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

2018-09-24

Contents

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:

service_date	day_type	bus	rail_boardings	total_rides
2017-02-01	W	574,425	588,036	1,162,461
2017-02-02	W	865,433	695,663	1,561,096
2017-02-03	W	905,953	910,906	1,816,859
2017-02-06	W	740,684	373,851	1,114,535
2017-02-07	W	847,219	911,476	1,758,695

Note: For the purpose of comparison, we transformed the data to normalize the distributions for equal variances between them. This will allow us to perform both nonparametric and parametric tests to determine whether the two sample distributions are different. The variances of the samples were unequal, so we cannot nonparametric tests such as the Wilcoxon Rank Sum Test. With our new data, the variances are close enough that the distributions can then be compared.

Bootstrap Re-sampling

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: 0.647"
```

```
## [1] "Spearman's Correlation: 0.604"
```

```
## [1] "Kendall's Tau: 0.433"
```

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 standard deviations for all correlation statistics were similar, and were slightly higher for spearman's rank:

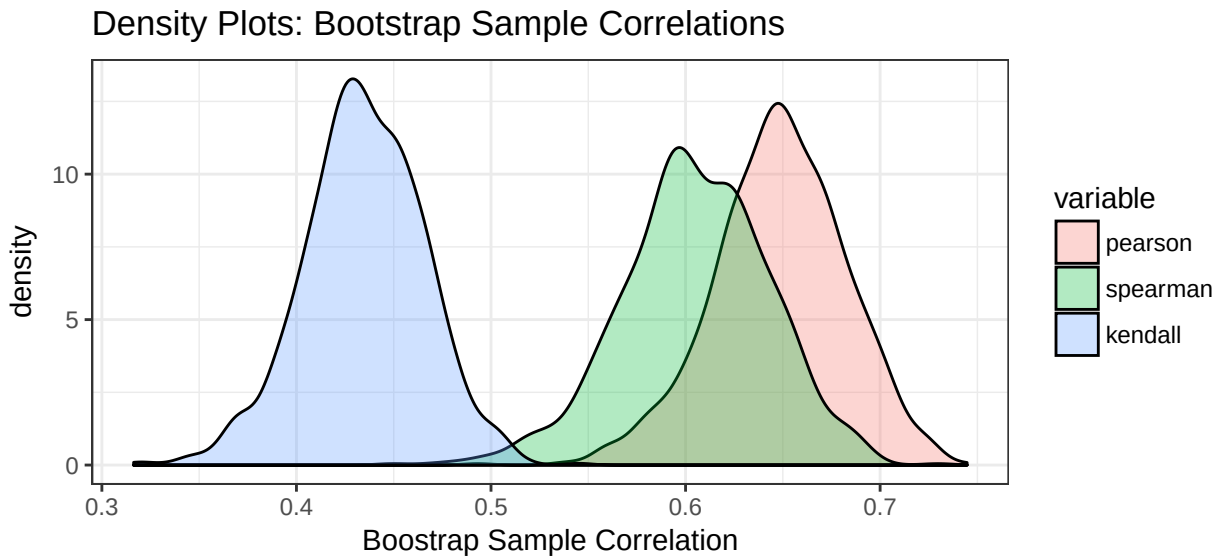
```
#set the number of bootstrap samples (with replacement - n = 318)
B <- 1000
#replicate process of collecting sample from original dataset, and compute sample stats
#results stored in: cor_result, with 1000 obs of bootstrap sample stats
#note this process takes a few seconds on a standard laptop
cor_result <- 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
0.0337	0.0378	0.0307

Density plots of these statistics are shown here, and provide visual evidence that the spearman rank correlation results and the parametrics pearson's correlation have more similar results than the kendall's tau correlation which was consistently lower:

```
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_bw() +
  xlab('Boostrap Sample Correlation') + ggtitle('Density Plots: Bootstrap Sample Correlations')
```



Signed Rank Tests

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.

Continuing on with the data given by the CTA on bus and rail ridership in 2017 and 2018, we can utilize the Wilcoxon Signed-Rank Test to determine whether there is a difference between paired data, but certain assumptions must first be met. In using the Wilcoxon Signed Rank Test, the data must be paired, the pairs must be independent, and the paired differences must be distributed symmetrically about 0 (the assumption that the pairs are drawn from the same population). The samples would also be better if from a non-normal distribution (comparable parametric tests are more reliable only if data are normally distributed).

The Signed Rank Test (for two samples) looks at the difference in distributions between paired values, and our data are not paired in this case. Instead, the better test to use in the case of our CTA data would be the Wilcoxon Ranked Sum Test. The goal of the Wilcoxon Ranked Sum Test is to effectively compare two populations with a continuous response variable and non-normal distributions. Thus, the Wilcoxon Ranked Sum Test is essentially analagous to a nonparametric form of a two sample t-test. Assuming the ridership count between the bus and rail systems are independent of each other and random, we can compare the two samples using the Wilcoxon Ranked Sum Test even without the assumption of paired data as was required in the Wilcoxon Signed Rank Test.

The Mann Whitney U Test accomplishes essentially the exact same purpose as the Wilcoxon Ranked Sum Test. They are even performed using the exact same function in R, with the "U" test statistic being equivalent to the "W" test statistics of the Wilcoxson Ranked Sum Test.

