Nonparametric Rating and Speed Analysis

Sean O'Malley

Non-Parametric Testing Procedures

Describe the differences between a parametric and a non-parametric statistical approach.

Non-parametric tests are distribution free because unlike parametric analysis, they do not assume that your data follow a specific distribution. To put in laymen's terms, parametric analysis uses the mean to determine insights on group data that is normally distributed; nonparametric analysis uses the median in order to control more easily for non-normal distributions of group data.

What are the advantages of a parametric approach over a non-parametric approach on regression?

Parametric tests can perform well with continuous data that are non-normal if you can satisfy sample guidelines, such as samples of 20 for 1 sample t-test and 15 per group for 2 sample t tests. For non-parametric tests that compare groups, a common assumption is that the data for all groups must have the same spread.

At the end of the day, parametric tests usually have more statistical power, while non-parametric tests, though less predictive overall, do a much better job working with messier, non-normally distributions of data.

What are the disadvantages of a parametric approach?

Nonparametric tests use the median as the best measure of central tendency, thus dealing best with non-normal distributions, small sample size and non-continuous data. Typical parametrics can be significantly affected by outliers. Conversely, some nonparametric tests can handle ordinal data, ranked data, and not be seriously affected by outliers. f your groups have different spread, datatypes, and insufficient sample size the nonparametric tests might not provide valid results.

Souperb: Wilcoxon Signed-Rank Test

Using the Wilcoxon Signed-Rank Test, we will decide whether the corresponding data population distributions are identical without assuming them to follow the normal distribution.

We observe 15 observations from a population of survey responders.

Assumptions:

- n observations are independent
- The observations come from a continuous population which as a median and which is symmetric

Souperb Import Data

```
souperb <- read.xlsx("/Users/SeanOMalley1/Desktop/MSDS_660_Stats/nonparametric_class_dat
a.xlsx",
                     sheetIndex = 1,
                     colClasses = "numeric")
```

Souperb EDA

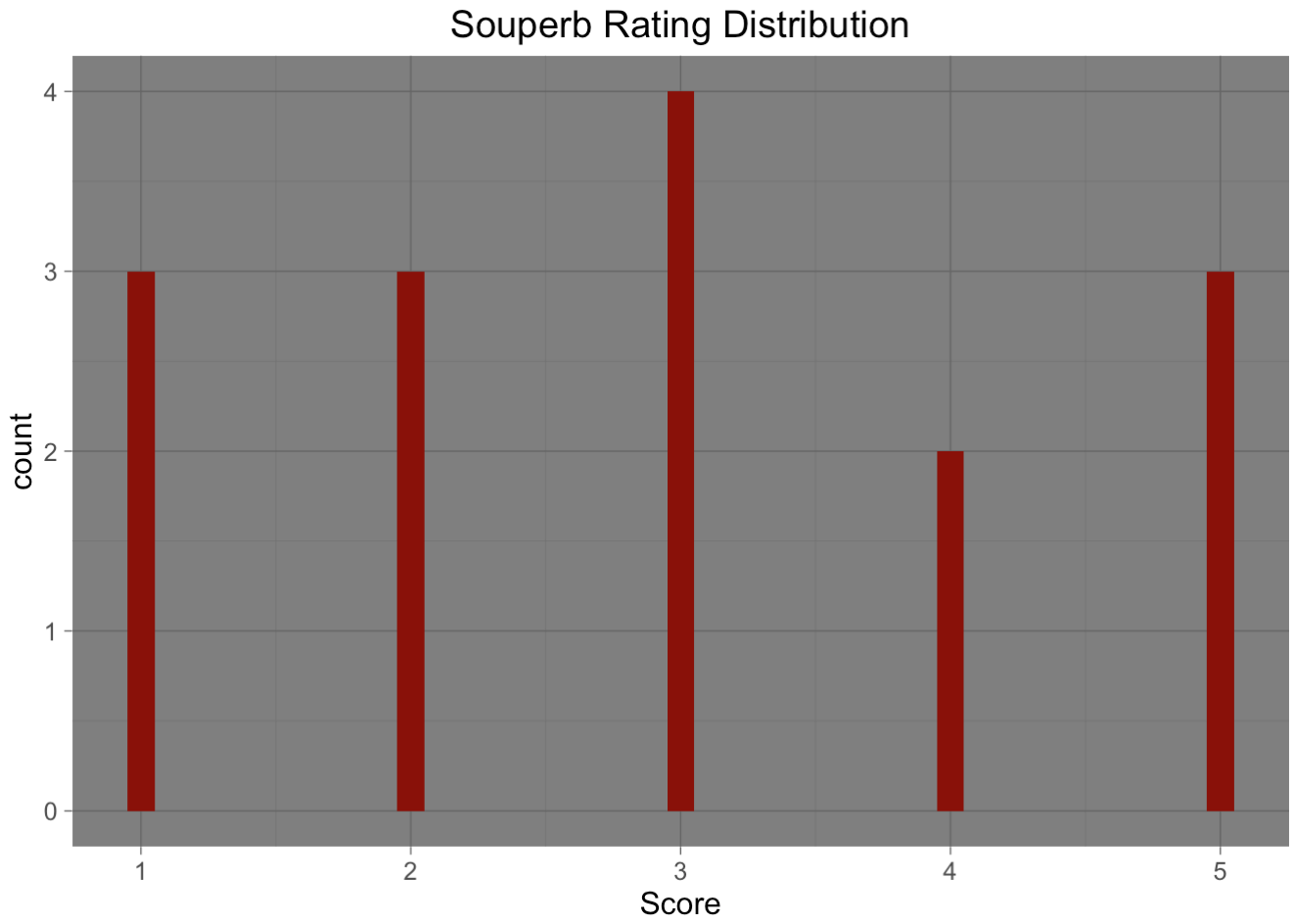
```
glimpse(souperb)
```

```
## Observations: 15
## Variables: 2
## $ Person <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15
## $ Score  <dbl> 5, 3, 2, 1, 4, 3, 5, 1, 5, 2, 3, 4, 2, 1, 3
```

```
summary(souperb$Score)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   3.000   2.933   4.000   5.000
```

```
ggplot(data=souperb, aes(Score)) +
      geom_histogram(binwidth = .1, fill = "#8b0000") +
      theme_dark() +
      ggtitle("Souperb Rating Distribution")
```



