Enrichment Project #1: Rank-based Methods

Yuan Gao, Kevin Lee, Akshay Govindaraj
Yijun (Emma) Wan, Peter Williams, Ruixuan Zhang
ygao390, kylee20, ywan40, agovindaraj6, pwilliams60, rzhang438 | @gatech.edu

01 September 2018

1) Two-sample Studies (40%):

Locate a data set in the field of your interest, e.g., eCommerce, medical study, drug development, supply-chain/logistics operations, for applying the following procedures for two-sample studies.

For our two-sample study, we have selected a dataset published by the City of Chicago Transit Authority (CTA) that details daily ridership numbers (system-wide boarding totals) for both bus and rail services. As a pre-processing step, we take a subset of the original dataset that consists of weekday ridership only (weekends, and holidays excluded) for 2017 and 2018. Below is a quick preview of the dataset and pre-processing steps:

```r
rider_data <- read.csv(
  'data/CTA_-_Ridership_-_Daily_Boarding_Totals.csv',
                       header = T,stringsAsFactors = F)
#subset data to just weekdays
rider_data <- rider_data[rider_data$day_type == 'W',]
rider_data <- transform(rider_data, service_date = as.Date(service_date, '%m/%d/%y'))
#subset data to 2017 and 2018 rides only
rider_data <- rider_data[format(rider_data$service_date, '%Y') %in% c('2017','2018'),]
knitr::kable(head(rider_data,n=5),row.names = F)
```

service_date	day_type	bus	rail_boardings	total_rides
2017-02-01	W	883845	750500	1634345
2017-02-02	W	829125	736551	1565676
2017-02-03	W	782899	690415	1473314
2017-02-06	W	838312	713858	1552170
2017-02-07	W	827885	734586	1562471

1. Calculate Pearson and Spearman coefficient of correlation and Kendall's Tau. Use a Bootstrap resampling procedure with B = #bootstrap-samples = 1000 to assess the standard deviation (sd) of three estimates. Comment on your findings.

To begin we compute all three statistics using methods in base R, reported here:

```
## [1] "Pearson's Correlation: 0.858"
```

```
## [1] "Spearman's Correlation: 0.657"
```

```
## [1] "Kendall's Tau: 0.499"
```

We then compute these statistics using $B = 1000$ bootstrap samples based on our original dataset, which consists of 318 observations. Code and details are below. As shown in the table of results, bootstrap sample statistics for spearman's rank were more dispersed than for other statistics.

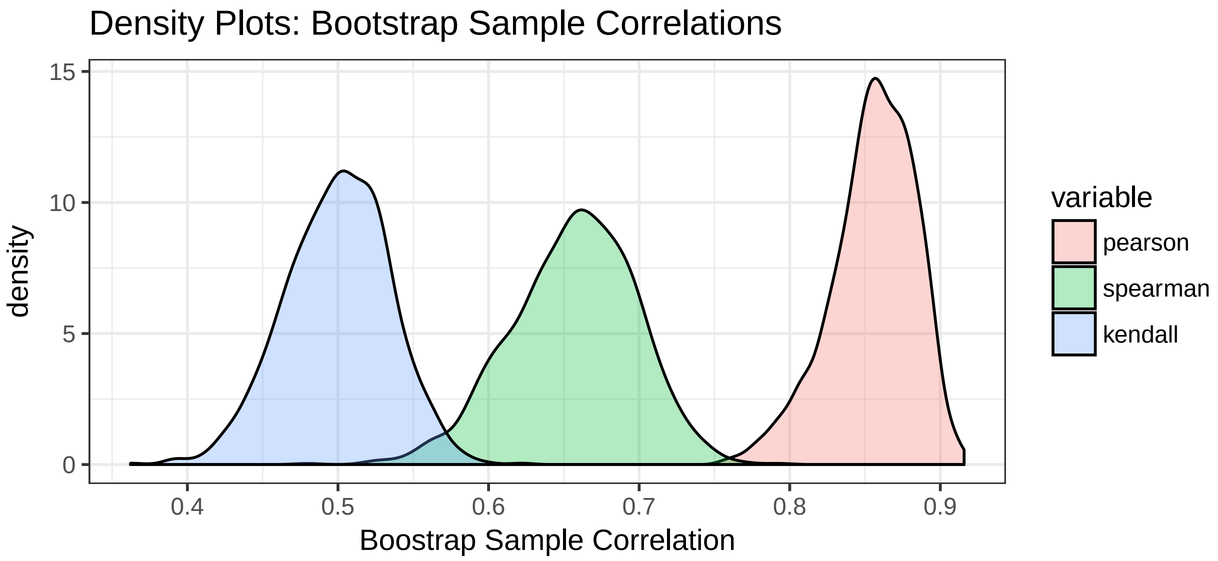
```r
#set the number of bootstrap samples (with replacement - n = 318)
B <- 1000
#replicate process of collecting sample from original dataset, and compute sample stats
#results stored in: cor_result, with 1000 obs of bootstrap sample stats
#note this process takes a few seconds on a standard laptop
cor_result <- do.call('rbind',lapply(1:B, function(x){
  bootstrap_sample <- rider_data[sample(1:nrow(rider_data),
                                        size=nrow(rider_data), replace=T),]
  data.frame(pearson = cor(bootstrap_sample$bus,bootstrap_sample$rail_boardings,
                          method = 'pearson'),
             spearman = cor(bootstrap_sample$bus,bootstrap_sample$rail_boardings,
                           method = 'spearman'),
             kendall = cor(bootstrap_sample$bus,bootstrap_sample$rail_boardings,
                          method = 'kendall'))
}))
knitr::kable(round(data.frame(lapply(cor_result, sd)),digits=4),row.names=F,
             caption = 'Sample Standard Deviation: Bootstrap Sample Correlation')
```