If we plot the density, we can see the distributions seem to resemble normal distributions, and they do resemble each other in terms of their "shape".

```
#Performing the Wilcoxon Rank Sum Test/Mann Whitney U Test
with(rider_data, wilcox.test(bus,rail_boardings, paired=F))
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  bus and rail_boardings
## W = 77742, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```

The results of the Wilcoxon Ranked Sum Test show the p-value of the test to be small enough to be significant, meaning we reject the null hypothesis that the distributions of the two samples are identical. The fact that the Wilcoxon Ranked Sum Test and the Mann Whitney U Test are essentially the same test allows us to also state that the medians of the two distributions are also different.

For the calculation of the p-value, one must first be able to determine the test statistic being used in each test. For the Wilcoxon Rank Sum test, we find the

$$W_n = \sum_{i=1}^{n} iS_i(X, Y)$$

where n_1 is the sample size of the first sample (observations in bus ridership) , n_2 is the sample size of the second sample (observations in rail boarding), and $n_1 + n_2 = n$

S_i is an indicator function with value is 1 if the i^{th} ranked observation is from the first sample or 0 if from the second sample.

The expected value of the W_n statistic is where the distribution should be centered, and the variance of the statistic will be the symmetric around that center:

$$E(W_n) = [n_1(n + 1)]/2$$

$$Var(W_n) = [n_1 n_2(n + 1)]/12$$

Since the test statistic is a linear rank statistic, we can then deduce that the W statistic is distributed approximately normally as long as the sample sizes are large enough (at least 10 observations per sample). We can then calculate the right-tail probability using Z-score with normal approximation with the following distribution:

$$W_n \sim N([n_1(n + 1)]/2, [n_1(n_2)(n + 1)]/12)$$

The Mann-Whitney U statistic is used to calculate p-value much in the same manner of the Wilcoxon Rank Sum Test. The test statistic can be calculated as such:

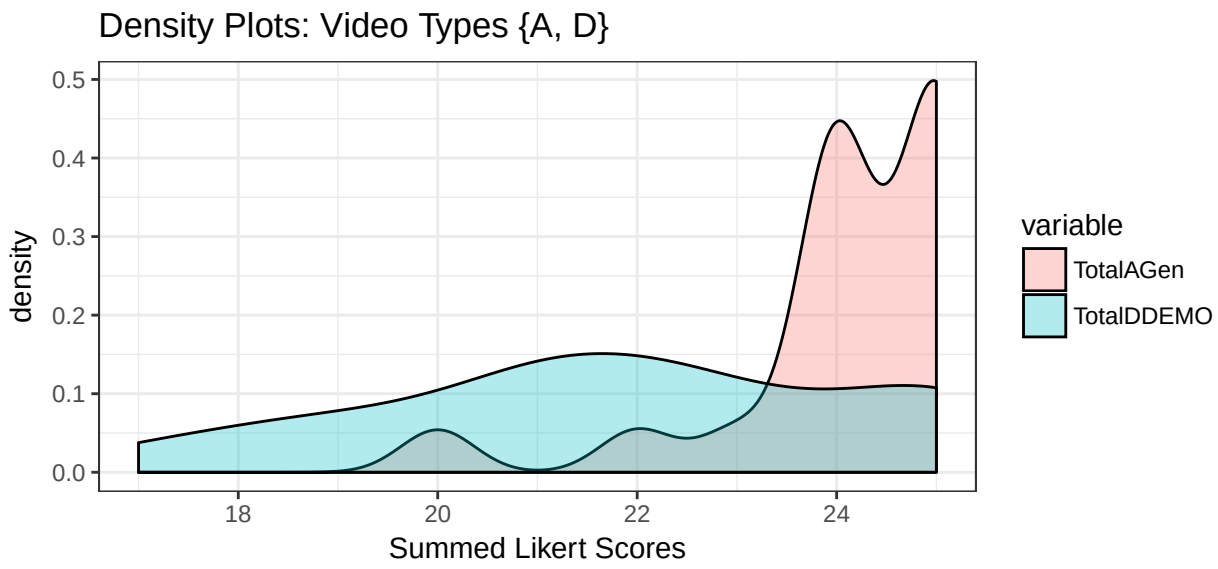
$$U = \sum_{i=1}^{n1} \sum_{j=1}^{n2} D_{ij}$$

The test statisic U and W end up becoming the same value, used in the same approximately normal distribution as the W_n statistic, which means we can use the same tail probability to find the p-value.

In order to apply the Wilcoxon Signed-Rank Test, we will need to use a different set of data to demonstrate the test in R. For this purpose, we will use the data from this site: https://www.sheffield.ac.uk/polopoly_fs/1.569449!/file/stcp-Rdataset-Video.csv.

A professor at the University of Sheffield collected data using "Likert" style questions to determine which of three new videos are most effective in informing the public of medical conditions. The four videos in questions are deemed the following: A, a new general video; B, a new medical profession video; C, the old video; and D, a demonstration using props. In this dataset, there are two particular variables of interest, and they are "TotalAGen" and "TotalDDEMO". These two variables are essentially the overall summed "Likert Scores" for each of the different types of videos. Our paired data will be between the scaled overall scores between video A's reception and video D's reception.

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  video_data$TotalAGen and video_data$TotalDDEMO
## V = 167.5, p-value = 0.003558
## alternative hypothesis: true location shift is not equal to 0
```



We can assume the independence of each observation (person) as well as the fact that they should come from the same population (differences should be symmetric about 0). The data are also paired, so we can follow through with the test now that the assumptions are met. Through the Wilcoxon Signed Rank Test, we can see that the test statistic V = 167.5, which leads us to a p-value below our assumed alpha level of 0.05. Thus, we can reject the null hypothesis that the two sample distributions (Scores of Video A compared to scores of Video D) are the same.

