Non-Parametric Tests

Lynn Waterhouse

August 2, 2017

Non-Parametric Statistical Tests

When you are unable to make the assumption about the distribution of the data one can use non-parametric statistical tests.

These are the most common parametric tests and their non-parametric tests:

One-sample:

1. One sample T-test –> Wilcoxon signed rank sum test

Two-sample, paired:

1. Paired T-test –> Wilcoxon matched pairs signed rank sum test

Two-sample, unpaired:

1. Two sample T-test –> Mann Whitney U-test

Three or more samples:

1. One-way ANOVA –> Krusall-Wallis Test

Note: When you conduct non-parametric tests you no longer will get a confidence interval around your estimate, because that assumes a distribution.

Also, in statistics, thanks to the central limit theorem when you have a sample size that is 30 or more you can assume a normal distribution.

One Sample Sign Test (Binomial Test)

If there are no differences, on average, between the sample values and the hypothesized specific value, we would expect and equal number of observations above and below the specific value. To test this, we can use the Binomial distribution (or the Normal approximation to it) to evaluate the probability of the observed frequencies above and below when the true probability of being above the specific value is p=1/2.

To conduct this test in R, let's get some data to explore. We need to load the datasets package.

```
library(datasets)
```

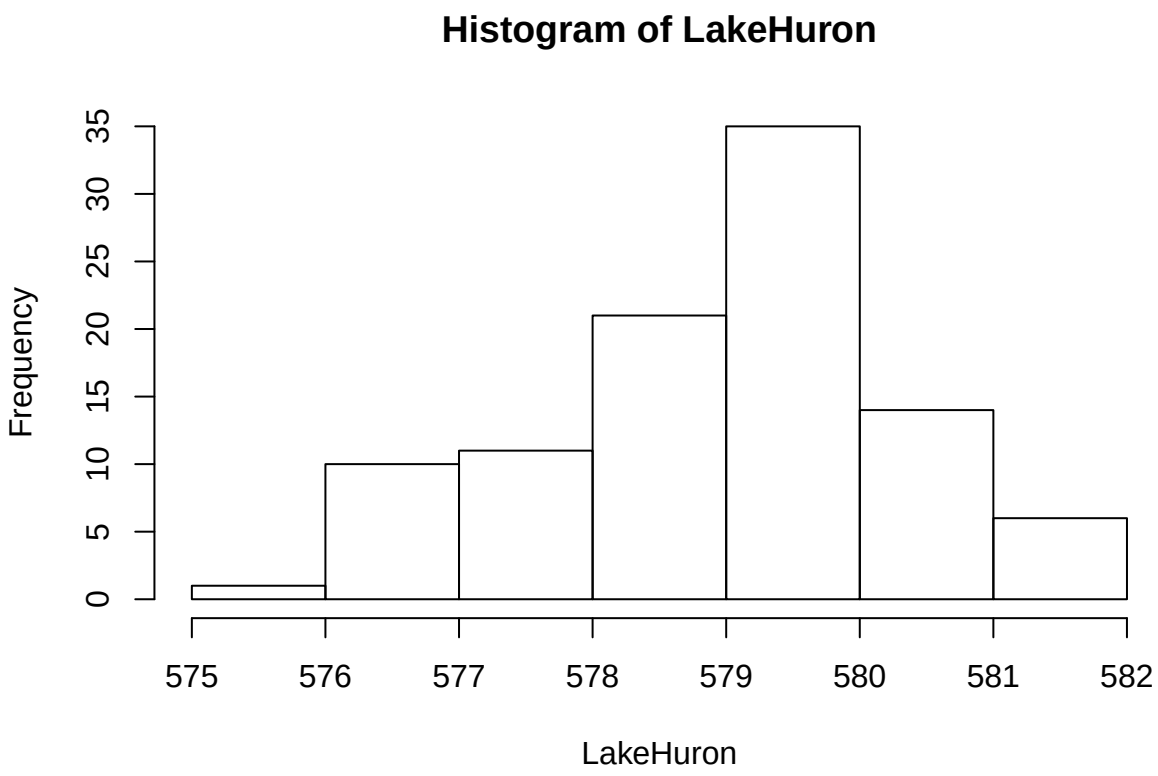
Example We have heights of the water levels in Lake Huron from 1875 to 1972. We are interested in if the average water levels is 578.5. We are unwilling to make an assumption of normality.

