Biostatistics Task 1

Danyu Zhang & Daniel Alonso

May 1st, 2021

```
library(coin)
library(vcd)
library(dplyr)
library(mosaic)
library(qvalue)
library(broman)
```

Exercise 2

```
groupA <- c(324,275,349,604,566,810,340,295,357,580,344,655,380,503,314)
groupB <- c(558,108,291,863,303,640,358,503,646,689,250,540,630,190)
groups_d <- factor(c(rep('A',length(groupA)), rep('B', length(groupB))))
values <- c(groupA, groupB)
groups <- data.frame(group=groups_d, ASO=values)
```

Performing a Shapiro-Wilk normality test tells us that the Group A likely is not normally distributed.

```
shapiro.test(groupA)
#>
#>  Shapiro-Wilk normality test
#>
#> data:  groupA
#> W = 0.86339, p-value = 0.02702
shapiro.test(groupB)
#>
#>  Shapiro-Wilk normality test
#>
#> data:  groupB
#> W = 0.9586, p-value = 0.7
```

Standard Procedure

We don't reject the null hypothesis, therefore we can say there's no significant difference between the population median of each group.

```
wilcox.test(ASO ~ group, data=groups)
#>
#>  Wilcoxon rank sum test with continuity correction
#>
#> data:  ASO by group
#> W = 101.5, p-value = 0.8958
#> alternative hypothesis: true location shift is not equal to 0
```

Resampling Method: Permutation Test

As with the standard procedure, we can say the population medians of both groups are not different.

```
independence_test(ASO ~ group, data=groups, distribution=approximate(nresample=10000))
#>
#>  Approximative General Independence Test
#>
#> data:  ASO by group (A, B)
#> Z = -0.32484, p-value = 0.7558
#> alternative hypothesis: two.sided
```

Exercise 3

```
treatment <- c(rep('conventional',23), rep('alternative',24))
result <- c(rep('relapse',2),rep('no relapse',21), rep('relapse',8), rep('no relapse',16))
backpain <- data.frame(treatment=treatment, result=result)
backpain_tally <- tally(~treatment+result, data=backpain)
```

Standard Method

We don't reject the null hypothesis, therefore there's no significant effect on relapse when treating using conventional methods versus alternative methods.

```
chisq.test(backpain_tally)
#>
#>  Pearson's Chi-squared test with Yates' continuity correction
#>
#> data:  backpain_tally
#> X-squared = 2.9125, df = 1, p-value = 0.0879
```

Resampling method

The resampling method confirms our results with the standard method. There's no significant differences.

```
chisq.test(backpain_tally, simulate.p.value=TRUE, B=1000)
#>
#>  Pearson's Chi-squared test with simulated p-value (based on 1000
#>  replicates)
#>
#> data:  backpain_tally
#> X-squared = 4.2563, df = NA, p-value = 0.06294
```

Exercise 4

(i)

```
cont_table = matrix(c(17,298,230,428),nrow= 2,dimnames=list("seatbelt"=c("Y","N"),"head_injury"=c("Y","
cont_table
#>          head_injury
#> seatbelt   Y   N
#>        Y  17 230
#>        N 298 428
```

(ii)

Without running any tests we can see following: The rate of having head injuries while wearing sealbelt (0.07) is much lower than while not wearing seatbelt (0.41).

```
17/247
#> [1] 0.06882591
298/726
#> [1] 0.4104683
```

(iii)

The expected counts for the contingency table are as follows:

```
chisq.test(cont_table)$expected
#>          head_injury
#> seatbelt         Y       N
#>        Y  79.96403 167.036
#>        N 235.03597 490.964
```

(iv)

Standard procedures

We can reject the null hypothesis, therefore there is a significant effect of wearing a seatbelt in order to prevent head injuries.

```
chisq.test(cont_table)
#>
#>  Pearson's Chi-squared test with Yates' continuity correction
#>
#> data:  cont_table
#> X-squared = 96.7, df = 1, p-value < 2.2e-16
```

