week 5, report

Diana Hilleshein

2/15/2022

```
setwd("C:/R scripts/DWA2022")
library(haven)

# import data
data <- read_sav("C:/R scripts/DWA2022/ESS9e03_1.sav")


save(data,file="data.csv")
head(data)
```

```
## # A tibble: 6 x 572
##   name  essround edition proddate  idno cntry    nwspol netusoft netustm ppltrst
##   <chr>    <dbl> <chr>   <chr>    <dbl> <chr+lb> <dbl+> <dbl+lb> <dbl+l> <dbl+l>
## 1 ESS9~        9 3.1     17.02.2~    27 AT [Aus~     60 5 [Ever~     180   2 [2]
## 2 ESS9~        9 3.1     17.02.2~   137 AT [Aus~     10 5 [Ever~      20   7 [7]
## 3 ESS9~        9 3.1     17.02.2~   194 AT [Aus~     60 4 [Most~     180   5 [5]
## 4 ESS9~        9 3.1     17.02.2~   208 AT [Aus~     45 5 [Ever~     120   3 [3]
## 5 ESS9~        9 3.1     17.02.2~   220 AT [Aus~     30 1 [Neve~      NA   5 [5]
## 6 ESS9~        9 3.1     17.02.2~   254 AT [Aus~     45 2 [Only~      NA   8 [8]
## # ... with 562 more variables: pplfair <dbl+lbl>, pplhlp <dbl+lbl>,
## #   polintr <dbl+lbl>, psppsgva <dbl+lbl>, actrolga <dbl+lbl>,
## #   psppipla <dbl+lbl>, cptppola <dbl+lbl>, trstprl <dbl+lbl>,
## #   trstlgl <dbl+lbl>, trstplc <dbl+lbl>, trstplt <dbl+lbl>, trstprt <dbl+lbl>,
## #   trstep <dbl+lbl>, trstun <dbl+lbl>, vote <dbl+lbl>, prtvtcat <dbl+lbl>,
## #   prtvtdbe <dbl+lbl>, prtvtdbg <dbl+lbl>, prtvtgch <dbl+lbl>,
## #   prtvtbcy <dbl+lbl>, prtvtecz <dbl+lbl>, prtvede1 <dbl+lbl>, ...
```

```
dim(data)#everything is right
```

```
## [1] 49519   572
```

From some previous courses recall topics: cross tabulation, Chi-Square test of independence and test for mean(s). If you find good sources, please add them to our database: https://moodle.helsinki.fi/mod/data/view.php?id=2537669.

Preperations

As well as in a demo file, I will create the restricted data. I beleive it will will significantly increase speed of processing.

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.5     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.1.0     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(dplyr)
library(ggplot2)

#create sum var
data$img_sum<-rowSums(select(data,imsmetn, imdfetn, impcntr), na.rm = TRUE)
head(data$img_sum)
```

```
## [1] 6 8 7 8 9 6
```

```
#create mean var
data$img_avg<-rowMeans(select(data,imsmetn, imdfetn, impcntr), na.rm = TRUE)
head(data$img_avg)
```

```
## [1] 2.000000 2.666667 2.333333 2.666667 3.000000 2.000000
```

```
data5 <- data %>% filter(cntry == "FI") %>%
  select(., idno, cntry, agea, gndr, impcntr, img_sum, img_avg)
head(data5)
```

```
## # A tibble: 6 x 7
##    idno cntry             agea       gndr         impcntr img_sum img_avg
##   <dbl> <chr+lbl>    <dbl+lbl>  <dbl+lbl>       <dbl+lbl>   <dbl>   <dbl>
## 1    19 FI [Finland]        71 1 [Male]   2 [Allow some]        6    2
## 2    57 FI [Finland]        29 1 [Male]   3 [Allow a few]       9    3
## 3    86 FI [Finland]        77 1 [Male]   3 [Allow a few]       7    2.33
## 4   120 FI [Finland]        46 1 [Male]   2 [Allow some]        7    2.33
## 5   164 FI [Finland]        57 2 [Female] 2 [Allow some]        6    2
## 6   238 FI [Finland]        39 2 [Female] 3 [Allow a few]       8    2.67
```

