MSDS 660 Week 7 Homework Assignment

Jeremy Beard

2022-08-17

Introduction In this week's assignment, we will explore the concept of nonparametric statistics. Nonparametric statistics come into play when traditional assumptions about a dataset are not able to be made e.g. the data is not assumed to be normal, or the sample size is small. These nonparametric statistic tests can be performed regardless of the distribution of the data. However, given that these tests make less assumptions about the data and can be applied more widely, the conclusions they're able to draw are less powerful as a result.

In this assignment, we will be using a dataset that compares the effect of a placebo with the effect of a new drug. The sample size is only five (5) for each variable so they are small datasets. Because they are small datasets, not many assumptions can be made about their distributions so nonparametric statistics will come into play.

```
# First, we will load the libraries and the dataset
library(data.table)
library(BSDA)
```

```
## Loading required package: lattice
```

```
##
## Attaching package: 'BSDA'
```

```
## The following object is masked from 'package:datasets':
##
##     Orange
```

```
library(agricolae)

# Is there a significant difference in the medians of
# the placebo dataset vs. the medians of the new drug dataset?

# Load the data:
dt <- read.csv("C:\\Users\\jerem\\OneDrive\\Documents\\School\\_REGIS\\2022-05_Summer\\MSDS660\\Week7\\
dt <- as.data.table(dt)
```

Methods The statistical methods and tests we will be performing in this week's assignment are a Shapiro-Wilks test to try to glean information about the distribution of the data. This test will tell us if we can safely reject the null hypothesis or not and will see if the data is non-normal or if it may be normal. True normality cannot be concluded from this test however.

We will also perform a Wilcoxon Signed test to compare the two (2) datasets as a whole. This will give us an indicator of if the two (2) populations are significantly different or not.

We will also perform a Sign test which will give us another indicator for comparing the two (2) datasets. The Sign test will tell us if the medians of the two (2) datasets are significantly different or not. This is similar to a t-test but a t-test compares population means while the Sign test compares population medians.

First, below, we will examine the data and create some preliminary charts to observe the shape of the data.

```
# View the structure of the data
dt
```

```
##    Placebo New.Drug
## 1:       7        3
## 2:       5        6
## 3:       6        4
## 4:       4        2
## 5:      12        1
```

```
summary(dt)
```

```
##     Placebo        New.Drug
##  Min.   : 4.0   Min.   :1.0
##  1st Qu.: 5.0   1st Qu.:2.0
##  Median : 6.0   Median :3.0
##  Mean   : 6.8   Mean   :3.2
##  3rd Qu.: 7.0   3rd Qu.:4.0
##  Max.   :12.0   Max.   :6.0
```

```
str(dt)
```

```
## Classes 'data.table' and 'data.frame':   5 obs. of  2 variables:
##  $ Placebo : int  7 5 6 4 12
##  $ New.Drug: int  3 6 4 2 1
##  - attr(*, ".internal.selfref")=<externalptr>
```