Souperb Hypothesis

H^0 : Median is greater than or equal to 3

H^1 : Median is less than 3

Souperb Hypothesis Testing

```
SIGN.test(souperb$Score, md = 3, y = NULL, alternative = "less", conf.level = 0.95)
```

```
##
##  One-sample Sign-Test
##
## data:  souperb$Score
## s = 5, p-value = 0.5
## alternative hypothesis: true median is less than 3
## 95 percent confidence interval:
##  -Inf    4
## sample estimates:
## median of x
##           3
```

```
##                   Conf.Level L.E.pt U.E.pt
## Lower Achieved CI     0.9408   -Inf      4
## Interpolated CI       0.9500   -Inf      4
## Upper Achieved CI     0.9824   -Inf      4
```

Souperb Conclusion

Given our alternative hypothesis of the median being lower than 3, we have computed a one sided sign test, at a 95% confidence interval. The S statistic tells us the number of positive differences between the data and the hypothesized median, in this case 5. The p-value of 0.5 is greater than the 0.05 p-value, we therefore cannot reject the null hypothesis that the median is less than 3.

Operating System Rates: Wilcoxon Rank Sum Test

To best test the difference in two population means I will use the Wilcoxon Rank-Sum Test. We will obtain n + m observations from two groups and compare their medians.

Assumptions:

- Samples from each population were random samples
- Samples are independent
- Both populations are continuous

OS Import Data

```
os_rates <- read.xlsx("/Users/SeanOMalley1/Desktop/MSDS_660_Stats/nonparametric_class_da
ta.xlsx",
                      sheetIndex = 2)
```