The p-value of the signed rank test is based on a W test statistic and tabulated critical values when dealing with lower sample sizes, such that it can be calculated as follows:

$$W = \sum_{i=1}^{n} sign(x_{2,i} - x_{1,i}, |R_i|)$$

Where R_i is the rank of the observation and the statistic is the formula for the sum of the signed ranks. With this test statistic, one can determine a distribution with

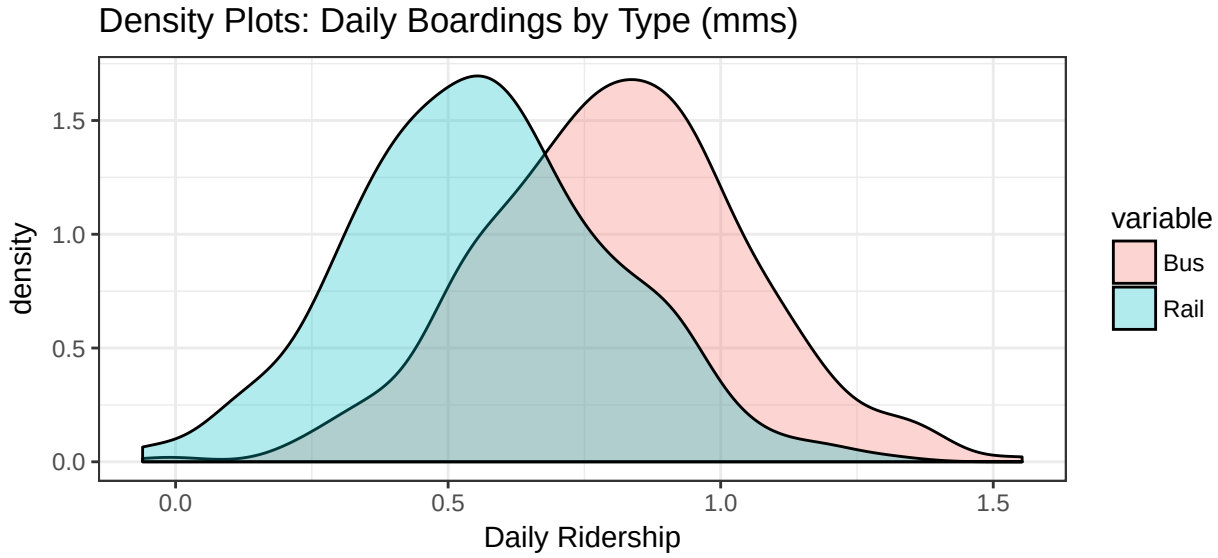
$$E(W) = 0$$
$$Var(T) = [n(n+1)(2n+1)/6]$$

where n is the number of non-tied ranked observations. With this information, one can look at a table of set critical values dependent on sample size and quantiles with which to determine whether or not to reject the hypothesis that the

With larger sample sizes though, the distribution of the W statistic for the signed rank test also tends to follow normal approximation, which allows the use of z-score approximation to determine p-value through tail probability.

Fligner Test for Equal Variances

3. Use Fligner test for equal variances in these two samples. Explain how to calculate its p-value.
We again show a density plot that visualizes the counts of daily bus and rail ridership illustrating that variance of daily ridership seems homogenous across ride modality. Both ridership types also seems to be spread symmetrically around their central point as shown here:



To explore this further we utilize the Fligner 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 'fligner.test' function from R (package::car) is shown below, with our chosen $\alpha = 0.05$:

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  boardings and factor_level
## Fligner-Killeen:med chi-squared = 0.0021818, df = 1, p-value =
## 0.9627
```

Since our test statistic $\chi_K^2* = 0.002$, is less than (3.8414588), we fail to reject our null hypothesis, that is $H_0 : \sigma_X^2 = \sigma_Y^2$.

The calculation of p-value is based on our realized χ^2 statistic which is computed by:

$$\chi_K^2* = \frac{\sum_{j=1}^{k} n_j(\bar{a_j} - \bar{a})^2}{V^2}$$

Where k = number of groups, n_j is the number of observations for the j^{th} group, $\bar{a_j}$ is the mean of the median centered, ranked, and subsequently normalized observations for the j^{th} group. $\bar{a}$ is the mean of all median centered, ranked and normalized observations, and V^2 is the sample variance of the same normalized observations. If the assumptions are met, the distribution of this test statistic follows approximately the Chi-squared distribution with degrees of freedom $k - 1$. We can use the Chi-square distribution to get the p-value of Fligner test given our realized sample statistic.

Parametric F-test For Equal Variances

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).

F-test for testing equality of variance is used to test the hypothesis of the equality of two population X and Y's variances. Let $A_1, ..., A_n$ and $B_1, ..., B_m$ be independent and identically distributed samples from two populations which each have a normal distribution where the expected values for the two populations can be different. The test statistic can be obtained by computing the ratio of the two variances S_A^2 and S_B^2.

$$S_A^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(A_i - \overline{A}\right)^2 \text{ and } S_B^2 = \frac{1}{m-1}\sum_{i=1}^{m}\left(B_i - \overline{B}\right)^2$$

$$F = \frac{S_A^2}{S_B^2}$$

The degrees of freedom are $n_A - 1$ (for the numerator) and $n_B - 1$ (for the denominator). And, the more this ratio deviates from 1, the stronger the evidence for unequal population variances.

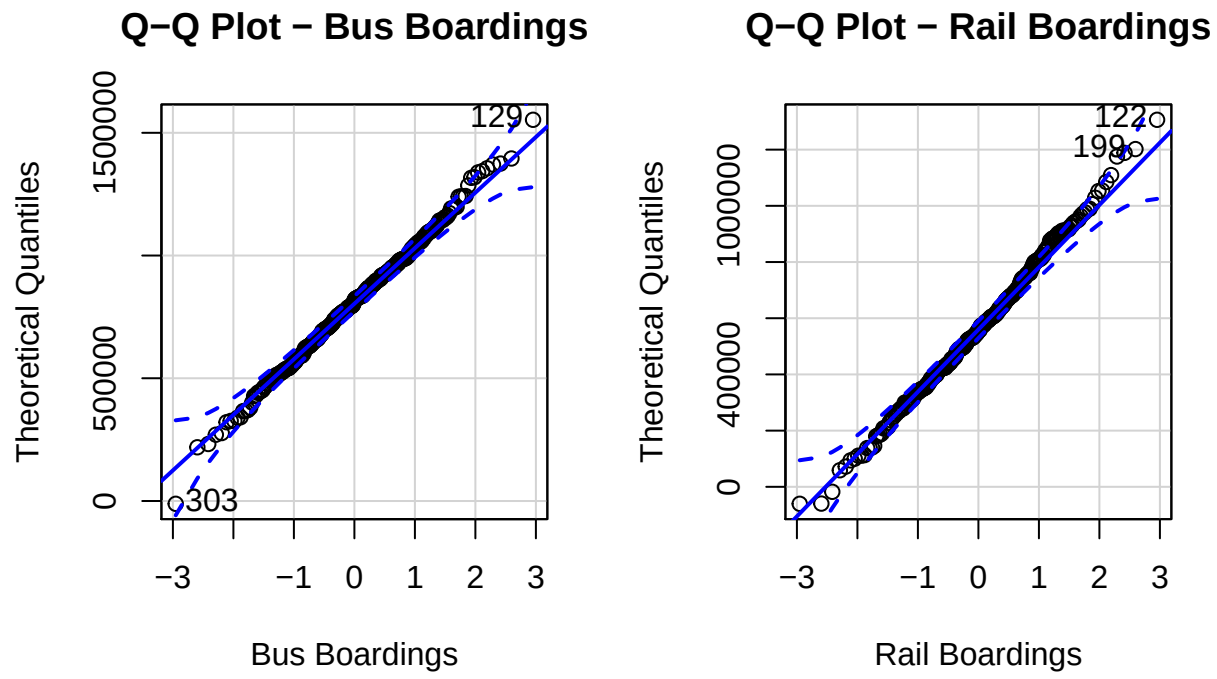
```
##
##  F test to compare two variances
##
## data:  rider_data$bus and rider_data$rail_boardings
## F = 1.0082, num df = 317, denom df = 317, p-value = 0.9422
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.8086292 1.2569753
## sample estimates:
## ratio of variances
##            1.00818
```

The result shows that the p-value = 0.9422 and the ratio = 1.00818 which is significantly greater than our $\alpha = 0.05$ and close to 1. In conclusion, we can accept our null hypothesis that the two data have the same variance.

Assumptions for Conover and F-test:

According to the deviration above,the F-test is known to be extremely sensitive to non-normality and F-tests for the equality of variances can be used in practice, with care, particularly where both graphical and formal checks of the assumption.We use Q-Q plot (quantile-quantile plot) to graphically evaluate the normality of a variable.

```
## [1] 303 129
```




```
## [1] 122 199
```

As we can see, Almost all the points fall into the reference line, for both rail and bus boardings, so, we can assume normality for either sample under Q-Q plot test. Therefore, we can use the F-test to test the equal variance of the two sample.

However, the Conover is a nonparametric test of homogeneity (equal variance) based on ranks. It computes the Conover-Iman test for stochastic dominance and gives the results among multiple pairwise comparisons. The test does not assume that all populations are normally distributed and is recommended when the normality assumption is not viable.

Results for Conover and F-test: Although the results are consistent based on this sample, Conover test is recommended when the normality assumption is not viable. As a result, use F-test on the data has shortcomings compared to use Conover test.

Parametric Two-Sample T-test

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).

The aim of the two-sample t-test is to find the difference between the two sample means in comparing their respective populations. In utilizing the two sample t-test, some assumptions must first be met. One must assume that the observations of our data are independent of each other, similar to the nonparametric tests. Another such assumption is for the response variable to be continuous, approximately normal distribution. Our normalized ridership data allows us to be able to perform the parametric (pooled) two sample t-test. Similar to the nonparametric tests performed (Mann-Whitney U), the data must be continuous.