Research question: Lake Huron typically has more water than 578.5.

Measurement: Water level in Lake Huron.

First, let's look at a histogram of the data.

```
hist(LakeHuron)
```



Now, let's find out how many values we have, and how many are greater than 578.5.

```
length(LakeHuron)
```

```
## [1] 98
```

```
length(which(LakeHuron>578.5))
```

```
## [1] 65
```

Of the 98 values, 65 are more than 578.5. We can now compute the binomial probability of getting this.

```
pbinom(q=65,size=98,prob=1/2) #this is P(X<=65)
```

```
## [1] 0.9996151
```

This p-value is 0.9996151, but this is the probability of getting less than or equal to 65 out of 98. Instead we want 65 or more.

```
1-pbinom(q=64,size=98,prob=1/2)
```

```
## [1] 0.0008002901
```

Our new p-value is 0.0008002901 and we reject the null hypothesis, in favor of the alternative that the median is above 578.5.

If we wanted to look at the parametric equivalent of this:

```
t.test(LakeHuron, alternative="greater", mu=578.5)
```

```
##
##  One Sample t-test
```

```
##
## data:  LakeHuron
## t = 3.7853, df = 97, p-value = 0.000133
## alternative hypothesis: true mean is greater than 578.5
## 95 percent confidence interval:
##  578.7829      Inf
## sample estimates:
## mean of x
##  579.0041
```

The p-value was 0.000133, which we means we would reject the null and conclude the average is greater than 578.5.

Wilcox Signed Rank Test

There is a build in function in R to perform the Wilcox signed rank test.

This test has two assumptions: (1) the random variable X is continuous (2) the probablility density function of X is symmetric

The Lake Huron data certainly meets assumption (1) but probably not (2).

How this test works is it takes all the values in the dataset and calculates the absolute value of each minus the median (578.5), and ranks then smallest to largest (these are the ranks- the Ri values). Then it takes the ranks and makes them plus or minus (based on if the value minus the median was positive or negative)- and then recodes them as 1 or 0(if negative)– these are the "Zi" values. The statistic is then the sum of the products of the RiZi's. This is a lot more detail than you need to know.

Here is the code to perform it in R.

```
wilcox.test(x=LakeHuron,alternative="greater",mu=578.5)
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  LakeHuron
## V = 3409.5, p-value = 0.0002459
## alternative hypothesis: true location is greater than 578.5
```

From this test, we have a p-value of 0.0002459, so we reject the null in favor of the alternative (that the median is greater than 578.5).

Wilcoxon signed rank test: 0.0002459 Binomial test:0.0008002901 T-test: 0.000133

We can see that the Binomial is the most conservative test (has the fewest assumptions), followed by the Wilcoxon signed rank test (assumes symmetry), and then the T-test (assumes normality).

Two-sample paired: Wilcoxon matched pairs signed rank sum test

Example In the built-in data set named immer, the barley yield in years 1931 and 1932 of the same field are recorded. The yield data are presented in the data frame columns Y1 and Y2.

```
library(MASS)          # load the MASS package
```

```
## Warning: package 'MASS' was built under R version 3.3.3
```