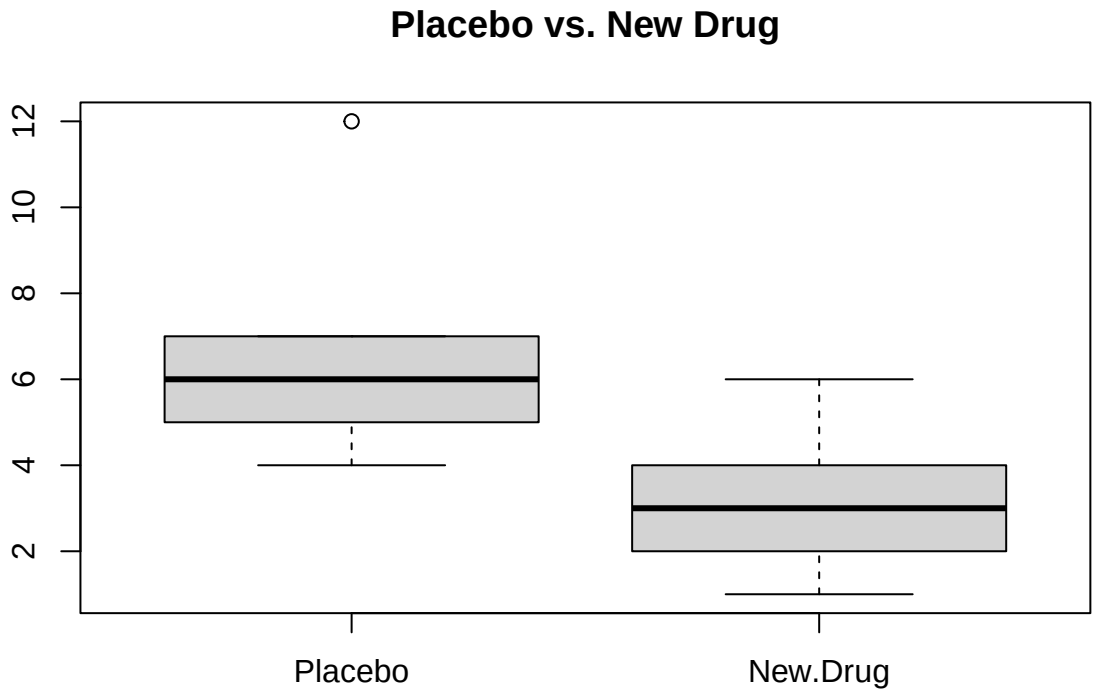
```
nrow(dt)
```

```
## [1] 5
```

```
ncol(dt)
```

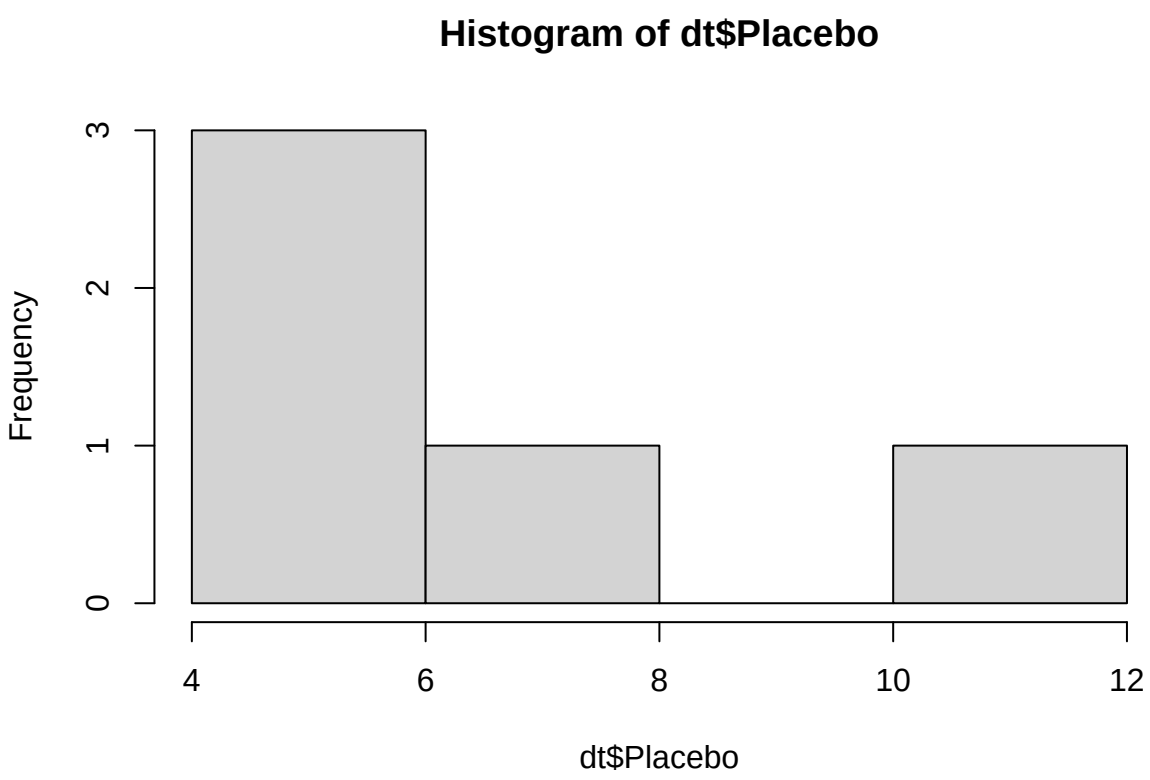
```
## [1] 2
```

```
# Create boxplots and histograms
boxplot(dt)
title("Placebo vs. New Drug")
```