```
##
##  Two Sample t-test
##
## data:  rider_data$bus and rider_data$rail_boardings
## t = 12.917, df = 634, p-value < 0.00000000000000022
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
```

```
##  206160.5 280083.7
## sample estimates:
## mean of x mean of y
##  807865.8  564743.7
```

As we can see from the results of the test, the p-value is quite low, showing that there is a significant difference between the two means. In comparing the two tests, the parametric t-test tests for significance in difference between the means of the two samples, while the nonparametric tests (Mann-Whitney U) test for significance in differences between the distributions (and medians) of the two samples. In terms of results, both the Mann-Whitney U test and the two sample t-test show a significant difference in mean and median for the two samples. Despite the fact that they both could determine whether there are differences between the two sample distributions, the two sample t-test is a more reliable test if only for the fact that the Mann-Whitney U test is better suited for nonnormal distributions (while we transformed our data to become more normal so as to meet the assumption of normality for the t-test). The t-test is a more reliable test compared to its nonparametric analog when the distributions are approximately normal.

Goodness of Fit Tests

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.

First, Normality test is sensitive to sample size. Small samples most often pass normality tests. Therefore, it's important to combine visual inspection and significance test in order to take the right decision. Second, apply Kolmogorov-Smirnov, Anderson-Darling, Cramer-Von Mises, Shapiro-Wilk Tests for normality as below:

Kolmogorov-Smirnov:

```
ks.test(rider_data$bus,"pnorm",mean(rider_data$bus),sd(rider_data$bus))
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  rider_data$bus
## D = 0.027154, p-value = 0.9731
## alternative hypothesis: two-sided
```

```
ks.test(rider_data$rail_boardings,"pnorm",mean(rider_data$rail_boardings),
        sd(rider_data$rail_boardings))
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  rider_data$rail_boardings
## D = 0.032738, p-value = 0.885
## alternative hypothesis: two-sided
```

Anderson-Darling:

```
ad.test(rider_data$bus)
```

```
##
##  Anderson-Darling normality test
##
## data:  rider_data$bus
## A = 0.16818, p-value = 0.9357
```

```
ad.test(rider_data$rail_boardings)
```

```
##
##  Anderson-Darling normality test
##
## data:  rider_data$rail_boardings
## A = 0.32229, p-value = 0.5267
```

Cramer-Von Mises:

```
cvm.test(rider_data$bus)
```

```
##
##  Cramer-von Mises normality test
##
## data:  rider_data$bus
## W = 0.023255, p-value = 0.9324
```

```
cvm.test(rider_data$rail_boardings)
```

```
##
##  Cramer-von Mises normality test
##
## data:  rider_data$rail_boardings
## W = 0.054268, p-value = 0.4491
```

Shapiro-Wilk:

```
shapiro.test(rider_data$bus)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  rider_data$bus
## W = 0.99794, p-value = 0.965
```

```
shapiro.test(rider_data$rail_boardings)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  rider_data$rail_boardings
## W = 0.99658, p-value = 0.7343
```

Comparison of results:

Although all the above methods (Kolmogorov-Smirnov, Anderson-Darling, Cramer-Von Mises, Shapiro-Wilk Tests) have the same result, i.e. accept H_0, the realized p-value of each test was different.

Appropriate situation for the four tests:

Shapiro-Wilk test is the most powerful test for all types of distribution which is customized from Normal whereas Kolmogorov-Smimov test is the least powerful test which perform well only for uniform distribution that don't need boundary. The performance of Anderson-Darling test is quite comparable with Shapiro-Wilk test. Cramer-Von Mises test is an alternative to the Kolmogorov–Smirnov test. However, the power of Shapiro-Wilk test is still low for small sample size.

Kolmogorov-Smimov is based on the empirical distribution function (ECDF), and the maximum distance between these two curves. So, it is indepedent with the underlying cumulative distribution function being tested. But, it is sensitive on the center of the distribution than at the tails.It is suitable for small samples, ties

are no problem and has omnibus test, but it is low power if prerequisites are not met. The Kolmogorov-Smimov test statistic is defined as

$$D = \max_{1 \le i \le N} (F(Y_i) - \frac{i}{N}, \frac{i}{N} - F(Y_i))$$

The Cramer-Von Mises test is an alternative to the Kolmogorov–Smirnov test.

Anderson-Darling testis used to test samples with a specific distribution. It is a modification of the Kolmogorov-Smirnov (K-S) test and gives more weight to the tails than does the K-S test which sharpens the test. It is high power when testing for normal distribution but is statistic based on squares.But its critical values must be calculated for each hypothesized distribution. The Anderson-Darling test statistic is defined as

$$A^2 = -N - S$$

where

$$S = \sum_{i=1}^{N} \frac{(2i-1)}{N} [\ln F(Y_i) + \ln (1 - F(Y_{N+1-i}))]$$

The Shapiro-Wilk test relies on a test statistic, W, that tests whether a random sample, $x_1, x_2, \ldots, x_n$ comes from (specifically) a normal distribution. It has high power among all tests for normality, but its application is limited to tests for normality only.

The W statistic is calculated as follows:

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Cramer-Von Mises is higher power than KS test, but it's a statistic based on squares. Cramer-Von Mises statistic is defined as:

$$U^2 = T - n(\overline{F} - \frac{1}{2})^2$$

where

$$\overline{F} = \frac{1}{n} \sum F(x_i)$$

As a result, due to our data is normally distributed and we have more than 300 data points, so in our opinion, it is more reasonable to choose Shapiro-Wilk test. Because we normalized our ridership data which allows us to be able to perform the parametric F-test and t-test in (iv) and (v).So, in our case, it is 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.

Kruskal-Wallis Test

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

The Kruskal-Wallis (KW) test is a logical extension of the Wilcoxon-Mann-Whitney test. It is a nonparametric test used to compare three or more samples. It is used to test the null hypothesis that all populations have identical distribution functions against the alternative hypothesis that at least two of the samples differ only with respect to location (median), if at all.

We found a dataset relating to red variants of the Portuguese "Vinho Verde" wine. Input variables are 1 - fixed acidity 2 - volatile acidity 3 - citric acid 4 - residual sugar 5 - chlorides 6 - free sulfur dioxide 7 - total sulfur dioxide 8 - density 9 - pH 10 - sulphates 11 - alcohol Output variable (based on sensory data): 12 - quality (score between 0 and 10).

Data Source: https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009

Often the quality of wine is attributed to the physical properties like density, color. We can use the KW test to check if that is true.

We used 8-density and 12-quality to perform the KW test.

