# BMI713 Problem Set 3

**Instructions:**

Please submit this problem set before class on Tuesday, November 14. Problem sets may be submitted within a week past the due date at a 20% penalty; each person is allowed to submit one problem late (within a week) without penalty. Please comment your code indicating what your functions do and any relevant passage (not necessarily every line of code), because it is part of the requirements of each exercise. Missing comments will not allow the full score.

If you have any questions, please post on the piazza site. This problem set was prepared by Tiziana Sanavia and Giorgio Melloni, so they will be most prepared to answer questions.

## 1. Non-Parametric Testing Part 1 (35 points)

A pharmaceutical company is testing a new soporific drug that is supposed to be more effective than the state-of-the-art medication. 10 subjects are recruited and the hours of extra sleep are reported. The null hypothesis $H_0$ is that there is no difference in extra hours of sleep between the two drugs.

| Subject | New_Drug | Old_Drug |
|---|---|---|
| 1 | 0.7 | 1.9 |
| 2 | 0.8 | -1.6 |
| 3 | 1.1 | -0.2 |
| 4 | 0.1 | -1.3 |
| 5 | -0.2 | -0.1 |
| 6 | 4.4 | 3.4 |
| 7 | 5.5 | 3.7 |
| 8 | 1.6 | 0.8 |
| 9 | 4.6 | 0.0 |
| 10 | 3.4 | 2.0 |

### (a) Calculate the Wilcoxon signed-rank T statistic (5 points)

```
# Calculate the difference in hours slept
diff_vec <- New_Drug - Old_Drug
# Calculate the rank
rankDiff <- rank(abs(diff_vec))
# Sum the ranks where the difference is positive
Tstat <- sum( rankDiff[ diff_vec > 0 ] )
Tstat
```

```
## [1] 50
```

### (b) Calculate $\mu_T$ and $\sigma_T$ under the Null hypothesis (5 points)

The expected mean under the null hypothesis is

$$\mu_T = \frac{n(n+1)}{4} = 27.5$$

While $\sigma_T$ under the null hypothesis is equal to

$$\sigma_T = \sqrt{\frac{n(n+1)(2n+1)}{24}} \approx 9.81$$

```
n <- length(New_Drug)
muT <- (n * (n+1))/4
sigmaT <- sqrt( n * (n+1) * (2*n + 1) / 24)
```

```
## [1] "Mean of T under the Null hypothesis: 27.5"
```

```
## [1] "Standard Deviation of T under the Null hypothesis: 9.81070843517429"
```

**(c) Calculate the pvalue under the normal approximation, using T, $\mu_T$ and $\sigma_T$ and comment the result obtained. (5 points)**

```
2*( 1 - pnorm( Tstat , muT , sigmaT) )
```

```
## [1] 0.02182428
```

Alternatively

```
wilcox.test( New_Drug , Old_Drug , paired = TRUE
             , exact = FALSE , correct = FALSE
             , alternative = "two.sided")
```

```
##
##  Wilcoxon signed rank test
##
## data:  New_Drug and Old_Drug
## V = 50, p-value = 0.02182
## alternative hypothesis: true location shift is not equal to 0
```

By rejecting $H_0$ at $\alpha = 0.05$ the new drug is significantly better than the standard of care.

**(d) Calculate the pvalue using the built in R function for Wilcoxon signed-rank test. Are the pvalue different? Are the conclusions different? (5 points)**

```
wilcox.test( New_Drug , Old_Drug , paired = TRUE , alternative = "two.sided")
```

```
##
##  Wilcoxon signed rank test
##
## data:  New_Drug and Old_Drug
## V = 50, p-value = 0.01953
## alternative hypothesis: true location shift is not equal to 0
```

The pvalues are similar and they both lead to conclude that $H_0$ should be rejected.

**(e) Calculate the exact p-value "by hand" and show all the steps in order to obtain it. (10 points)**

We should consider all the possible rank sign assignments (+ and -) for 10 subjects which is equal to $2^{10} = 1024$. Then we need to calculate the number of possible assignments with a T statistic equal or more extreme than

the one we found. Because it is a two sided test, we should calculate both of the T statistics (where the differences are positive and where the differences are negative). The pvalue will be the ratio between these two numbers.

$$\text{p-value} = \frac{\text{assignments where } T \text{ is } \geq T_1 \text{ or } \leq T_2}{\text{total number of assignments}}$$

```
T1 <- sum( rankDiff[ diff_vec > 0 ] )
T2 <- sum( rankDiff[ diff_vec < 0 ] )
T1
```

```
## [1] 50
```

```
T2
```

