Felix Torres
STAT 672: Nonparametric Statistics
Fall 2019

EXAM 1 TAKE-HOME

October 23, 2019

STAT 672 Exam 1 Take-Home
October 23, 2019

Contents

1 Problem 1

Let X have a Cauchy distribution and let Y have a standard normal distribution. Prof. Bailey has simulated a random sample from each of these distributions to use as two datasets, using the `rcauchy()` and `rnorm()` R functions. Using `read.table()`, we import the desired dataset.

```
library(tidyverse)
library(exactRankTests)
datpath <- "https://edoras.sdsu.edu/~babailey/stat672/testdat.txt"
testdat <- as_tibble(read.table(datpath, header = T))
```

1.1 Part A

We first consider the random sample from the Cauchy distribution (X's). We define a **Cauchy distribution** $f(x; x_0, \gamma)$ in statistical terms as the distribution of the ratio of two independent normally distributed random variables with mean zero. Using a Wilcoxon Sign Rank statistic, we test whether the population median is equal to zero against a two-sided alternative. These hypotheses can be represented as:

$$\mathrm{H}_0 : \theta = 0$$
$$\mathrm{H}_1 : \theta \neq 0$$

Where θ is the median of population X, which can be estimated simply by the median of the Walsh averages– that is, the median of the pairwise differences of observations between paired samples given by:

$$\hat{\theta} = \text{median } \{\frac{Z_i + Z_j}{2}, i \leq j = 1, ..., n\} \tag{1}$$

A $(1 - \alpha)\%$ confidence interval can then be estimated around $\hat{\theta}$, using:

$$C_\alpha = \frac{n(n+1)}{2} + 1 - t_{\alpha/2} \tag{2}$$

$$\theta_L = W^{C_\alpha}, \theta_U = W^{t_{\alpha/2}} \tag{3}$$

Where $t_{\alpha/2}$ is the upper $(\alpha/2)$th percentile point of the null distribution of T^+ and W is the ordered set of the Walsh averages as defined in (1). Below, we test whether the X population median is equal to zero against a two-sided alternative at the $\alpha = 0.10$ significance level.

```
xmedwtest <- wilcox.exact(testdat$x,
             mu = 0,
             alternative = "two.sided")
print(xmedwtest)

##
##  Asymptotic Wilcoxon signed rank test
##
## data:  testdat$x
## V = 2926, p-value = 0.168
## alternative hypothesis: true mu is not equal to 0
```

Even at the $\alpha = 0.10$ level, we fail to reject the null hypothesis and conclude that the median of the X sample is not significantly different than 0.

To support these findings, we calculate a $(1 - \alpha) * 100 = 90\%$ confidence interval using Walsh averages below.

```
conf.level <- 0.90
z <- sort(testdat$x)
n <- length(z)
walsh <- outer(z, z, "+") / 2
walsh <- sort(walsh[lower.tri(walsh, diag = TRUE)])
m <- length(walsh)
median(walsh)

## [1] 0.3108975

alpha <- 1 - conf.level
k <- qsignrank(alpha / 2, n)
if (k == 0) k <- k + 1
cat("achieved confidence level:",
    1 - 2 * psignrank(k - 1, n), "\n")

## achieved confidence level: 0.9006407

c(walsh[k], walsh[m + 1 - k])

## [1] -0.05541285  0.72767725
```

We find that an estimator of the median X is $\hat{\theta} = 0.311$, which has an approximate 90% confidence interval of $[-0.0554, 0.728]$. We also see that the exact achieved confidence level of this interval is 0.9006407. Since 0 is included in the interval, the confidence interval supports the results of the Wilcoxon Sign Rank test. To ensure that the Wilcoxon Signed Rank procedure is appropriate, we check its assumptions below.

1. **Assumption 1: Z_is are mutually independent**: Because `rcauchy()` uses random generation to generate a Cauchy vector and therefore, this assumption is satisfied.
2. **Assumption 2: Each Z_i comes from a continuous population that is symmetric about a common median θ**: This assumption is satisfied. `rcauchy()` sets `location = 0` and `scale = 1` as the default, and therefore every Z indeed comes from a continuous population that is symmetric about a common median.

1.2 Part B

We have two random samples from different distributions(X - Cauchy distribution, Y - Normal distribution). To test whetherthere is a true location difference between samples, we use a Wilcoxon Rank Sum test. The hypotheses are stated below.

$$H_0 : \Delta = 0$$
$$H_1 : \Delta \neq 0$$

A Wilcoxon Rank Sum statistic W is used to test these hypotheses, defined by:

$$W = \sum_{j=1}^{n} S_j,$$

Where S_j is the rank of Y_j and therefore W the sum of ranks assigned to the Y values.

```r
x <- testdat$x
y <- testdat$y
wilcox.exact(x, y,
             paired = F,
             alternative = "two.sided",
             conf.int = T,
             conf.level = 0.90)
```

```
##
##  Asymptotic Wilcoxon rank sum test
##
## data:  x and y
## W = 5511, p-value = 0.2118
## alternative hypothesis: true mu is not equal to 0
## 90 percent confidence interval:
##  -0.08042713  0.67177739
## sample estimates:
## difference in location
##              0.2868192
```

The p-value for this test is 0.2118 and we therefore fail to reject the null hypothesis, concluding that the medians are not significantly different between the two samples. To see whether or not the Wilcoxon Rank Sum procedure is appropriate, we check its assumptions below.