OS EDA

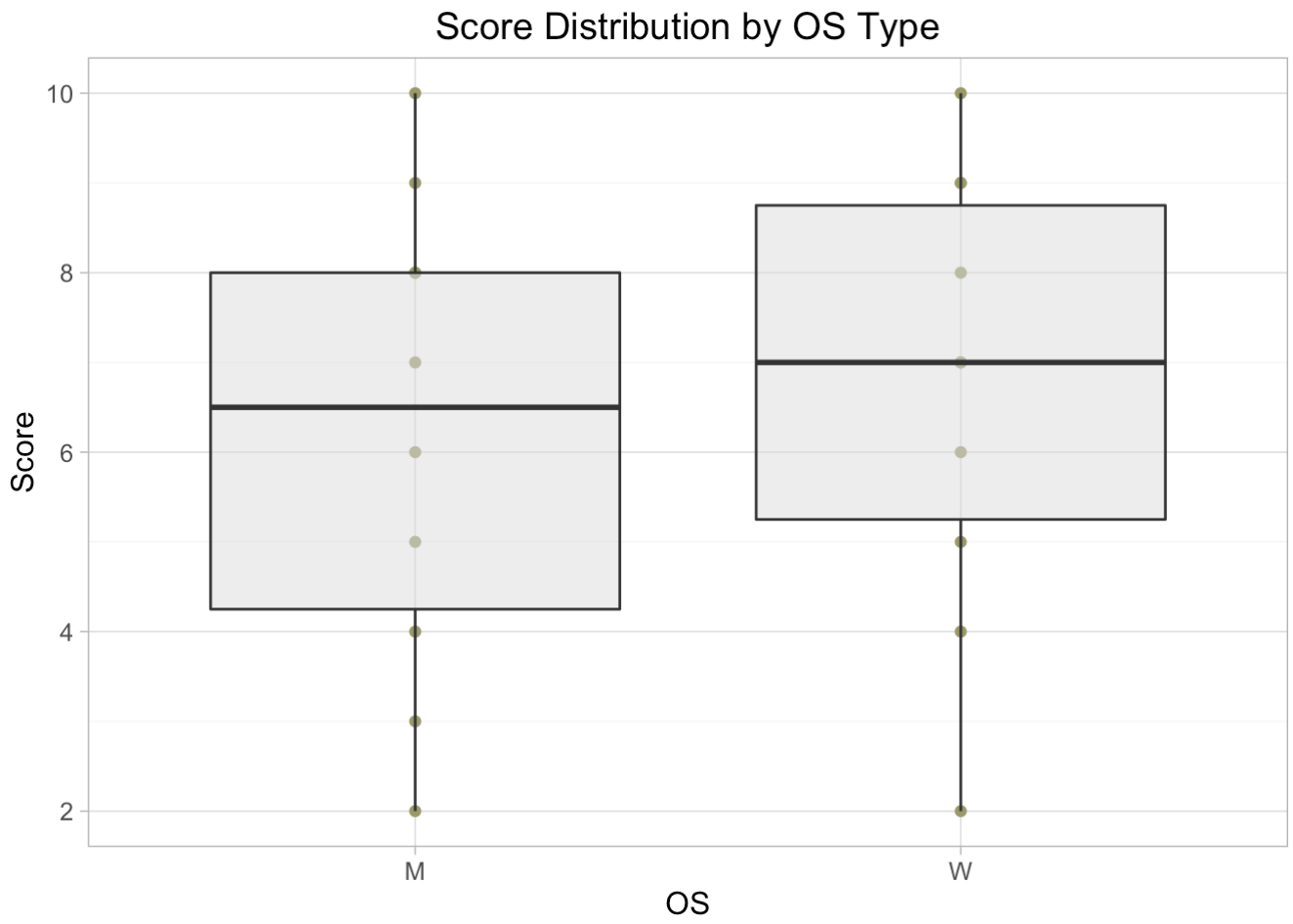
```
glimpse(os_rates)
```

```
## Observations: 20
## Variables: 2
## $ OS    <fctr> M, M, M, M, M, M, M, M, M, M, W, W, W, W, W, W, W, W, W, W
## $ Score <dbl> 9, 8, 5, 3, 6, 10, 4, 2, 8, 7, 7, 6, 8, 2, 9, 5, 4, 7, 1...
```

```
summary(os_rates$Score)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.00    4.75    7.00    6.45    8.25   10.00
```

```
ggplot(data = os_rates, aes(y = Score, x = OS)) +
      geom_point(color = "#999966") +
      geom_boxplot(outlier.colour = "#8b0000", fill = "#DCDCDC", notch = "gray", alpha =
 0.5) +
      theme_light() +
      ggtitle("Score Distribution by OS Type")
```



OS Hypothesis

H^0 : The populations of M and W Operating systems are the same

H^1 : The operating system W has a higher median than the M operating system

OS Hypothesis Testing

```
os_rates1 <- os_rates %>% group_by(OS)

M <- as.numeric(os_rates$Score[1:10])

W <- as.numeric(os_rates$Score[11:20])

os_rates2 <- as.data.frame(cbind(M,W), row.names = c("M","W"))

wilcox.test(data = os_rates, x =  M, y = W, paired = FALSE, alternative = "greater")
```

```
## Warning in wilcox.test.default(data = os_rates, x = M, y = W, paired =
## FALSE, : cannot compute exact p-value with ties
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  M and W
## W = 44.5, p-value = 0.676
## alternative hypothesis: true location shift is greater than 0
```

OS Conclusion

Given our alternative hypothesis of the median value of W being higher than M, we have computed a one sided sign test, with the desire of a 0.05 p value. The p-value of 0.676 is greater than the 0.05 p-value, we therefore cannot reject the null hypothesis that the median W is more than M. This number is additionally affirmed by a high value of the W (or U) statisti of 44.5.