Table 2: Sample Standard Deviation: Bootstrap Sample Correlation

pearson	spearman	kendall
0.0278	0.0412	0.0344

Density plots of these statistics are shown here, further visual confirmation that the spearman rank correlation results are more dispersed than others:

```r
library(ggplot2)
library(reshape2)
melt_cor_result <- melt(cor_result, measure.vars = c('pearson','spearman','kendall'))
ggplot(melt_cor_result, aes(x = value, fill = variable)) + geom_density(alpha = 0.3) + theme_b
```

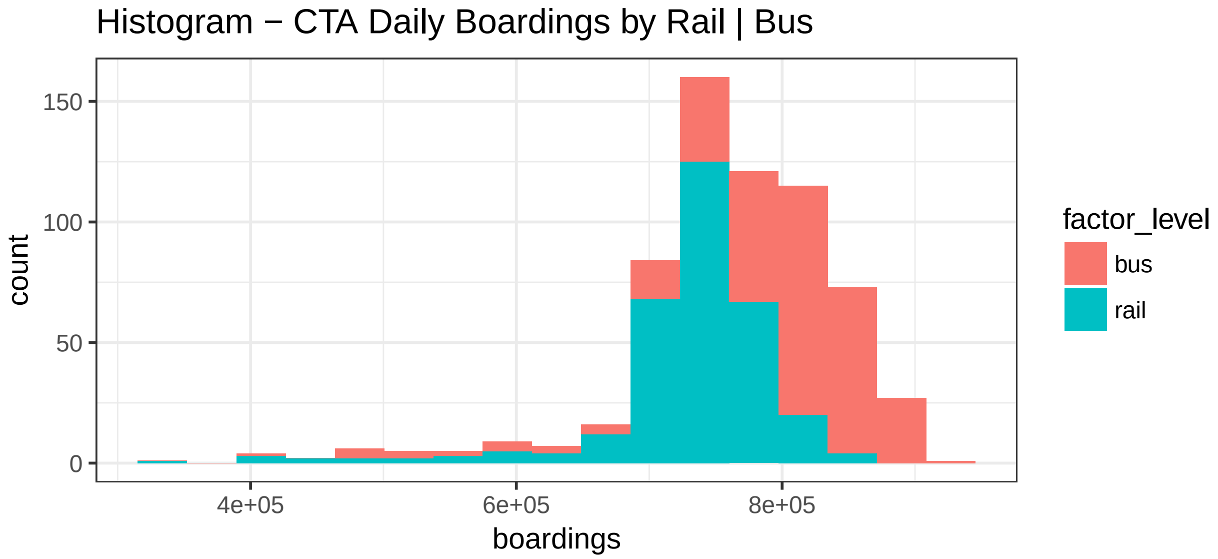


2. Apply Wilcoxon Signed Rank Test, Wilcoxon Sum Rank Test, Mann-Whitney U Test to compare two samples. For each test please state clearly what distribution is used to calculate the p-value.

