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

03 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:

```
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.

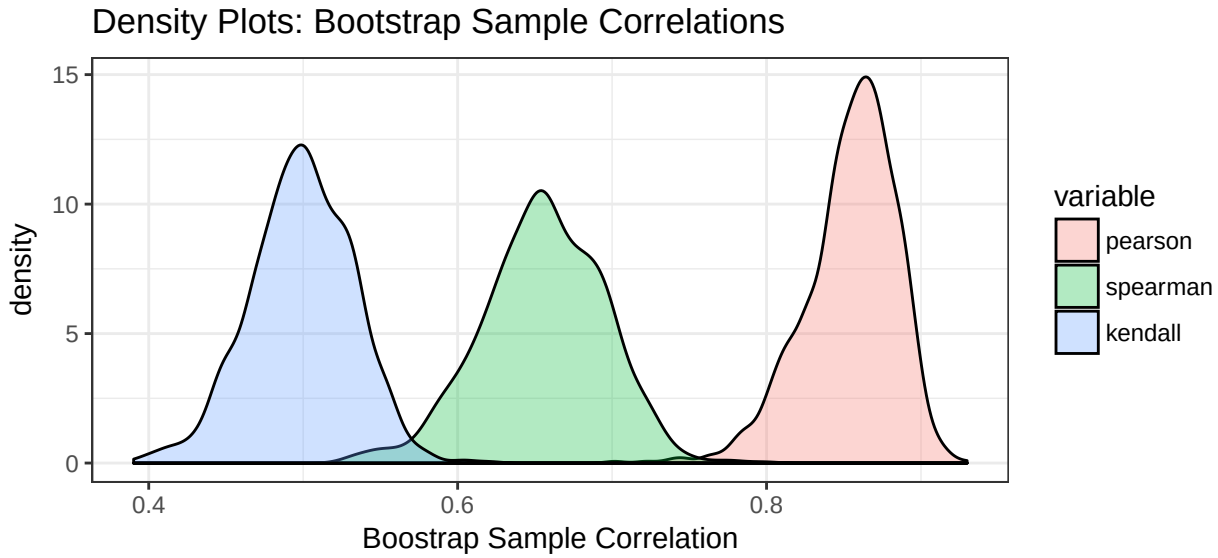
```
#set the number of bootstrap samples (with replacement - n = 318)
B <- 1000
#replicate process of collecting sample from original dataset, and compute sample stats
#results stored in: cor_result, with 1000 obs of bootstrap sample stats
#note this process takes a few seconds on a standard laptop
cor_result <- 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
0.0296	0.0395	0.033

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

```
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_bw
```