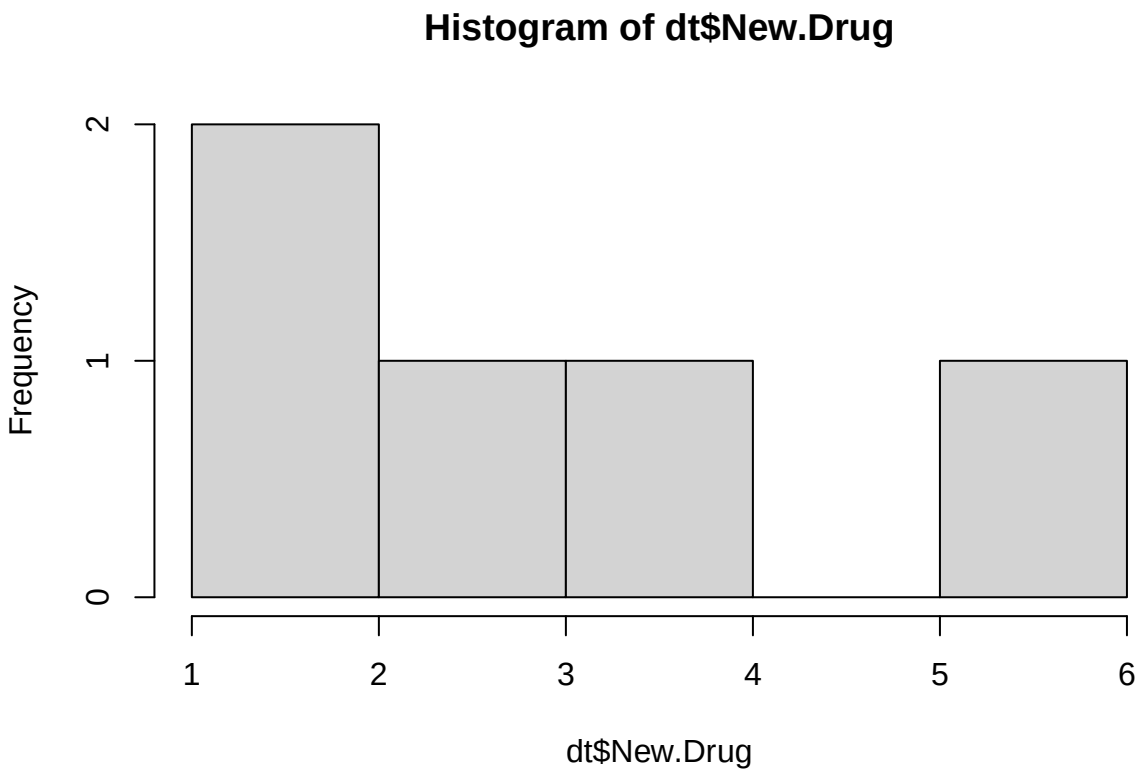



```
#Hmm should I remove the outlier 12? There are only 5 datapoints....hmm.....I will not

hist(dt$Placebo)
```




```
hist(dt$New.Drug)
```



Results The Shapiro-Wilk normality test gave a p-value of 0.25 for the Placebo variable, and 0.93 for the New Drug variable. This indicates the populations cannot be said to be not normal, so they may be normal distributions.