Resampling procedures

Just as the standard procedure explains, we can conclude that wearing a seatbelt does indeed prevent head injuries.

```
chisq.test(cont_table, simulate.p.value=TRUE, B=10000)
#>
#>  Pearson's Chi-squared test with simulated p-value (based on 10000
#>  replicates)
#>
#> data:  cont_table
#> X-squared = 98.255, df = NA, p-value = 9.999e-05
```

Exercise 5

```
cont_table = matrix(c(6,17,37,2,13,44),nrow=2,ncol=3,byrow=T,dimnames=list("type"=c("active_drug","Plac
cont_table
#>              ocular_discomfort
#> type          2  3  4
#>   active_drug 6 17 37
#>   Placebo     2 13 44
```

Standard procedures

We can see that there are no significant differences between active drug and placebo patients for this test.

```
chisq.test(cont_table)
#>
#>  Pearson's Chi-squared test
#>
#> data:  cont_table
#> X-squared = 3.1301, df = 2, p-value = 0.2091
```

Resampling procedures

The resampling procedure also corroborates with this result, therefore we can see no significant difference between active drug and placebo patients through this method either.

```
chisq.test(cont_table, simulate.p.value=TRUE, B=10000)
#>
#>  Pearson's Chi-squared test with simulated p-value (based on 10000
#>  replicates)
#>
#> data:  cont_table
#> X-squared = 3.1301, df = NA, p-value = 0.2177
```

Exercise 6

```
cont_table = matrix(c(23,45,17,1064,25,51,19,1043),nrow=2,ncol=4,byrow=T,dimnames=list("type"=c("Angiopl
cont_table
#>                  Status
#> type              Cardiac death Other death Unknown cause Alive
#>   Angioplasty                23          45            17  1064
#>   Medical therapy            25          51            19  1043
```

Standard procedures

We can see that there doesn't seem to be any statistically significant association between treatment group and the outcome.

```
chisq.test(cont_table)
#>
#>  Pearson's Chi-squared test
#>
#> data:  cont_table
#> X-squared = 0.72586, df = 3, p-value = 0.8671
```

Exercise 7

```
id <- 1:10
before <- c(6.7, 7.4, 9.2, 9.6, 7.4, 8.1, 10.8, 7.1, 7.9, 10.8)
after <- c(7, 7.4, 8.6, 8.1, 6.8, 7, 8.5, 7.7, 9.7, 7.7)
diff <- before-after
ba <- data.frame(id=id, before=before, after=after, diff=diff)
```

Paired t-test: Standard method

We can conclude that there are no differences among the results before and after the use of the insulin pump on HgbAlc.

```
t.test(ba$before, ba$after ,paired=TRUE)
#>
#>  Paired t-test
#>
#> data:  ba$before and ba$after
#> t = 1.4319, df = 9, p-value = 0.186
#> alternative hypothesis: true difference in means is not equal to 0
#> 95 percent confidence interval:
#>  -0.3768688  1.6768688
#> sample estimates:
#> mean of the differences
#>                    0.65
```

Resampling procedure for paired t-test

Our resampling procedure returns a very similar p-value, therefore it corroborates our results.

```
paired.perm.test(ba$diff)
#> [1] 0.1914062
```

Exercise 8

```
caloric_intake <- c(50,70,90,120,40,100,150,110,75,160)
VO2 <- c(7,8,10.5,11,9,10.8,12,10,9.5,11.9)
```

Testing normality

Testing normality using Shapiro-Wilk test results in enough statistical significance to claim that the samples come from a normal distribution.

```
shapiro.test(caloric_intake)
#>
#>  Shapiro-Wilk normality test
#>
#> data:  caloric_intake
#> W = 0.96646, p-value = 0.8562
shapiro.test(VO2)
#>
#>  Shapiro-Wilk normality test
#>
#> data:  VO2
#> W = 0.95434, p-value = 0.7199
```

Pearson's correlation test

Looking at the result of our correlation test, we obtain a high correlation of ~ 0.88, which suggests there is a significant relationship between the two variables. Our p-value also rejects the null hypothesis that the true correlation is equal to zero.

```r
cor.test(caloric_intake, VO2)
#>
#>  Pearson's product-moment correlation
#>
#> data:  x and y
#> t = 5.2158, df = 8, p-value = 0.0008067
#> alternative hypothesis: true correlation is not equal to 0
#> 95 percent confidence interval:
#>  0.5586333 0.9711674
#> sample estimates:
#>       cor
#> 0.8790656
```