```
dim(data5)
```

```
## [1] 1755    7
```

```
summary(data5$agea)#average age is 51 years, minimum is 15 maximum is 90
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    15.0    36.0    53.0    50.9    66.0    90.0
```

```
summary(data5$impcntr) #average is 2.5
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   1.000   2.000   3.000   2.521   3.000   4.000      37
```

Contingency table (aka cross tabulation)

Choose two variables and create a contingency table (also known as a cross tabulation or crosstab). Discuss your findings.

The interpretation for results:

1) **gndr**-Gender (1 = Male, 2 = Female)

2) **agea** -Age of respondent, calculated (integer)

3) **cntry** - Country of respondent (AT = Austria, BE = Belgium, BG = Bulgaria, CH = Switzerland, CY = Cyprus, CZ = Czechia, DE = Germany, DK = Denmark, EE = Estonia, ES = Spain, FI = Finland, FR = France, GB = United Kingdom, HR = Croatia, HU = Hungary, IE = Ireland, IS = Iceland, IT = Italy, LT = Lithuania, LV = Latvia, ME = Montenegro, NL = Netherlands, NO = Norway, PL = Poland, PT = Portugal, RS = Serbia, SE = Sweden, SI = Slovenia, SK = Slovakia)

4) **impcntr** -Allow many/few immigrants from poorer countries outside Europe (scale from 1 to 4, 1 = Allow many to come and live here, 4 = Allow none)

Contingency table

```
t1 <- table(data5$gndr, data5$impcntr)
t1
```

```
##
##       1   2   3   4
##   1  62 275 436  55
##   2 115 308 408  59
```

In the crosstab we can see the number of responses that fall under one of the categories. The number of observations in each cell in sufficient.

Contingency table can be visualized using the function balloonplot() [in gplots package]. This function draws a graphical matrix where each cell contains a dot whose size reflects the relative magnitude of the corresponding component.

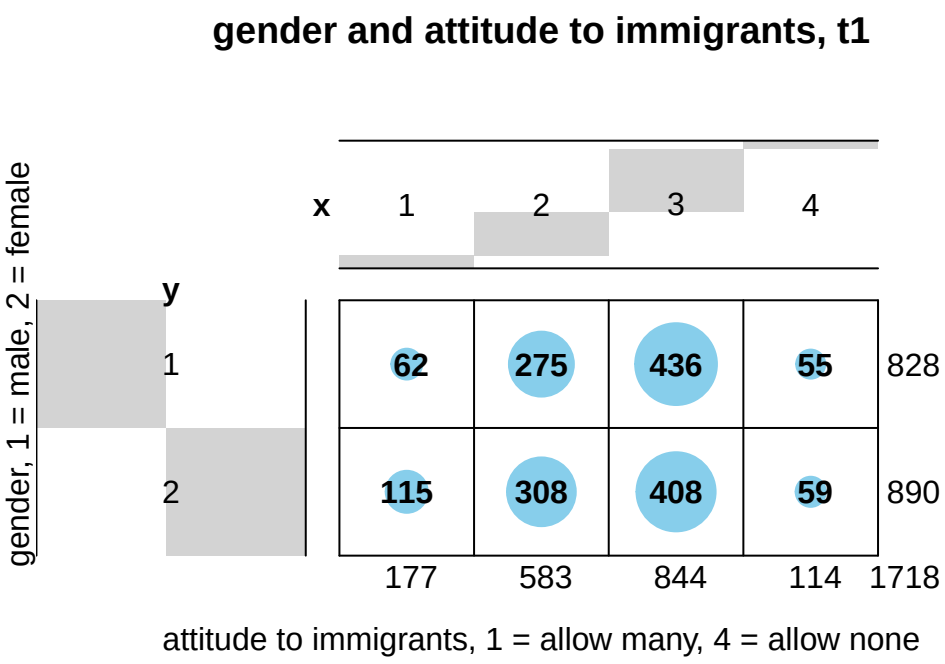
*This website helped me a lot with the topic: http://www.sthda.com/english/wiki/chi-square-test-of-independence-in-r.

