# BRM1_IBA_2021: Week 3

Demi Maric, Ashley, Merle, teamnumber 117

11/3/2020

### Before you start, install your working environment

```
getwd()
```

```
## [1] "/Users/Demi/github/BRM117"
```

```
rm(list=ls())
df<-read.csv("GRC_PHL.csv")
```

## Assignment week 3

This week you will continue to work on the WVS data. Read in the data file.

### 1. *t*-test

Test the differences in emancipation `Q31` between the genders measured by `Q260`. To learn what the answers to these questions mean, have a look at the coding manual. Report the *t*-value, *p*-value, and significance. Which gender is most positive about female business executives? Think about what totally agree (1) means!

```
t.test(df$Q31~df$Q260)
```

```
##
##  Welch Two Sample t-test
##
## data:  df$Q31 by df$Q260
## t = -9.0208, df = 2304.8, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.3625363 -0.2330616
## sample estimates:
## mean in group 1 mean in group 2
##        2.582234        2.880033
```

## 2. Hypothesis

To know more about emancipation within the countries, you would need to know whether males support female business executives. Perform the t-test on emancipation for males across the two countries, and inspect the results. Use indexing! In which country are males more positive about female business executives? Remember, that differences in means only are meaningful for the wider population if significant.

```
t.test(df$Q31[df$Q260==1]~df$Country[df$Q260==1])
```

```
##
##  Welch Two Sample t-test
##
## data:  df$Q31[df$Q260 == 1] by df$Country[df$Q260 == 1]
## t = 7.8149, df = 1118.2, p-value = 1.263e-14
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.2810443 0.4694771
## sample estimates:
## mean in group GRC mean in group PHL
##          2.780261          2.405000
```

Write down the one-sided hypothesis that males in country 1 are more positive about female business executives than males in country 2 in mathematical notation. *Note*: country 1 comes first alphabetically.

## 3. One-sided $t$-test

Test the hypothesis under 2 using a regular $t$-test, and report $t$-value, $p$-value and interpret the significance (use 95% confidence interval). Use indexing!

```
t.test(Q31~Country, df[df$Q260==1,], mu=2.780261,conf.level=0.95, alternative="greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  Q31 by Country
## t = -50.085, df = 1118.2, p-value = 1
## alternative hypothesis: true difference in means is greater than 2.780261
## 95 percent confidence interval:
##  0.2962119       Inf
## sample estimates:
## mean in group GRC mean in group PHL
##          2.780261          2.405000
```

Reflect on the t-test you have conducted above. Under which conditions is the $t$-test you conducted unreliable?

Check the gender-specific means of each country using `tapply` and `list`, and explain why you think emancipation is higher or lower in country 1. Interpret the mean using the original categories (see codebook). Use 2.5 as the cutoff point between agree and disagree.

```
tapply(df$Q31, list(df$Q260,df$Country), mean, na.rm=T)
```

```
##        GRC      PHL
## 1 2.780261 2.405000
## 2 3.139159 2.612688
```

In which country are males more positive about female business executives?

## 4. Assumption: Equal variances

Test whether the variances are equal among males and females in country 2, using the formula below.

$$F = \frac{N-k}{k-1} * \frac{\sum_{i=1}^{k} N_i (Z_{i.} - Z_{..})^2}{\sum_{i=1}^{k} \sum_{j=1}^{N_i} (Z_{ij} - Z_{i.})^2}$$

where $N$ is total number of observations, $N_i$ the observations in group $i$, $k$ is the number of groups, $Z_{ij}$ are the absolute differences with the median in group $i$, $Z_{i.}$ is the mean in group $i$, and $Z_{..}$ is the overall mean across all groups. Be aware that this formula is based on list-wise deletion. Follow the steps below.