```
##
##  Kruskal-Wallis rank sum test
##
## data:  density by quality
## Kruskal-Wallis chi-squared = 65.329, df = 5, p-value =
## 0.000000000000958
```

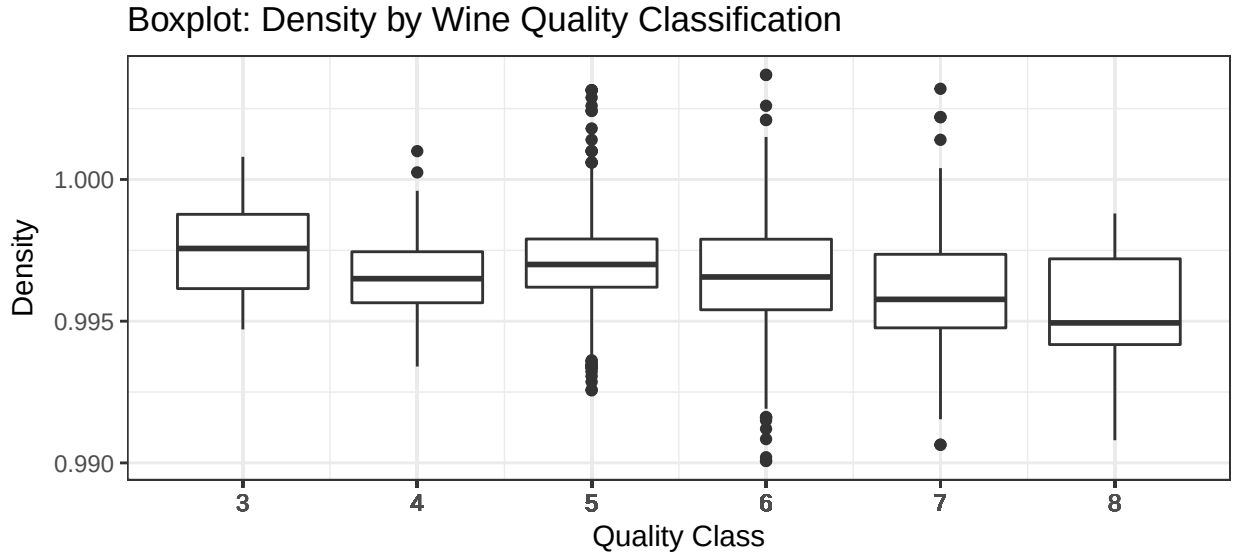
The result shows that p-value is small enough to reject the null hypothesis, which means density and quality do not have identical distribution functions and therefore aspects other than the physical property also contribute to the quality of wine.

Since the KW test detects the differences of red variants, we can determine if two particular red variants group are different at level alpha. Then we performed a K-W pairwise comparisons to the above dataset.

```
##
##  Pairwise comparisons using Conover's-test for multiple
##                           comparisons of independent samples
##
## data:  density and quality
##
##   3      4      5              6      7
## 4 0.8753 -      -              -      -
## 5 1.0000 0.1607 -              -      -
## 6 0.8753 1.0000 0.000006349494 -      -
## 7 0.2021 0.8753 0.000000000038 0.0049 -
## 8 0.1314 0.5016 0.0058         0.1806 0.8753
##
## P value adjustment method: holm
```

The above test gives us more insight into how density on different qualities relate with each other. In many cases wine of certain quality appears to have similar density as other similar quality wine, like in the case of quality 7 and 8, as well as quality 3 and 4. But for the pairs (5,6) and (6,7) there is good reason the believe that they are unrelated and therefore it may be the case that the seemingly related qualities are just coincidental.

To visualize potential differences in the data, we use the boxplot below to show some of the relations in the data described above:



The boxplot highlights that for (3)-citric acid, (4)-residual sugar, (5)-chlorides, and (6)-free sulfur dioxide are normally distributed. (5)-chlorides has lots of outliers.

Friedman Test

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.

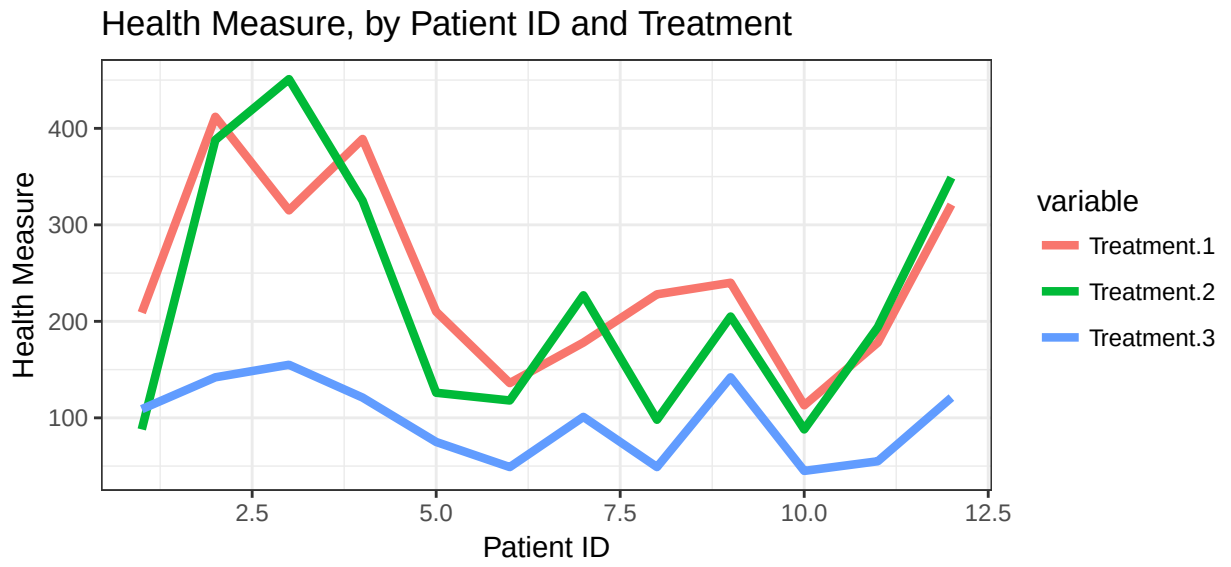
To perform the Friedman test, we utilize a dataset retrieved from a website that provides tutorial's on performance statistical tests in R, that is shown below. The dataset consists of generic health measurements of patients who have been exposed to three different anonymized treatments, shown here:

Table 3: Patient & Treatment Data: 2.2

Patient	Treatment.1	Treatment.2	Treatment.3
1	209	88	109
2	412	388	142
3	315	451	155
4	389	325	121
5	210	126	75
6	136	118	49
7	178	227	101
8	228	98	49
9	240	205	142
10	113	88	45
11	178	194	55
12	321	349	121

Source: http://www.statisticshowto.com/friedmans-test/

A line plot, is shown below, where the x-axis represents each patient id, and the y-axis represents health measures, and the color of each line represents the three treatments on which health measures were collected:



Using the Friedman test will help us identify whether the 3 kinds of treatments are from the same distribution or not. The test setup is:

H_0: Health measures across groups are the same
H_α: At least one measure differs across groups