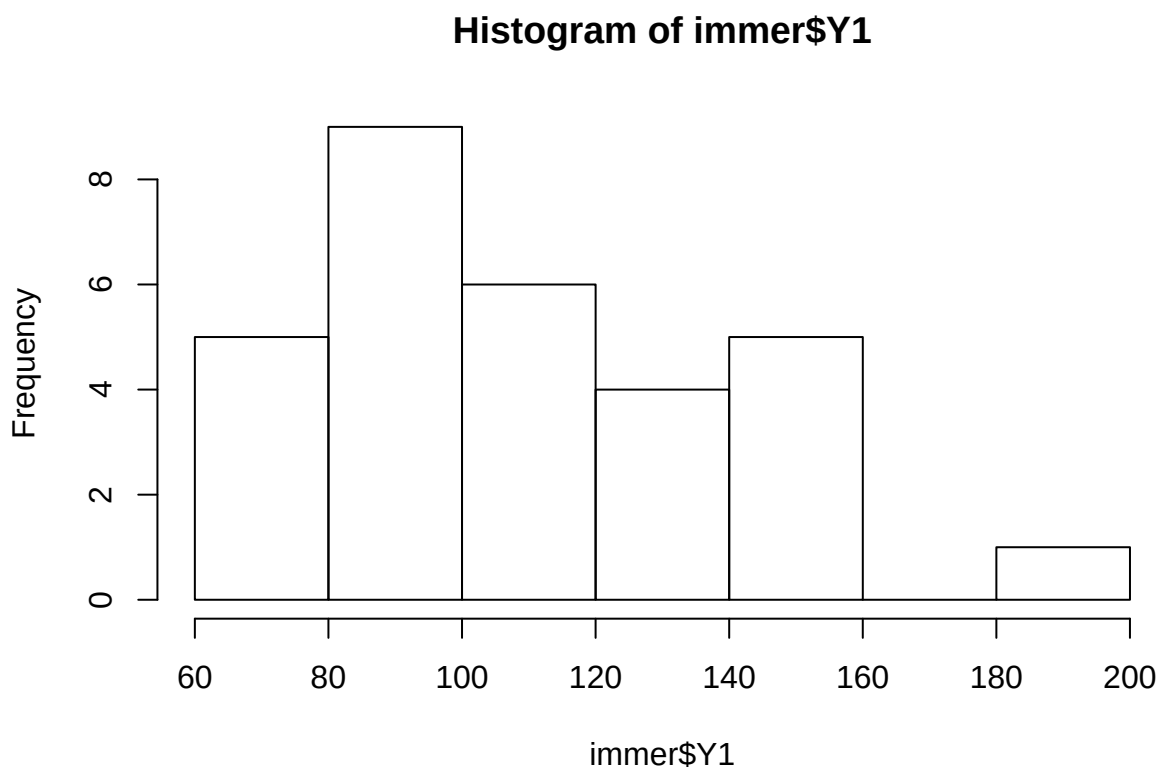
```
head(immer)
```

```
##   Loc Var    Y1    Y2
## 1  UF   M  81.0  80.7
## 2  UF   S 105.4  82.3
## 3  UF   V 119.7  80.4
## 4  UF   T 109.7  87.2
## 5  UF   P  98.3  84.2
## 6   W   M 146.6 100.4
```