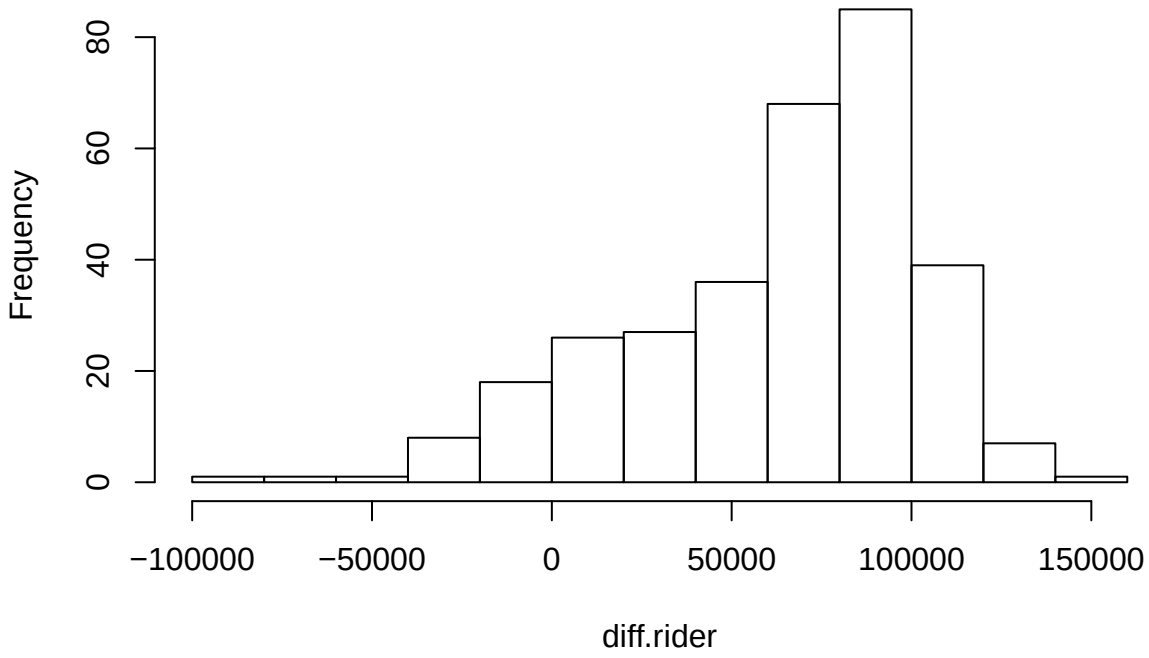


2. Apply Wilcoxon Signed Rank Test, Wilcoxon Sum Rank Test, Mann-Whitney U Test to compare two samples. For each test please state clearly what distribution is used to calculate the p-value. ##Work in progress. Feel free to edit along! ##NOTE: Still need to write distributions used to calculate P-values.

Continuing on with the data given by the CTA on bus and rail ridership in 2017 and 2018, we can attempt to 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 must also be from a nonnormal distribution.

We can see with these histograms that the data is definitely not distributed normally, so at least one of the assumptions thus far is met. Despite this, the Signed Rank Test looks at the median difference between paired values, and our data is not considered "paired". Below is utilization of the Signed Rank Test function on R in spite of the inapplicability of the test.

```
#Checking symmetry of distribution (differences between ridership for a given day)
diff.rider= rider_data$bus-rider_data$rail_boardings
hist(diff.rider)
```




```
#not symmetric around 0, so fails the assumption
#Wilcoxon Signed Rank Test
wilcox.test(rider_data$bus, rider_data$rail_boardings, paired=TRUE)
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  rider_data$bus and rider_data$rail_boardings
## V = 49316, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```

```
# V = 49316, p-value < 2.2e-16
# alternative hypothesis: true location shift is not equal to 0
```

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 nonnormal distributions. 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, without the assumption of paired data as required in the Wilcoxon Signed Rank Test. The Mann Whitney U Test accomplishes essentially the exact same purpose as the Wilcoxon Ranked Sum Test. The only differences are a differently labeled test statistic "U" compared to that of the Wilcoxon ranked sum test statistics "W" as well as the assumption that the shapes of the two population distributions are different (when using tests involving population means). They are even performed using the exact same function, with the "U" test statistic being equivalent to the "W" test statistics of the Wilcoxson Ranked Sum Test.

```
#Wilcoxon Ranked Sum Test
wilcox.test(rider_data$bus, rider_data$rail_boardings, paired=F)
```

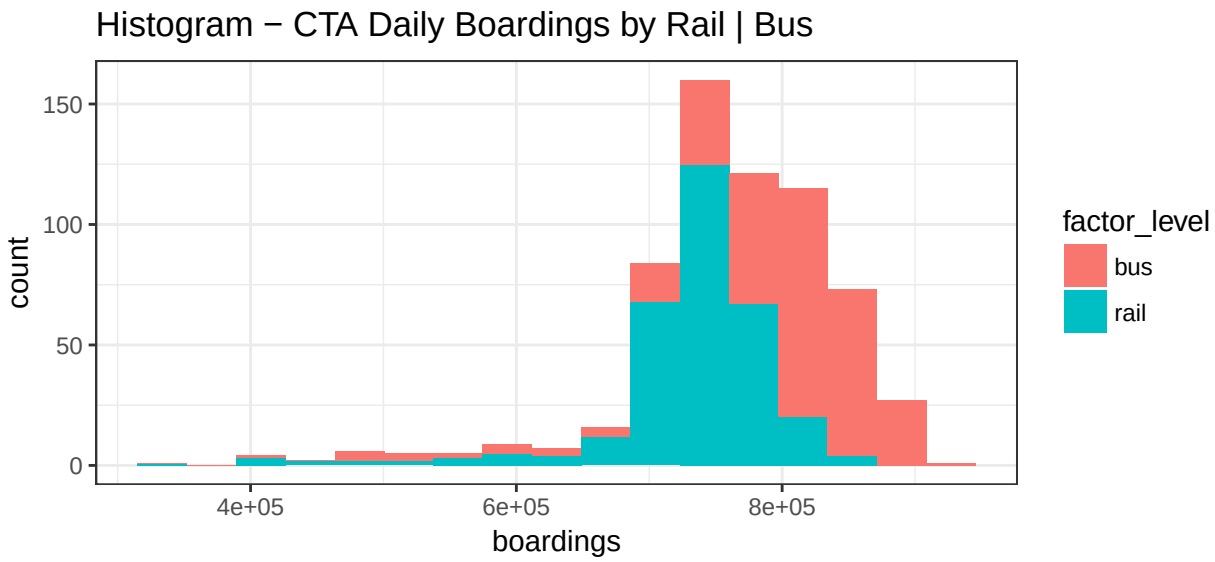
```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  rider_data$bus and rider_data$rail_boardings
## W = 81592, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```

```
#W = 81592, p-value < 2.2e-16
#alternative hypothesis: true location shift is not equal to 0
```