Exercise 9

```r
dbp <- read.table('./dbp.txt', header=TRUE)
```

(i)

Testing normality

According to our Shapiro-Wilk normality test, we obtain that these samples from either group A or B are probably not coming from a normally distributed population.

```r
shapiro.test(dbp[dbp$TRT == 'A','DBP'])
#>
#>  Shapiro-Wilk normality test
#>
#> data:  dbp[dbp$TRT == "A", "DBP"]
#> W = 0.95088, p-value = 0.0009485
shapiro.test(dbp[dbp$TRT == 'B','DBP'])
#>
#>  Shapiro-Wilk normality test
#>
#> data:  dbp[dbp$TRT == "B", "DBP"]
#> W = 0.9668, p-value = 0.01264
```

Wilcoxon-Mann-Whitney test using resampling

According to our test we reject the null hypothesis, therefore there are differences between the diastolic blood pressure of group A and group B.

```r
wilcox_test(DBP~factor(TRT),data=dbp,distribution=approximate(nresample=10000))
#>
#>  Approximative Wilcoxon-Mann-Whitney Test
#>
#> data:  DBP by factor(TRT) (A, B)
#> Z = -5.4917, p-value < 1e-04
#> alternative hypothesis: true mu is not equal to 0
```

Wilcoxon-Mann-Whitney test using resampling over time

Same results are obtained through the Wilcoxon-Mann-Whitney test using resampling over time (stratified by month).

```r
dbp$TRT <- factor(dbp$TRT)
dbp$month <- factor(dbp$month)
dbp <- as.data.frame(dbp)
wilcox_test(DBP~TRT|month,data=dbp,distribution=approximate(nresample=10000))
#>
#>  Approximative Wilcoxon-Mann-Whitney Test
#>
#> data:  DBP by TRT (A, B)
#>   stratified by month
#> Z = -7.9385, p-value < 1e-04
#> alternative hypothesis: true mu is not equal to 0
```

Exercise 10

```r
cont_table = matrix(c(13,37-13,170-13,699-170-37+13),nrow= 2, byrow=T,dimnames=list("control"=c("Y","N"
```

McNemar's test

Standard method

According to our test, there are difference among the hazard and control groups. Therefore the use of a cellular telephone while driving is associated with a significant increase in car accident rate.

```r
mcnemar.test(cont_table, correct = FALSE)
#>
#>  McNemar's Chi-squared test
#>
#> data:  cont_table
#> McNemar's chi-squared = 97.729, df = 1, p-value < 2.2e-16
```

Resampling method

The resampling test returns the same results as the standard test, therefore we can decisively conclude that there's statistical evidence to state that it is most likely significantly more dangerous to use the phone while driving than not.

```r
mh_test(as.table(cont_table), distribution = approximate(nresample =10000))
#>
#>  Approximative Marginal Homogeneity Test
#>
#> data:  response by
#>   conditions (control, case)
#>   stratified by block
#> chi-squared = 97.729, p-value < 1e-04
```

Exercise 11

Resume the main ideas about the False Discovery Rate (FDR).

False discoveries are the number of false positives, which is the type I error (when you incorrectly reject the null hypothesis). Consequently, false discovery rate is the expectation of the proportion of false discoveries among all the discoveries (rejections of all the null hypothesis).

Proportion of false discoveries:

$$Q = \frac{V}{R} = \frac{V}{(V+S)}$$

where V is the number of false discoveries, S is the number of true positives (true discoveries), and

$$R = V + S$$

which is the total number of rejected null hypotheses (discoveries).

$$FDR = Q_e = E[Q]$$

Formally, it can be written as follows,

$$FDR = E[\frac{V}{R}|R > 0] * P[R > 0]$$

Additionally, false discovery rate (FDR) can be used as a method of conceptualizing the rate of type I error in null hypothesis testing when conducting multiple comparisons.