3. Use Conover test for equal variances in these two samples. Explain how to calculate its p-value.

A histogram visualizing the counts of daily bus and rail ridership illustrates that there are likely differences in the variance of daily ridership. Rail ridership tends to be more tightly group around its central point, than bus ridership which is more spread out as shown here:

```
boardings <- c(rider_data$bus, rider_data$rail_boardings)
factor_level <- factor(c(rep('bus',nrow(rider_data)), rep('rail',nrow(rider_data))))
ggplot(NULL,aes(x = boardings, fill = factor_level)) +
  geom_histogram(bins = sqrt(nrow(rider_data)) ) + theme_bw() +
  ggtitle("Histogram - CTA Daily Boardings by Rail | Bus")
```



To explore this further we utilize the two-sided Conover test to test the hypothesis $H_0 : \sigma_X^2 = \sigma_Y^2$, where samples $X_1, ..., X_{n1}$, and $Y_1, ..., Y_{n_2}$ refer to samples from daily ridership of bus and rail. The output of the test utilizing the 'conover.test' function from R is shown below:

```
#alt = 0 -> test hypothesis sigma^2_x \neq sigma^2_y
con_test <- conover.test(x = boardings, g = factor_level,alpha = 0.05, altp = 0, method = 'bon
```

```
##   Kruskal-Wallis rank sum test
##
## data: boardings and factor_level
## Kruskal-Wallis chi-squared = 179.3705, df = 1, p-value = 0
##
##
##                        Comparison of boardings by factor_level
##                                      (Bonferroni)
## Col Mean-|
## Row Mean |        bus
## ---------+-----------
##     rail |   15.79844
##          |    0.0000*
##
## alpha = 0.05
## Reject Ho if p <= alpha/2
```

```
con_test
```

```
## $chi2
## [1] 179.3705
##
## $T
## [1] 15.79844
##
## $P
## [1] 5.931375e-48
##
## $P.adjusted
## [1] 5.931375e-48
##
## $comparisons
## [1] "bus - rail"
```

Since our test statistic $T* = $ 15.798, is outside the region (-1.959964,1.959964), we can reject H_0 in favor of the alternative, that is $\sigma^2_X \neq \sigma^2_Y$.

4. Use the parametric F-test for equal variances to the data; comment on the difference of the assumptions and results compared to them in (iii).

5. Depending on the outcomes from the F-test in (iv), apply an appropriate parametric two- sample t-test to the data; comment on the difference of the assumptions and results compared to them in (ii).

6. Apply Kolmogorov-Smirnov, Anderson-Darling, Cramer-Von Mises, Shapiro-Wilk Tests for normality to the two samples separately; comment on the findings by comparing results obtained from these four tests. Make a statement about the situation that a particular procedure might be *more appropriate. Moreover, based on the results learned here, comment on whether the parametric methods used in (iv) and (v) are appropriate.*

2) Multiple-Sample (ANOVA) Studies (60%):

Locate one data set each for the two problems below in the field of your interest, e.g., eCommerce, medical study, drug development, supply-chain/logistics operations, for applying the following procedures for ANOVA studies.

1. Apply Kriskal-Wallis Test for an one-way ANOVA study. If it is suitable, perform a K- W pairwise comparisons. Make conclusions about your findings.

2. Use Friedman test and also the F-Test discussed in the textbook page 148 for the study of one-way ANOVA with one blocking variable. Comment on your findings. If it is suitable, perform a K-W pairwise comparisons. Make conclusions about your findings.

3. Conduct a variance test based on the procedure (Conover test) given in Section 8.3 textbook. Comment on your findings.

4. Repeat the same studies in (i), (ii) and (iii) using parametric approaches (also include the possible pairwise comparisons). State the assumptions needed for the parametric approaches. Compare the results here against those in (i), (ii) and (iii), respectively. Note that if there are certain assumptions (e.g., normality and equal-variance) required in the parametric studies, please apply appropriate procedures to "test" the assumption.