independent samples may lose the inherent pair correlation structure between the matched samples. Kuan and Huang's paper [2] summarizes five statistical analysis methods used for analyzing partially matched samples. We study the details of two of them.

4 Statistical Analysis Methods for Partially Matched Samples

4.1 Kim et al.'s Modified t-Statistic

Let $\overline{D}$ and S_D be the mean and the standard deviation of the difference of Sample 1 and Sample 2 in `data1`, respectively. Also, let $\overline{T}$ and S_T be the mean and the standard deviation of Sample 1 in `data2`, respectively. Similarly define $\overline{N}$ and S_N for Sample 2 in `data3`, respectively. Moreover, let n_H be the harmonic mean of n_2 and n_3. Then, Kim et al.'s modified t-statistic [1] is given by

$$t_3 = \frac{n_1\overline{D} + n_H(\overline{T} - \overline{N})}{\sqrt{n_1 S_D^2 + n_H^2(S_T^2/n_2 + S_N^2/n_3)}}. \tag{1}$$

Note that the distribution of t_3 approximately follows a standard normal distribution under the null hypothesis [2]. Therefore, p-values and confidence intervals can be obtained from t_3.

4.2 Looney and Jones's Corrected Z-Test

Let $\overline{T}^*$ and S_T^* be the mean and the standard deviation of Sample 1 in the combined data of `data1` and `data2`, respectively. Similarly, define $\overline{N}^*$ and S_N^* for Sample 2 in the combined data of `data1` and `data3`, respectively. Moreover, let S_{TN_1} be the sample covariance of Sample 1 and Sample 2 in `data1`. Then, Looney and Jones's corrected Z-test [3] is given by

$$Z_{\text{corr}} = \frac{\overline{T}^* - \overline{N}^*}{\sqrt{{S_T^*}^2/(n_1+n_2) + {S_N^*}^2/(n_1+n_3) - 2n_1 S_{TN_1}/[(n_1+n_2)(n_1+n_3)]}}. \tag{2}$$

Note that the distribution of Z_{corr} follows a standard normal distribution. Therefore, p-values and confidence intervals can be obtained from Z_{corr}.

5 Grading Rubric

1. (40 points) Correctness
   - Any function submitted that does not compile and run **will receive an automatic grade of zero on this criterion**. No exceptions will be made for this rule. To achieve even a single point on this criterion, your function must minimally build (compile to an executable) in R.
   - The function should work correctly on all arguments/inputs and meet all the specifications.
2. (30 points) Completeness
   - The function should show evidence of excellent case analysis, and all possible cases for partially matched samples are handled appropriately.
3. (10 points) Readability[1]
   - Variables, expressions, and function(s) should have meaningful names.
   - The code should be clean, easy to follow, and well-organized into chunks where appropriate.

[1] Negative scores are possible for this section.

- There should be an appropriate amount of white space so that the code is readable, and indentation should be consistent.

4. (5 points) Documentation

   - The code should be appropriately commented and clearly explains what the code is accomplishing and how.
   - However, not every line should be commented because that makes your code overly busy. Think carefully about where comments are needed.

5. (10 points) Modularity

   - The function should be decomposed into coherent and reusable units, and unnecessary repetition should be removed.
   - There are many ways to write the same functionality into your code, and some of them are needlessly slow or complicated. For example, if you are repeating the same code, it should be inside creating a new function, for loop, or consider using vectorized operations.
   - Hard coding should be minimal. The code could be reused as a whole, or each routine could be reused.

6. (5 points) Efficiency

   - The code is extremely efficient without sacrificing readability and understanding.

References

[1] B. S. Kim, I. Kim, S. Lee, S. Kim, S. Y. Rha, and H. C. Chung. Statistical methods of translating microarray data into clinically relevant diagnostic information in colorectal cancer. *Bioinformatics*, 21(4):517–528, 2004.

[2] P. F. Kuan and B. Huang. A simple and robust method for partially matched samples using the p-values pooling approach. *Statistics in medicine*, 32(19):3247–3259, 2013.

[3] S. W. Looney and P. W. Jones. A method for comparing two normal means using combined samples of correlated and uncorrelated data. *Statistics in Medicine*, 22(9):1601–1610, 2003.