1. **Assumption 1: Observations X_i come from a population of X's that are independent and identically distributed. Observations Y_j also come from a population of Y's that are i.i.d.**: This assumption is satisfied; `rnorm(100)` and `rcauchy(100)` were used to generate these samples.

2. **Assumption 2: X's and Y's are mutually independent**: This assumption is satisifed because it is reasonable to assume that the outcome of one randomly generated X value has no influence on another randomly generated Y value.

3. **Assumption 3: Populations 1 and 2 are continuous.** This assumption is also satisfied. `rnorm(100)` and `rcauchy(100)` both generate random variables from continuous distributions.

Since all three assumptions are satisifed, the Wilcoxon Rank Sum procedure was appropriate to test 2-sample location equality.

2 Problem 2

In this problem we consider linear rank statistics as a general class of distribution-free statistics. A general definition for L and definitions the expected value (mean) and variance of a special case are given in Part A. In Part B, we test whether a given sample comes from a random sample with a parabolic trend.

2.1 Part A

It can be shown that a linear rank statistic L and its mean μ_L are generally

$$L = \sum_{i=1}^{N} a_i c(R_i)$$

$$\begin{aligned}\mu_L &= E[\sum_{i=1}^{N} a_i c(R_i)] = \sum_{i=1}^{N} a_i E[c(R_i)] \\ &= \sum_{i=1}^{N} a_i \bar{c} = N\bar{a}\bar{c}\end{aligned}$$

It is given that a statistic that could be used to test the alternative parabolic trend hypothesis is

$$L= \textstyle\sum_{i=1}^{6} a_i R_i.$$

Therefore, it can be seen that for this example, $c(R_i) = R_i$. We are also given that $a_1 = a_9 = 9, a_2 = a_5 = 4, a_3 = a_4 = 1$, such that $\vec{a} = \{9, 4, 1, 1, 4, 9\}$. We find the expected value (mean) of the linear rank statistic below.

$$\begin{aligned}N &= 6 \\ \bar{a} &= \frac{9+4+1+1+4+9}{6} = 4.\bar{6} \\ \bar{c} &= \frac{\sum_{i=1}^{N} c(R_i)}{N} = \frac{\sum_{i=1}^{N} R_i}{N} \\ &= \frac{N(N+1)}{2N} = \frac{6*7}{12} = 3.5\end{aligned}$$

$$\mu_L = N\bar{a}\bar{c} = 6(4.667)(3.5) \approx 98.$$

To find the variance of the linear rank statistic σ_L^2, we apply the observation that

$$\mathrm{Var}_0(L) = \sigma_L^2 = \frac{1}{N-1}(\sum_{i=1}^{N}(a_i - \bar{a})^2(\sum_{k=1}^{N}(c_k - \bar{c})^2)$$

Below, we verify μ_L, $\bar{a}$, and $\bar{c}$, and calculate σ_L^2 programmatically.

```r
N <- 6
ai <- c(9,4,1,1,4,9)
ci <- c(1,2,3,4,5,6)
abar <- sum(ai)/length(ai)
print(cbar <- sum(ci)/length(ci))

## [1] 3.5

print(cbar2 <- (N * (N + 1)) / (2 * N))

## [1] 3.5

afac <- (ai - abar)^2
cfac <- (ci - cbar)^2
muL <- N * abar * cbar
varL <- (1/(N - 1)) * sum(afac) * sum(cfac)
muL

## [1] 98

varL

## [1] 228.6667
```

Therefore, $\mu_L = 98$ and $\sigma_L^2 = 228.667$.

2.2 Part B

We start with the set of observations $X = \{1.8, 1.0, 1.5, 1.4, 1.9, 1.6\}$. The linear rank statistic is found below.

```r
x <- c(1.8, 1, 1.5, 1.4, 1.9, 1.6)
print(xranks <- rank(x))

## [1] 5 1 3 2 6 4
```

```
ai <- c(9,4,1,1,4,9)
Li <- c()
for (i in 1:length(ai)) {
  Li[i] <- ai[i] * xranks[i]
}
print(Lstat <- sum(Li))

## [1] 114

pnorm(q = Lstat, mean = muL, sd = sqrt(varL), lower.tail = F)

## [1] 0.1450095
```

Using our previously calculated μ_L and σ_L^2, we find that $p = 0.145$ for this set of observations. Therefore, we fail to reject the null hypothesis at the $\alpha = 0.05$ significance level.