The median of the Placebo variable was six (6) while the median of the New Drug variable was three (3).

The Wilcoxon Sign test gave a p value JUST above 0.05, at 0.059. This indicates that the populations are not significantly different.

The Sign test gave a p value of 0.97 which indicates the medians of the populations are not significantly different. Nothing seems to be significant!

Given the few tests we performed on the data, the null hypothesis that the two populations are not significantly different cannot be rejected. All tests resulted in p-values greater than 0.05 which indicates that the results are not significantly different.

From the results, it seems like more tests need to be performed. Or really, there needs to be more data! Only having 5 datapoints to work with is really not much and I would say NO strong conclusions can be gained from this information.

```
# Use the shapiro-wilks test to see if they are from a normal distribution
shapiro.test(dt$Placebo)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  dt$Placebo
## W = 0.86696, p-value = 0.2543
```

```r
shapiro.test(dt$New.Drug)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  dt$New.Drug
## W = 0.97872, p-value = 0.9276
```

```r
# we have a small sample size so determining the distribution of the data is important
# using the shapiro-wilks test, the p values of each variable were above 0.05 which
# does not indicate non-normality. The data may be called normal

# Are the data paired or unpaired?
# The data is paired because each placebo datapoint can be paired with the new.drug datapoint
# this will make comparison easier

# i will perform a Wilcoxon Signed Test to see if the two populations have a significant different
median(dt$Placebo)
```

```
## [1] 6
```

```r
median(dt$New.Drug)
```

```
## [1] 3
```

```r
#the medians certainly seem significantly different! But we will see

wilcox.test(dt$Placebo, dt$New.Drug)
```

```
## Warning in wilcox.test.default(dt$Placebo, dt$New.Drug): cannot compute exact p-
## value with ties
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  dt$Placebo and dt$New.Drug
## W = 22, p-value = 0.05855
## alternative hypothesis: true location shift is not equal to 0
```

```r
# the wilcoxon test produced a p value greater than 0.05, so no conclusions can be made from this test

# Let's perform a sign test now
SIGN.test(x = dt$Placebo, y = dt$New.Drug, alternative = 'less')
```

```
##
##  Dependent-samples Sign-Test
##
## data:  dt$Placebo and dt$New.Drug
## S = 4, p-value = 0.9687
## alternative hypothesis: true median difference is less than 0
## 95 percent confidence interval:
##   -Inf 10.16
## sample estimates:
## median of x-y
##             2
##
## Achieved and Interpolated Confidence Intervals:
```

```
##
##                   Conf.Level L.E.pt U.E.pt
## Lower Achieved CI     0.8125   -Inf   4.00
## Interpolated CI       0.9500   -Inf  10.16
## Upper Achieved CI     0.9688   -Inf  11.00
```

```
# This p value is huge, so we cannot say that the two median's are any different really.
# We fail to reject the null hypothesis in this case
```

Conclusion From all the information above, it can be said that the new drug does not have a significant effect on the population, compared to the placebo. The p-values on all tests performed were above 0.05 and no comparison yielded a significant effect.

Future steps with this data may be more tests to see if we can learn anything else about the distributions of the data, or learn anything useful when comparing the data to each other.

Another large critique I have of the data used in this week's assignment, also stated above, is that the sample size is only 5 for each of the variables! This is an extremely insignificant sample size and I would say regardless of what the statistical tests conclude, everything should be taken with a grain of salt as five (5) datapoints is not much to analyze. Usually drug trials are much more far-reaching and populations are much higher.

Thank you!

Jeremy

References

1) Quantifying Health. (n.d.). How to Report the Shapiro-Wilk Test. Retrieved August 17, 2022, from https://quantifyinghealth.com/report-shapiro-wilk-test/

2) GraphPad Software, LLC. (n.d.). Interpreting results: Wilcoxon signed rank test. GraphPad Prism 9 Statistics Guide. Retrieved August 17, 2022, from https://www.graphpad.com/guides/prism/latest/statistics/stat_interpreting_results_wilcoxon_.htm