```
library(gplots)
```

```
##
## Attaching package: 'gplots'
```

```
## The following object is masked from 'package:stats':
##
##     lowess
```

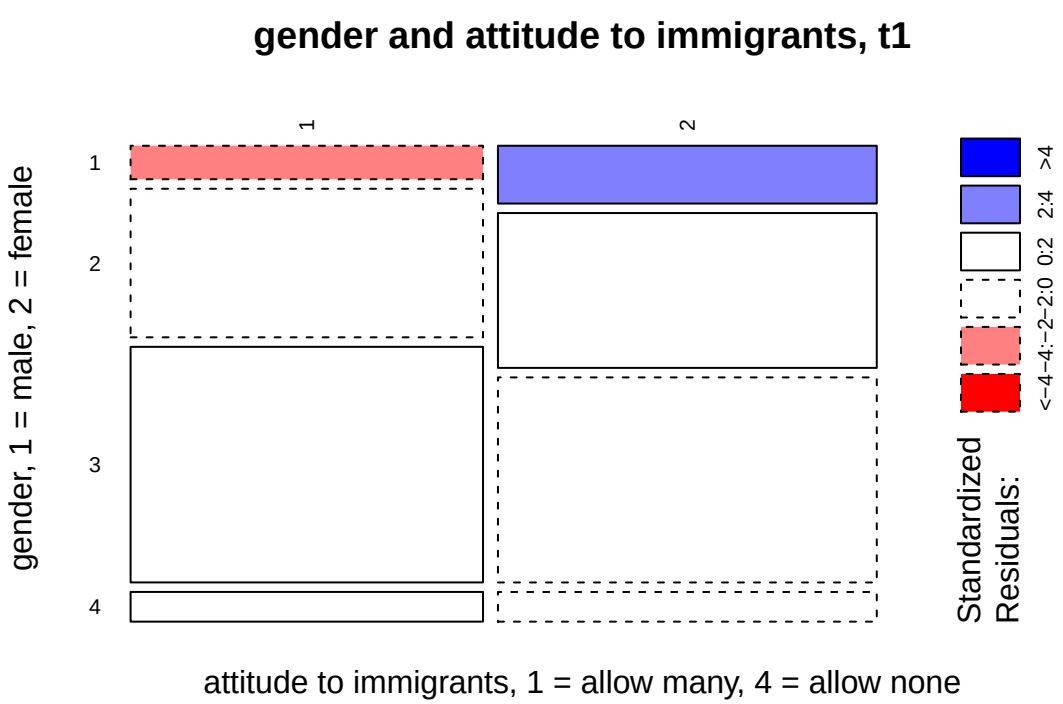
```
library(ggplot2)
library(tidyverse)

balloonplot(t(t1), main ="gender and attitude to immigrants, t1", mtext("attitude to
↪  immigrants, 1 = allow many, 4 = allow none", side = 1), mtext ("gender, 1 = male, 2 =
↪  female", side = 2),
            label = T, show.margins = T)
```



We can see a pattern here, women (2) tend to respond 1 or 2 while men (1) tend to respon 3 or 4. It is not reliable to count by number of response in each cell, but rather by persentage. I will make the same visualization but using persentages later.

It's also possible to visualize a contingency table as a mosaic plot. This is done using the function mosaicplot() from the built-in R package garphics:

```
mosaicplot(t1, shade = TRUE, las=2,
           main ="gender and attitude to immigrants, t1", xlab ="attitude to immigrants,
            ↪  1 = allow many, 4 = allow none", ylab="gender, 1 = male, 2 = female",
            label = T, show.margins = T)
```



- Blue color indicates that the observed value is higher than the expected value if the data were random
- Red color specifies that the observed value is lower than the expected value if the data were random

```
#install.packages("janitor")
library(janitor)
```

```
##
## Attaching package: 'janitor'
```

```
## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test
```

```
t2<- data5 %>% tabyl(impcntr, gndr, show_na=F)
t2 #same as table()
```

```
##  impcntr   1   2
##        1  62 115
##        2 275 308
```

```
##         3 436 408
##         4  55  59
```

```
t3 <- tabyl(data5$gndr, data5$impcntr, show_na=TRUE) #really nice package, good to see
↪  numbers and persetage at the same time. Also, NA is showed separately that is
↪  convinient.
t3
```

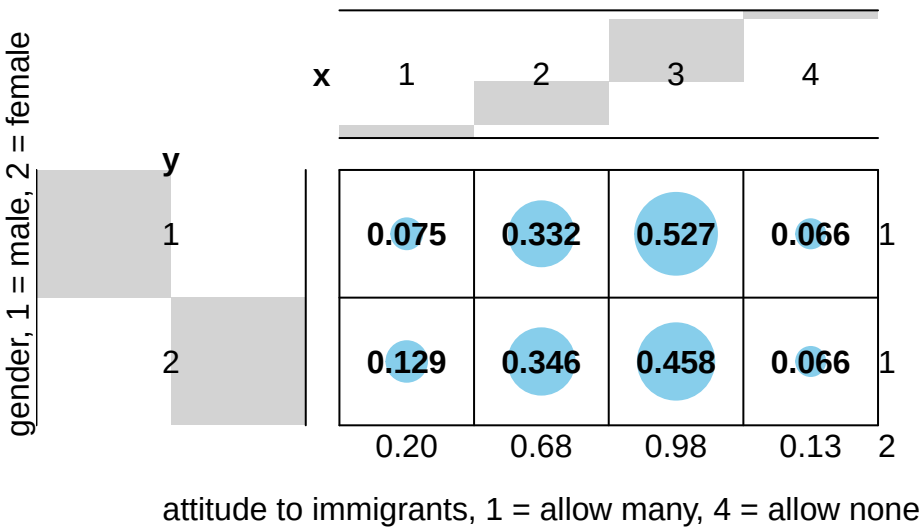
```
##  data5$gndr   n   percent
##           1 848 0.4831909
##           2 907 0.5168091
```

Number of responses are not always easy to observe especially when crosstabs are large. This is why is it beneficial to use prop.table(). I found some docuamntation on the function in using "?prop.table()".

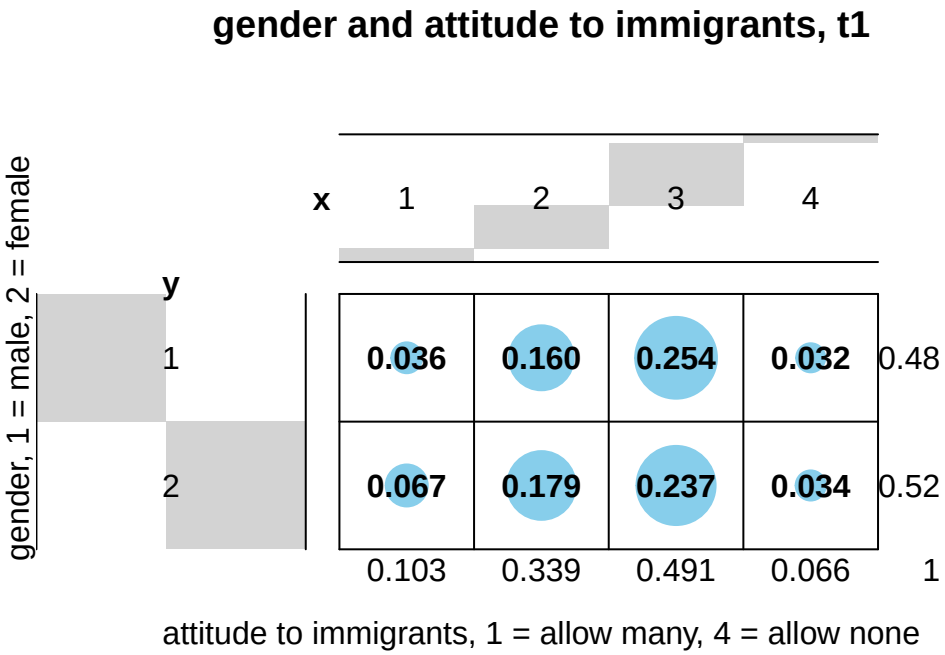
```
prop.table(t1)
```