```
tdat = tdat[,!names(tdat) %in% c("Patient")]
mymatrix = as.matrix(tdat)
result = friedman.test(t(mymatrix))
print(result);
```

```
##
##  Friedman rank sum test
##
## data:  t(mymatrix)
## Friedman chi-squared = 28.125, df = 11, p-value = 0.003097
```

As seen above we get the value of the test statistic (Friedman chi-squared) as 28.125 and the p-value less than 10^{-2}. We reject the null hypothesis in favor of the alternative, that is, we have evidence that at least one pair of the 3 samples differ.

For the F-test, we have:

```
S <- as.numeric(result$statistic)
b <- dim(t(mymatrix))[1];
k <- dim(t(mymatrix))[2];
Fstat <- (b-1)*S/(b*(k-1)-S);
pF <- 1-pf(Fstat,k-1,(b-1)*(k-1))
paste0('F stat: ', round(Fstat,digits = 2),', p-value: ' ,round(pF,digits = 8));
```

```
## [1] "F stat: 11.54, p-value: 0.00000094"
```

The results of the F-Test above, agree with the results of Friedman test that at least one pair of samples differ.

To explore this further we conduct some post hoc pairwise comparison, utilizing a K-W test:

```
##
##  Pairwise comparisons using Conover's-test for multiple
##                           comparisons of independent samples
```

```
##
## data:  value and variable
##
##             Treatment.1 Treatment.2
## Treatment.2 0.24451     -
## Treatment.3 0.00026     0.00479
##
## P value adjustment method: holm
```

As show above, we find that treatment 3, differs significantly from treatment 1, and treatment 2, which is also visually evident from the line graph presented above.

Finally, we also conduct an analysis of variance for this same dataset. Our test setup is:

$H_0 : \mu_1 = \mu 2 = ... = \mu_k$, i.e. means are the same across groups. $H_\alpha : \mu_i \neq \mu_j$ for some $i \neq j$, i.e. observed mean is different for some pairing.

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## tm1          2 150507   75254   8.338 0.00117 **
## Residuals   33 297844    9026
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The results from a one-way ANOVA, agree with the results of Friedman test that the mean treatment differs for at least one pair of the 3 samples of treatments.

Variance Testing

3. Conduct a variance test based on the Fligner procedure discussed in class. Comment on your findings.

Because the Conover Test in R is for Multiple Comparisons, which is not suitable for equal variance testing, we will use Fligner-Killeen Test instead, which can be found in "car" package in R. The Fligner-Killeen (median) Test is a non-parametric test for checking the homogeneity of variance between groups of samples which is very robust against departures from normality. We conduct this test using the red wine quality dataset described in section 2.1. Here is the hypothesis test setup:

Null hypothesis H_0: Variances equals in all groups. Alternative hypothesis H_α: At least one variance is different from others in sample groups.