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 $\sigma_X^2 \neq \sigma_Y^2$.

The calculation of p-value: First, the test does not assume that all populations are normally distributed and is recommended when the normality assumption is not viable. Second, Suppose there are g groups obey normal distribution with possibly different means and standard deviations $\sigma_1, \sigma_2, \ldots, \sigma_g$. Let $n_1, n_2, \ldots, n_g$ denote the number of subjects in each group, Y_{ki} denote response values, and N denote the total sample size of all groups. The test assumes that the data are obtained by taking a simple random sample from each of the g populations. The formula for the calculation of Conover test is:

$$T = \frac{1}{D^2}[\sum_{k=1}^{g} \frac{S_k^2}{n_k} - N\overline{S}]$$

Where

$$Z_{ki} = |Y_{ki} - \overline{Y}_k|$$

$$R_{ki} = Rank(Z_{ki})$$

$$S_k = \sum_{i=1}^{n_k} R_{ki}^2$$

$$\overline{S} = \frac{1}{N}\sum_{k=1}^{g} S_k$$

$$D^2 = \frac{1}{N-1}[\sum_{k=1}^{g}\sum_{i=1}^{n_k} R_{ki}^4 - N\overline{S}^2]$$

If the assumptions are met, the distribution of this test statistic follows approximately the Chi-squared distribution with degrees of freedom $g-1$. And then, we can use the p-value of Chi-square distribution to get the p-value of conover test.

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 variances. The test statistic can be obtained by computing the ratio of the two variances S_A^2 and S_B^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.

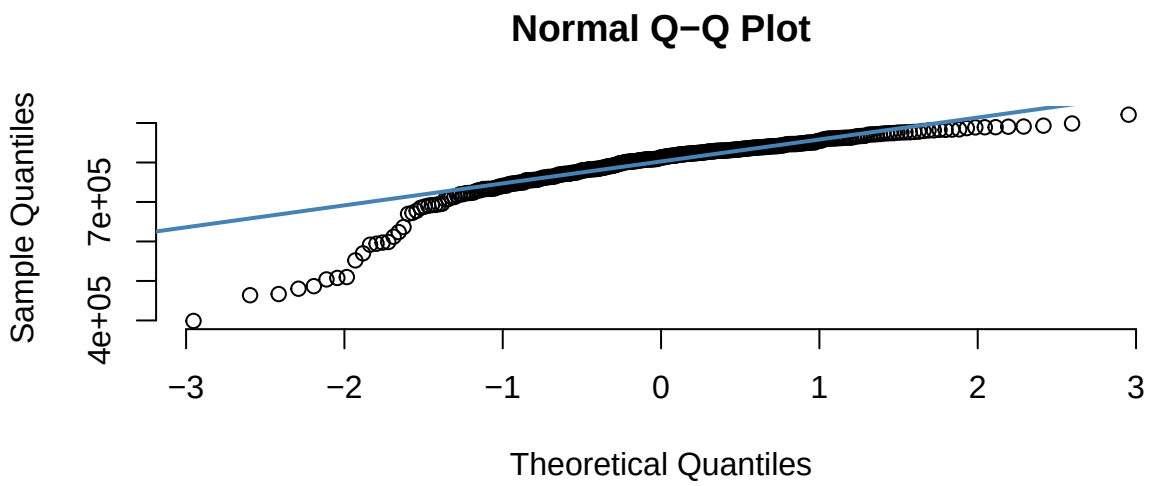
```
var.test(rider_data$bus, rider_data$rail_boardings, alternative = "two.sided")
```

```
##
##  F test to compare two variances
##
## data:  rider_data$bus and rider_data$rail_boardings
## F = 1.306, num df = 317, denom df = 317, p-value = 0.01772
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  1.047533 1.628339
## sample estimates:
## ratio of variances
##           1.306039
```

The result shows that the p-value $= 0.01772$ which is smaller than our $\alpha = 0.05$. In conclusion, there is significant difference between the two variances.