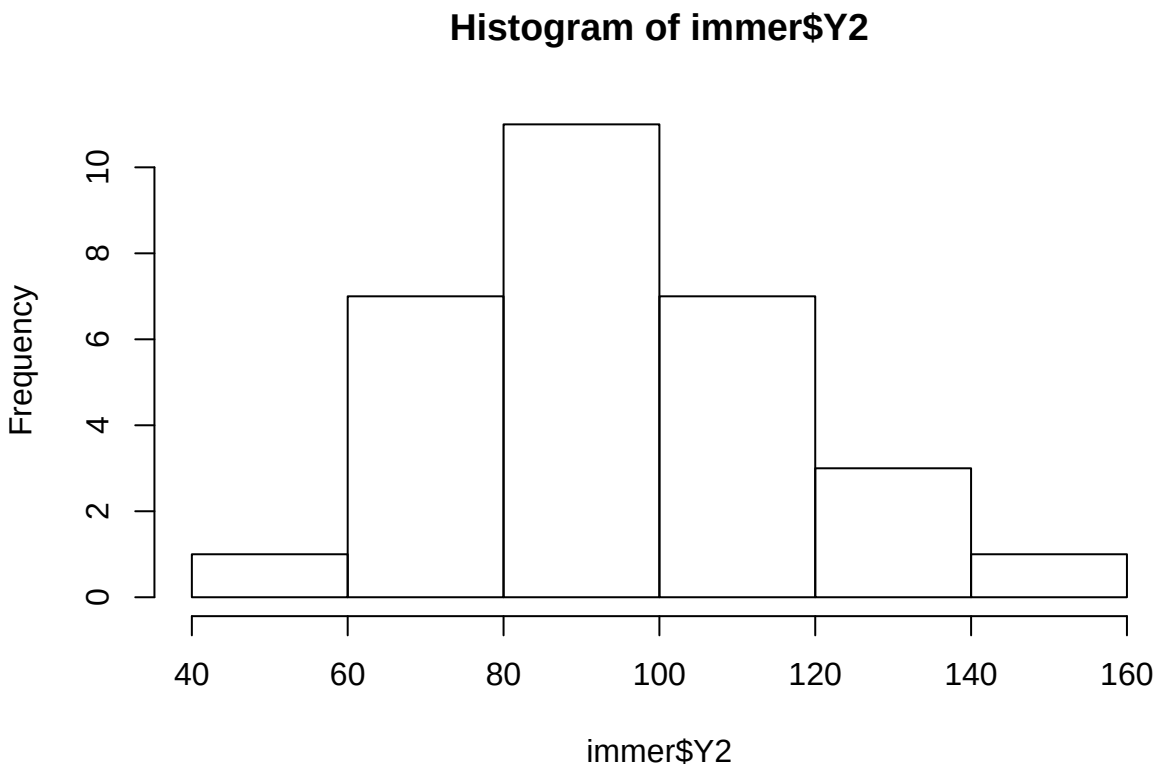
Problem Without assuming the data to have normal distribution, test at .05 significance level if the barley yields of 1931 and 1932 in data set immer have identical data distributions.

Let's first look at the data. What do you notice?

```
hist(immer$Y1)
```




```
hist(immer$Y2)
```



We use the same code as in the one-sample test.

```
wilcox.test(immer$Y1, immer$Y2, paired=TRUE)
```

```
## Warning in wilcox.test.default(immer$Y1, immer$Y2, paired = TRUE): cannot
## compute exact p-value with ties
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  immer$Y1 and immer$Y2
## V = 368.5, p-value = 0.005318
## alternative hypothesis: true location shift is not equal to 0
```

At .05 significance level, we conclude that the barley yields of 1931 and 1932 from the data set immer are nonidentical populations.

Two-sample unpaired: Mann Whitney U-Test

Example You want to see if the mean of goals suffered by two football teams over the years is the same. Are below the number of goals suffered by each team in 6 games for each year.

Team A: 6, 8, 2, 4, 4, 5 Team B: 7, 10, 4, 3, 5, 6

```
a = c(6, 8, 2, 4, 4, 5)
b = c(7, 10, 4, 3, 5, 6)

wilcox.test(a,b, correct=FALSE)
```

```
## Warning in wilcox.test.default(a, b, correct = FALSE): cannot compute exact
## p-value with ties
```

```
##
##  Wilcoxon rank sum test
##
## data:  a and b
## W = 14, p-value = 0.5174
## alternative hypothesis: true location shift is not equal to 0
```

The p-value is greater than 0.05, then we can accept the hypothesis H0 of statistical equality of the means of two groups.

Kruskall-Wallis

A collection of data samples are independent if they come from unrelated populations and the samples do not affect each other. Using the Kruskal-Wallis Test, we can decide whether the population distributions are identical without assuming them to follow the normal distribution.

Example In the built-in data set named airquality, the daily air quality measurements in New York, May to September 1973, are recorded. The ozone density are presented in the data frame column Ozone.