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

In fact, just one of the two Ts is necessary and we can then multiply by two:

$$\text{p-value} = \frac{\text{assignments where } T \text{ is } \geq T_1 \text{ or } \leq T_2}{\text{total number of assignments}} = \frac{2*(\text{assignments where } T \text{ is } \geq 50)}{\text{total number of assignments}}$$

The total sum of all the ranks is 55 which corresponds to a all + signs (10 +). With 9 + and 1 − the possible combinations with a value equal or greater than 50 are 5 in total (by assigning − sign to ranks 1 to 5). 8 + and 2 − is 4 (considering the following couples with a minus sign: 1-2 , 2-3 , 1-3 , 1-4). With 3 − signs, there are no combinations with a sum below or equal to 50. To Summarize:

| Sign_Composition | Number_of_Combinations |
|---|---|
| 10+ | 1 |
| 9+ and 1- | 5 |
| 8+ and 2- | 4 |
| 7+ and 3- and others | 0 |
| Total | 10 |

The p-value is then:

```
extremeAssignments <- sum(c(1 , 5 , 4))
p <- 2*(extremeAssignments)/2^10
p
```

```
## [1] 0.01953125
```

**(f) Calculate the p-value using an appropriate equivalent parametric test and comment the obtained results with respect to the 'Non-parametric' version. (5 points)**

```
t.test( New_Drug , Old_Drug , paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  New_Drug and Old_Drug
## t = 2.7804, df = 9, p-value = 0.02139
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
```

```
##  0.2497748 2.4302252
## sample estimates:
## mean of the differences
##                    1.34
```

The parametric test gives a very similar result and the conclusions are similar. The new drug is better.

## 2. Non-Parametric Testing Part 2 (30 points)

In this second part we are going to simulate a few data to check the difference between unpaired T-test and Wilcoxon rank sum test.

Imagine two vectors of length 10 from two different exponential distributions:

```
x <- rexp(10 , rate = 10)
y <- rexp(10 , rate = 40)
```

The hypothesis test is that $\mu_x$ is different than $\mu_y$ (two-sided $H_1$)

### (a) What is the most appropriate test in this case and why? (5 points)

The normality assumptions do not hold so an unpaired non-parametric test is more appropriate (Wilcoxon Rank Sum Test).

### (b) As a general rule, if the assumptions of CLT do not hold, a non parametric test is more appropriate and sometimes more powerful than its parametric counterpart. By running a simulation with 1000 random couples (x,y) like above, show that the fraction of rejected Null hypotheses at alpha = 0.01 is higher in the case of a non parametric test. NOTE alpha is 1%!! What are we showing with this simulation? (10 points)

```
myB <- 1000
tests <- t(replicate(myB , {
x <- rexp(10 , rate = 10)
y <- rexp(10 , rate = 40)
c( t.test(x , y , paired=FALSE , var.equal = TRUE)$p.value ,
   wilcox.test(x , y , paired=FALSE , exact = TRUE)$p.value)
}))
fracTest <- sum( tests[ , 1] < 0.01 )/myB
fracWilcoxon <- sum( tests[ , 2] < 0.01 )/myB
```

```
## [1] "Fraction of significant pvalues for the T Test: 0.33"
```

```
## [1] "Fraction of significant pvalues for the Wilcoxon rank sum: 0.443"
```

The fraction of rejected $H_0$ is $\sim 10\%$ higher using the Wilcoxon test. Since we know *a priori* that $H_1$ is true, this simulation demonstrate that Wilcoxon rank sum test is more powerful than T test in this specific case.

### (c) An old statistical adagio says "If the data don't behave, hit it with a log. If the data still don't behave, hit it with a log again". What happen if we log-transform the data? Run the same simulation with log(x) and log(y) and comment the results obtained? (10 points)

```
tests <- t(replicate(myB , {
x <- rexp(10 , rate = 10)
y <- rexp(10 , rate = 40)
c( t.test(log(x) , log(y) , paired=FALSE , var.equal = TRUE )$p.value ,
   wilcox.test(log(x) , log(y) , paired=FALSE , exact = TRUE )$p.value)
}))
fracTest <- sum( tests[ , 1] < 0.01 )/myB
fracWilcoxon <- sum( tests[ , 2] < 0.01 )/myB
```

```
## [1] "Fraction of significant pvalues for the T Test: 0.411"
```

```
## [1] "Fraction of significant pvalues for the Wilcoxon rank sum: 0.409"
```

The log-transformation "normalizes" the data, so that the fraction of the rejected Null hypothesis is now comparable or even higher for the T-test.