In medical testing, the FDR is getting a "positive" test result but without actually having the disease, it's the complement of the Positive Predictive Value(PPV), which tells the probability of a positive test result being accurate.

Explain the Benjamini-Hochberg and the q-Value procedures.

It is a FDR approach which adjusts the p-value for a series of tests.

P-value gives the probability of a false positive on one single test. Instead of that, q-value gives the proportion of false positives of all the tests done. It is a better critical value to use when running a large number of tests from small samples.

Q-value is the infumum of the probability that H0 is true given that H0 is rejected (the false discovery rate).

B-H procedure is a procedure that decreases the FDR, the steps are as follows:

1. Take the individual p-values in ascending order;
2. Assign ranks to each p-value with the smallest rank 1;
3. Calculate each p-values's B-H critical value by using formula $(\frac{i}{m}) * Q$ where i is the individual p-value's rank, m is the total number of tests done and Q is the false discovery rate chosen by the user;
4. Compare the original p-values to the critical B-H from step 3 and find the largest p-value that is smaller than the critical value.

Show examples of application with R with a comparison between both methods.

Suppose we have done a test related with food, and the data are as follows, the first column indicates each type of food and the second row shows the exact p-value of multiple test of each observation.

```
Input = ("
  Food               pvalues
  Blue_fish          .34
  Bread              .594
  Butter             .212
  Carbohydrates      .384
  Cereals_and_pasta  .074
  Dairy_products     .94
  Eggs               .275
  Fats               .696
  Fruit              .269
  Legumes            .341
  Nuts               .06
  Olive_oil          .008
```

```
  Potatoes           .569
  Processed_meat     .986
  Proteins           .042
  Red_meat           .251
  Semi-skimmed_milk  .942
  Skimmed_milk       .222
  Sweets             .762
  Total_calories     .001
  Total_meat         .975
  Vegetables         .216
  White_fish         .205
  White_meat         .041
  Whole_milk         .039
")
Data11 = read.table(textConnection(Input), header = TRUE)
```

We order the data by the p-values

```
Data11 = Data11[order(Data11$pvalues),]
```

Obtaining the adjusted p-values by using the method Benjamini-Hochberg and q-values, they are the same. We can observe that after adjusting the p-values, we reject only the first hypothesis test instead of rejecting 5.

```
Data11$BH = p.adjust(Data11$pvalues,method = "BH")
Data11$qvalues = qvalue(Data11$pvalues)$qvalues
Data11
#>                 Food pvalues        BH   qvalues
#> 20    Total_calories   0.001 0.0250000 0.0250000
#> 12         Olive_oil   0.008 0.1000000 0.1000000
#> 25        Whole_milk   0.039 0.2100000 0.2100000
#> 24        White_meat   0.041 0.2100000 0.2100000
#> 15          Proteins   0.042 0.2100000 0.2100000
#> 11              Nuts   0.060 0.2500000 0.2500000
#> 5  Cereals_and_pasta   0.074 0.2642857 0.2642857
#> 23        White_fish   0.205 0.4910714 0.4910714
#> 3             Butter   0.212 0.4910714 0.4910714
#> 22        Vegetables   0.216 0.4910714 0.4910714
#> 18      Skimmed_milk   0.222 0.4910714 0.4910714
#> 16          Red_meat   0.251 0.4910714 0.4910714
#> 9              Fruit   0.269 0.4910714 0.4910714
#> 7               Eggs   0.275 0.4910714 0.4910714
#> 1          Blue_fish   0.340 0.5328125 0.5328125
#> 10           Legumes   0.341 0.5328125 0.5328125
#> 4      Carbohydrates   0.384 0.5647059 0.5647059
#> 13          Potatoes   0.569 0.7815789 0.7815789
#> 2              Bread   0.594 0.7815789 0.7815789
#> 8               Fats   0.696 0.8700000 0.8700000
#> 19            Sweets   0.762 0.9071429 0.9071429
#> 6     Dairy_products   0.940 0.9860000 0.9860000
#> 17 Semi-skimmed_milk   0.942 0.9860000 0.9860000
#> 21        Total_meat   0.975 0.9860000 0.9860000
#> 14    Processed_meat   0.986 0.9860000 0.9860000
```