Check with `leveneTest` from `car` package. Report $F$-value, $p$-value and significance. *Note*: The `var.test` also tests variances is less robust to deviations of the normal distribution.

```
# First define N, k
# use na.omit to perform listwise deletion
# only for the two variables included!
D<-na.omit(df[c("Q31","Q260")])
N <- nrow(D)
k<-2

# then define the group specific numbers
# Ni and group means Zi
# label them N1, N2, and respectively Z1 and Z2
D1<-nrow(df[df$Q31 & df$Q260==1,])
N1<-na.omit(D1)
D2<-nrow(df[df$Q31 & df$Q260==2,])
N2<-na.omit(D2)
# calculate the Zij
# label them Z1j and Z2j


# finally calculate the overall group mean
# label this Z


# then calculate the numerator and denominator
# label them between_var, and within_var as they represent these variances


# check with leveneTest
library(car)
```

```
## Loading required package: carData
```

```
leveneTest(df$Q260~factor(df$Country == "PHL"),var.equal=T)
```

```
## Levene's Test for Homogeneity of Variance (center = median: T)
##         Df F value  Pr(>F)
## group    1  4.5769 0.03251 *
##       2398
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
t.test(df$Q260~factor(df$Country == "PHL"),var.equal=T)
```

```
##
##  Two Sample t-test
##
## data:  df$Q260 by factor(df$Country == "PHL")
## t = 1.5113, df = 2398, p-value = 0.1308
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.009173068  0.070839735
## sample estimates:
## mean in group FALSE  mean in group TRUE
##            1.530833            1.500000
```

## 5. ANOVA: Test mean differences

Perform an ANOVA to analyze how education groups measured by `Q275R` differ in their views on female emancipation. Use `aov` and `factor` to test the mean differences between the groups. Report the $F$-value, the $p$-value and interpret whether the differences are significant. Save the ANOVA test in an R object.

```
aov1<-aov(df$Q31~as.factor(df$Q275R))
summary(aov1)
```

```
##                       Df Sum Sq Mean Sq F value   Pr(>F)
## as.factor(df$Q275R)    3   47.4   15.81   24.71 9.68e-16 ***
## Residuals           2350 1503.7    0.64
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 46 observations deleted due to missingness
```

Use `$coeff` to get the means used in the ANOVA test. Which education group is the least positive about female emancipation?

```
aov1$coeff
```

```
##          (Intercept) as.factor(df$Q275R)2 as.factor(df$Q275R)3
##            2.5398496            0.2209159            0.2585375
## as.factor(df$Q275R)4
##            0.4222716
```

## 6. Tukey-B

Conduct a Tukey-B test to see which education groups are significantly different. Use `TukeyHSD`. Report the education category of the groups that differ significantly with the primary education group, or that there are no significant differences.

```
TukeyHSD(aov1)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = df$Q31 ~ as.factor(df$Q275R))
##
## $`as.factor(df$Q275R)`
##           diff          lwr       upr     p adj
## 2-1 0.22091593  0.118901578 0.3229303 0.0000002
## 3-1 0.25853747  0.105523568 0.4115514 0.0000860
## 4-1 0.42227159  0.291735243 0.5528079 0.0000000
## 3-2 0.03762155 -0.107638925 0.1828820 0.9099351
## 4-2 0.20135566  0.080000452 0.3227109 0.0001218
## 4-3 0.16373412 -0.002799409 0.3302676 0.0560048
```

## 7. ANOVA: Inspect mean differences

Repeat an ANOVA as you did under 5 for how age groups measured by `Q278R` differ in their views on female emancipation for **country 1**. Save the ANOVA test in an R object. Report the $F$-value, the $p$-value and interpret whether the differences are significant. Use `$coeff` to get the means used in the ANOVA test. Which group is the least positive about female emancipation in country 1?

## 7. Alternative hypothesis ANOVA

Think about the null hypothesis that you test in an ANOVA test. What is an alternative hypothesis of ANOVA?

## 8. Weighing

A weighted mean is used to correct for sampling. If a group occurs significantly more or less in the population than they do in the sample, you can use sampling weights to correct for this discrepancy. You can obtain the weighted means by multiplying the means calculated under 5) with the weight `W_WEIGHT`:

$\sum(m_i * w_i) / \sum(w_i)$,

where $m$ stands for mean, $w$ for weight, and $i$ for individual. Using the formula above, the sampling weight makes sure that at an aggregate level, the sample has the same characteristics as the population. Calculate the weighted means of emancipation `Q31`. Report the weighted mean for country 1. Mind the missings!

```
sum(df$Q31[df$Country=="GRC"] * df$W_WEIGHT[df$Country=="GRC"], na.rm = TRUE)/sum(df$W_WEIGHT[df$Country
```

```
## [1] 2.855069
```