```
fligner.test(density ~ quality,data=wdat)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  density by quality
## Fligner-Killeen:med chi-squared = 49.395, df = 5, p-value =
## 0.000000001842
```

Our observations of wine density are simply grouped by quality. It turns out that p-value is small and H_0 is rejected, so variances in density are considered inconsistent on different qualities. That is, there is no homogeneity in variance of wine density across quality classifications.

Similarly, we also perform the Fligner test for equal variance testing for the patient/treatment described in section 2.2. The results of the test are printed below:

```
fligner.test(value ~ variable,data=melt_patient)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
```

```
##
## data:  value by variable
## Fligner-Killeen:med chi-squared = 7.012, df = 2, p-value = 0.03002
```

In the dataset, there are 3 types of treatment used to 12 patient and scored relatively. In this case, number of patients become blocks and samples (scores) are grouped by different treatments. The F-K test turns out that H_0 is rejected, which provides evidence that there is not homogeneity in the variance of treatment scores.

Parametric Testing

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.

ANOVA test hypotheses: H_0: the means of the different groups are the same H_α: At least one sample mean is not equal to the others.

Assumptions: The observations are obtained independently and randomly from the population defined by the factor levels. The data of each factor level are normally distributed. These normal populations have a common variance. (Levene's test can be used to check this.)

Normality test - Shapiro test:

H_0: the population is normally distributed. H_α: the population is not normally distributed.

```
with(wdat, shapiro.test(x = quality))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  quality
## W = 0.85759, p-value < 0.00000000000000022
```

```
with(wdat, shapiro.test(x = density))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  density
## W = 0.99087, p-value = 0.00000001936
```

Since the p-value is less than the chosen alpha level ($\alpha = 0.05$)), in both cases, then the null hypothesis is rejected and there is evidence that the data tested are not normally distributed.

Equal-variance test: H_0: Sample groups have common variance. H_α: At least one group has different variance from others. Levene Test for studies in (i)

```
with(wdat, leveneTest(density,factor(quality)))
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##         Df F value         Pr(>F)
## group    5  9.7725 0.000000003274 ***
##       1593
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As a result, p-value in the Levene Test is too small to accept H_0, so we have evidence that the variance in each groups of wine density is not equal, which violates the assumption in parametric test.

In (ii), we got the same conclusion as in (i), H_0 is rejected according to small p-value, the 3 samples of treatments do not have same variance, which is also inconsistent with the equal-variance assumption. The samples have distinct variances, which agrees the conclusion that they do not come from same distribution in 2(i) and 2(ii). Basically, in a conclusion, these data are not suitable for parametric mean test because of violation of basic assumption, and nonparametric approaches perform better in this case.

Leveve Test for studies in (ii):

```
melted_tdat <- melt(tdat, id.vars = NULL)
with(melted_tdat, leveneTest(value,variable))
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value  Pr(>F)
## group  2  4.3513 0.02103 *
##       33
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In (ii), we got the same conclusion as in (i), H_0 is rejected according to small p-value, the 3 samples of treatments do not have same variance, which is also inconsistent with the equal-variance assumption. The samples have distinct variances, which agrees the conclusion that they do not come from same distribution in 2(i) and 2(ii). Basically, in a conclusion, these data are not suitable for parametric mean test because of violation of basic assumption, and nonparametric approaches perform better in this case.

One-way ANOVA test:

$H_0 : \mu_1 = \mu 2 = ... = \mu_k$, i.e. means are the same across groups. $H_\alpha : \mu_i \neq \mu_j$ for some $i \neq j$, i.e. observed mean is different for some pairing.

```
res.aov <- aov(density~factor(quality), data = wdat)
summary(res.aov)
```

```
##                   Df   Sum Sq    Mean Sq F value             Pr(>F)
## factor(quality)    5 0.000230 0.00004594    13.4 0.000000000000812 ***
## Residuals       1593 0.005462 0.00000343
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The output includes the columns F value and Pr(>F) corresponding to the p-value of the test. As the p-value is less than the significance level 0.05, we can conclude that there are significant differences between the groups highlighted with "*" in the model summary. Given our observed test statistic, we can reject our null hypothesis, in favor of our alternative, that observed means are different for some pairing in the data.

Factorial Design of Experiment (with one factor and one blocking variable):

The dataset used in 2.2, related to patient treatments, is used to test normality assumptions for parametric tests:

Here are the R outputs from the Shapiro-Wilk normality test for each of three patient treatments, where we test H_0: population data is is normally distributed$:

```
shapiro.test(x = tdat$Treatment.1)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  tdat$Treatment.1
## W = 0.93584, p-value = 0.4461
```

```r
shapiro.test(x = tdat$Treatment.2)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  tdat$Treatment.2
## W = 0.89272, p-value = 0.1278
```

```r
shapiro.test(x = tdat$Treatment.3)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  tdat$Treatment.3
## W = 0.89512, p-value = 0.1372
```

Based on the results above, we do not have evidence to reject the null hypothesis in any of these three cases, with $\alpha = 0.05$

To take it further, we use a parametric one-analysis of variance, where we test $H_0 : \mu_1 = \mu_2 = \mu_3$, against H_α : the mean of at least one group varies from the others:

```r
res.aov <- aov(value~factor(variable), data = melt_patient)
summary(res.aov)
```

```
##                  Df Sum Sq Mean Sq F value  Pr(>F)
## factor(variable)  2 150507   75254   8.338 0.00117 **
## Residuals        33 297844    9026
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Based on these results, we have evidence, to reject our null hypothesis, that is, a that treatment for at least one group differs from the others.

3) Workload Distribution

Below is a description of tasks and the distribution of work (%) by team member for this project:

Team Member	Task Description
Yuan Gao	Sections 1.4 & 1.6, Code Compilation
Kevin Lee	Description of transformed data, Sections 1.2 & 1.5
Akshay Govindaraj	2.2,2.4, one-way tests, and "DOE"
Yijun (Emma) Wan	2.1, normality tests and 2.4
Peter Williams	Proj. Management, Sections 1.1-1.3, Code Compilation, Debugging
Ruixuan Zhang	2.3 and 2.4 equal variance testing