Assumptions for Conover and F-test: According to the above, the Conover test does not assume that all populations are normally distributed. However, F-test is very sensitive to departure from the normal assumption.We use Q-Q plot (quantile-quantile plot) to graphically evaluate the normality of a variable.

```
qqnorm(rider_data$bus, pch = 1, frame = FALSE)
qqline(rider_data$bus, col = "steelblue", lwd = 2)
```



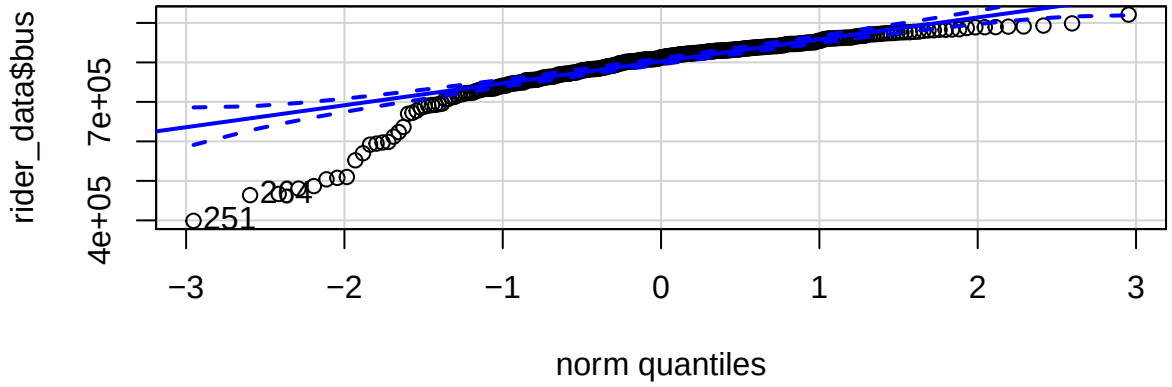

```
library("car")
```

```
## Warning: package 'car' was built under R version 3.4.4
```

```
## Loading required package: carData
```

```
## Warning: package 'carData' was built under R version 3.4.4
```

```
qqPlot(rider_data$bus)
```




```
## [1] 251 264
```

As we can see, there are many points don't fall approximately along this reference line, so, we cannot assume normality.

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.

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.

Kolmogorov-Smirnov:

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

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  rider_data$bus
## D = 1, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

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

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  rider_data$rail_boardings
## D = 1, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

Anderson-Darling:

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

```
##
##  Anderson-Darling normality test
##
## data:  rider_data$bus
## A = 13.108, p-value < 2.2e-16
```

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

```
##
##  Anderson-Darling normality test
##
## data:  rider_data$rail_boardings
## A = 18.67, p-value < 2.2e-16
```

Cramer-Von Mises:

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

Shapiro-Wilk:

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

Comparison of results: All the above methods (Kolmogorov-Smirnov, Anderson-Darling, Cramer-Von Mises, Shapiro-Wilk Tests) have the same result: rejest H0 under our sample. But the p-value

of each test is significantly different which means the measurement for each one is different.

Appropriate situation for the four tests: Shapiro-Wilk test is the most powerful test for all types of distribution whereas Kolmogorov-Smimov test is the least powerful test. 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

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

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. It is high power when testing for normal distribution but is statistic based on squares. The Anderson-Darling test statistic is defined as

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

where

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

Shapiro-Wilk test calculates a W statistic that tests whether a random sample, $x_1, x_2, \ldots, x_n$ comes from (specifically) a normal distribution. It is highest power among all tests for normality but test 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 satistic 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 not normally ditributed and we have more than 300 data points, so in our opinion, it is more reasonable to choose Anderson-Darling test.

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.