**(d) Is the log transformation useful for the wilcoxon test? if not, why? (5 points)**

The rank is invariant to monotonic transformations like the logarithm, so Wilcoxon test will give the same result for x,y or log(x),log(y).

## 3. Contingency tables (35 points)

A statistical analysis that combines the results of several studies on the same subject is called a meta-analysis. A meta-analysis compared aspirin with placebo on incidence of heart attack and of stroke, separately for men and from women (J. Am. Med. Assoc., 295: 306-313, 2006). For the Women's Health Study, heart attacks were reported for 198 of 19,934 taking aspirin and for 193 of 19,942 taking placebo. We are interested in whether aspirin was helpful in reducing the risk of heart attack.

**(a) State the null hypothesis and the alternative hypothesis. (2 points)**

Solution: Let $p_A$ be the true rate of heart attack among women who take aspirin and $p_P$ be the true rate of heart attack among women who are given placebo. Then $H_0 : p_A = p_P$ and $H_A : p_A \neq p_P$.

**(b) Construct the 2 x 2 contingency table that cross classifies the treatment (aspirin, placebo) and heart attack status (yes, no). (3 points)**

Solution: Following is the table of observed values with marginal sums:

| | Aspirin | Placebo | rowTotal |
|---|---|---|---|
| Heart attack | 198 | 193 | 391 |
| No heart attack | 19,736 | 19,749 | 39,485 |
| colTotal | 19,934 | 19,942 | 39,876 |

the matrix can be directly built in R:

```
o <- matrix(c(198,19736,193,19749),ncol=2)
o
```

```
##       [,1]  [,2]
## [1,]   198   193
```

```
## [2,] 19736 19749
```

**(c) Perform the chi-square test. Report the test statistic (5 points), the degrees of freedom (5 points) and calculate the p-value without using the R chisquare built-in function (10 points). What conclusion can you draw from this test? (5 points)**

Solution: To compute the table of expected values, since the marginals are fixed, we can calculate the expected value in each cell by:

1. multiplying the total number across the columns corresponding to the row of that cell per the total number across the rows corresponding to the column of that cell;
2. divide the resulting number in 1. by the total number of women.

Another way is to determine the proportion of heart attacks in the combined population:

$$\hat{p} = \frac{198 + 193}{198 + 193 + 19,736 + 19,749} = \frac{391}{39,876} \approx 0.0098$$

Using $\hat{p}$ and the number of patients in each of the Aspirin ($n_A$) and Placebo groups ($n_P$), we can generate the expected value for each cell in the previous table:

$$E = \begin{pmatrix} \hat{p} \\ 1-\hat{p} \end{pmatrix} \cdot (n_A, n_P) = \begin{bmatrix} \hat{p} \cdot n_A & \hat{p} \cdot n_P \\ (1-\hat{p}) \cdot n_A & (1-\hat{p}) \cdot n_P \end{bmatrix} \approx \begin{bmatrix} 195.46 & 195.54 \\ 19,738.54 & 19,746.46 \end{bmatrix}$$

The matrix product $\begin{pmatrix} \hat{p} \\ 1-\hat{p} \end{pmatrix} \cdot (n_A, n_P)$ is also called the outer product. The table of expected values above can be easily computed in R by using the *outer* function:

```
p.hat<-391/39876
n.A<-19934
n.P<-19942
e<-outer(c(p.hat,1-p.hat),c(n.A,n.P))
e
```

```
##            [,1]       [,2]
## [1,]   195.4608   195.5392
## [2,] 19738.5392 19746.4608
```

Now we can compute the $\chi^2$ statistic by summing $\frac{(O_{i,j}-E_{i,j})^2}{E_{i,j}}$, where we let the indices $i$ and $j$ run over each cell in the table. That is,

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}} \approx$$
$$\frac{(198 - 195.46)^2}{195.46} + \frac{(19,736 - 19,746.54)^2}{19,746.54} +$$
$$\frac{(193 - 195.54)^2}{195.54} + \frac{(19,749 - 19,746.46)^2}{19,746.46} = 0.0666546$$

Using the values $o$ and $e$ we computed in R, this sum can be computed in a compact R command:

```
o <- matrix(c(198,19736,193,19749),ncol=2)
p.hat<-391/39876
n.A<-19934
n.P<-19942
```

```
e<-outer(c(p.hat,1-p.hat),c(n.A,n.P))
sum((o-e)^2/e)
```

```
## [1] 0.06661375
```

The minor disagreement in value is due to the fact that the values computed by hand were rounded to two decimal places.