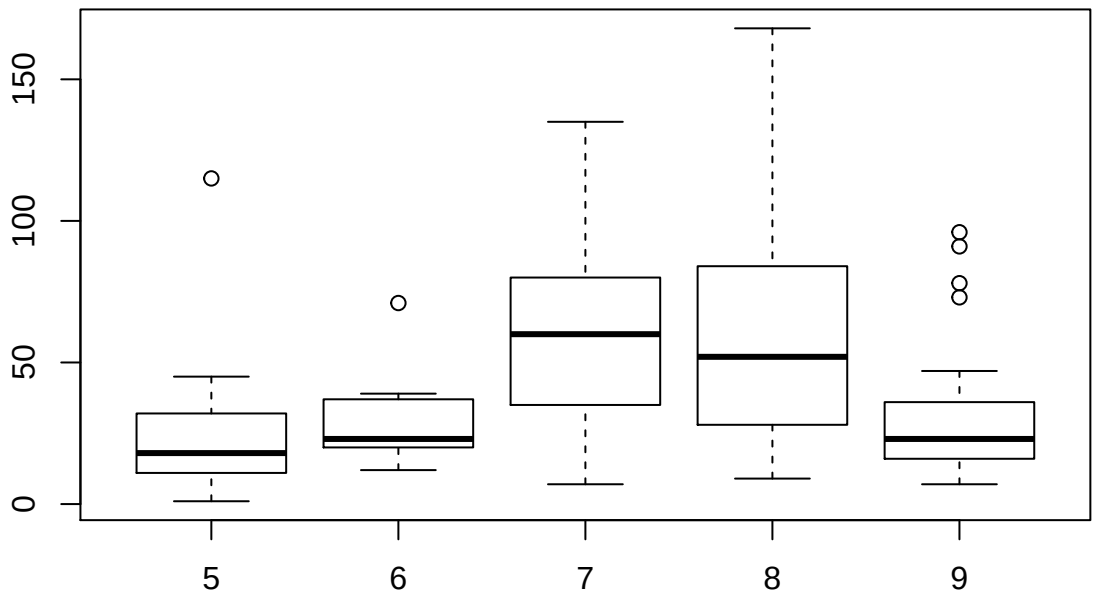
```
head(airquality)
```

```
##   Ozone Solar.R Wind Temp Month Day
## 1    41     190  7.4   67     5   1
## 2    36     118  8.0   72     5   2
## 3    12     149 12.6   74     5   3
## 4    18     313 11.5   62     5   4
## 5    NA      NA 14.3   56     5   5
## 6    28      NA 14.9   66     5   6
```

Problem Without assuming the data to have normal distribution, test at .05 significance level if the monthly ozone density in New York has identical data distributions from May to September 1973.

Let's look at some preliminary boxplots.

```
boxplot(formula=airquality$Ozone~airquality$Month)
```



Now let's conduct the Kruskall-Wallis test.

```
kruskal.test(Ozone ~ Month, data = airquality)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  Ozone by Month
## Kruskal-Wallis chi-squared = 29.267, df = 4, p-value = 6.901e-06
```

Solution The null hypothesis is that the monthly ozone density are identical populations. To test the hypothesis, we apply the kruskal.test function to compare the independent monthly data. The p-value turns out to be nearly zero (6.901e-06). Hence we reject the null hypothesis.

At .05 significance level, we conclude that the monthly ozone density in New York from May to September 1973 are nonidentical populations.

References

1. General non parametrics http://www.math.ntua.gr/~fouskakis/1-2.Non-%20Parametrics.pdf
2. Wilcox rank sum test https://onlinecourses.science.psu.edu/stat414/node/319
3. Paired wilcox rank sum test http://www.r-tutor.com/elementary-statistics/non-parametric-methods/wilcoxon-signed-rank-test
4. Mann Whitney U-Test https://www.r-bloggers.com/wilcoxon-mann-whitney-rank-sum-test-or-test-u/