In a 2 x 2 table with fixed margins, there is always exactly one degree of freedom. We can now compute the p-value:

```
o <- matrix(c(198,19736,193,19749),ncol=2)
p.hat<-391/39876
n.A<-19934
n.P<-19942
e<-outer(c(p.hat,1-p.hat),c(n.A,n.P))
1 - pchisq(sum((o - e)^2/e), df=1)
```

```
## [1] 0.7963325
```

The p-value is quite high, so we cannot reject the null hypothesis $H_0$.

**(d) Perform the chi-square test using R. NOTE: look at `chisq.test()` function. (5 points)**

```
o <- matrix(c(198,19736,193,19749),ncol=2)
chisq.test(o,correct=FALSE)
```

```
##
##  Pearson's Chi-squared test
##
## data:  o
## X-squared = 0.066614, df = 1, p-value = 0.7963
```

## Extra: Fisher's Exact Test (8 points)

Consider the following example of contingency table from a study evaluating the correlation between gender and diet:

| | Diet | Non Diet | rowTotal |
|---|---|---|---|
| Men | 2 | 10 | 12 |
| Women | 8 | 12 | 20 |
| colTotal | 10 | 22 | 32 |

We want to test whether men are less prone to start a diet than women.

**(a) Display the tables that are as 'extreme' as or more extreme than the observed table (5 points)**

**(b) Calculate the probabilities with these tables to obtain the p-value of the Fisher's Exact test (3 points)**

Solution (a and b). Given the following contingency table with the corresponding marginals:

| | Diet | Non Diet | rowTotal |
|---|---|---|---|
| Men | a | b | a+b |
| Women | c | d | c+d |
| colTotal | a+c | b+d | n=a+b+c+d |

The formula to calculate the probability associated to the table is:

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!b!c!d!n!}$$

p represents the "exact" probability of observing the values (a,b,c,d), given the margins (e.g. rowTotal and colTotal).

Using this formula, we can calculate the probabilities for both the observed table and its more extreme cases, keeping the margins **fixed**. Here the aim is to evaluate if men on diet are significantly less than women. Therefore we can create the extreme cases by lowering the number of men on diet as in the following:

| | Diet | Non Diet | rowTotal |
|---|---|---|---|
| Men | a=2 | b=10 | a+b=12 |
| Women | c=8 | d=12 | c+d=20 |
| colTotal | a+c=10 | b+d=22 | n=32 |

$$p_{a=2} = \frac{12! \cdot 20! \cdot 10! \cdot 22!}{2! \cdot 10! \cdot 8! \cdot 12! \cdot 32!} \approx 0.129$$

| Diet No | n Diet | rowTotal | |
|---|---|---|---|
| Men | a=1 | b=11 | a+b=12 |
| Women | c=9 | d=11 | c+d=20 |
| colTotal | a+c=10 | b+d=22 | n=32 |

$$p_{a=1} = \frac{12! \cdot 20! \cdot 10! \cdot 22!}{1! \cdot 11! \cdot 9! \cdot 11! \cdot 32!} \approx 0.0312$$

| | Diet | Non Diet | rowTotal |
|---|---|---|---|
| Men | a=0 | b=12 | a+b=12 |
| Women | c=10 | d=10 | c+d=20 |
| colTotal | a+c=10 | b+d=22 | n=32 |

$$p_{a=0} = \frac{12! \cdot 20! \cdot 10! \cdot 22!}{0! \cdot 12! \cdot 10! \cdot 10! \cdot 32!} \approx 0.0028$$

The p-value is therefore the sum of $p_{a=2}$, $p_{a=1}$, $p_{a=0}$:

$$\text{p-value} = p_{a=2} + p_{a=1} + p_{a=0} = 0.163$$

In R, we can use the function `fisher.test()` to obtain the same results:

```
mat<-matrix(c(2,8,10,12),nrow=2)
fisher.test(mat,alternative="less")
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  mat
## p-value = 0.163
## alternative hypothesis: true odds ratio is less than 1
## 95 percent confidence interval:
##  0.000000 1.652065
## sample estimates:
## odds ratio
##  0.3109654
```

In this example, the alternative hypothesis is that the odds ratio, corresponding to $(a/b)/(c/d)$, is less than 1, since we are evaluating how much the proportion of men on diet (i.e. $a/b$) is less than the proportion of women on diet (i.e. $c/d$). Therefore in the function we have to specify `alternative="less"`.