```
##
##              1          2          3          4
##   1 0.03608847 0.16006985 0.25378347 0.03201397
##   2 0.06693830 0.17927823 0.23748545 0.03434226
```

```
balloonplot(t(prop.table(t1, 1)), main ="gender and attitude to immigrants, t1",
↪  mtext("attitude to immigrants, 1 = allow many, 4 = allow none", side = 1), mtext
↪  ("gender, 1 = male, 2 = female", side = 2),
           label = T, show.margins = T)
```




```
balloonplot(t(prop.table(t1)), main ="gender and attitude to immigrants, t1",
↪  mtext("attitude to immigrants, 1 = allow many, 4 = allow none", side = 1), mtext
↪  ("gender, 1 = male, 2 = female", side = 2),
           label = T, show.margins = T)
```



The pattern didn't change that means that the plot made those calculations before, when we just put numbers as an input.

Here, what we used as an input data for the baloonplot().

```
prop.table(t1) #devide by all cases
```

```
##
##              1          2          3          4
##   1 0.03608847 0.16006985 0.25378347 0.03201397
##   2 0.06693830 0.17927823 0.23748545 0.03434226
```

```
prop.table(t1, 1) #devide by all cases in a row
```

```
##
##              1          2          3          4
##   1 0.07487923 0.33212560 0.52657005 0.06642512
##   2 0.12921348 0.34606742 0.45842697 0.06629213
```

```
prop.table(t1, 2) #devide by all cases in a column
```

```
##
##             1         2         3         4
##   1 0.3502825 0.4716981 0.5165877 0.4824561
##   2 0.6497175 0.5283019 0.4834123 0.5175439
```

```
# X^2-test
# H0: gender and happiness are statistically independent
# H1: gender and happiness are not statistically independent

chq1 <- chisq.test(data5$gndr, data5$impcntr)
chq2 <- data5 %>% tabyl(gndr, impcntr, show_na=F) %>% chisq.test()
chq1
```

```
##
##  Pearson's Chi-squared test
##
## data:  data5$gndr and data5$impcntr
## X-squared = 16.591, df = 3, p-value = 0.0008575
```

```
chq2
```

```
##
##  Pearson's Chi-squared test
##
## data:  .
## X-squared = 16.591, df = 3, p-value = 0.0008575
```

Two different ways of coding are showing the same result. The row and the column variables are statistically significantly associated (p-value < 0).

observed vs. expected frequencies

```
chq1$observed
```

```
##           data5$impcntr
## data5$gndr   1   2   3   4
##          1  62 275 436  55
##          2 115 308 408  59
```

```
chq2$expected
```

```
##  gndr        1        2        3        4
##     1 85.30617 280.9802 406.7707 54.94296
##     2 91.69383 302.0198 437.2293 59.05704
```

Expected and observed values differ a lot for column 1 (allow many).

The most contributing cells to the total Chi-square score:

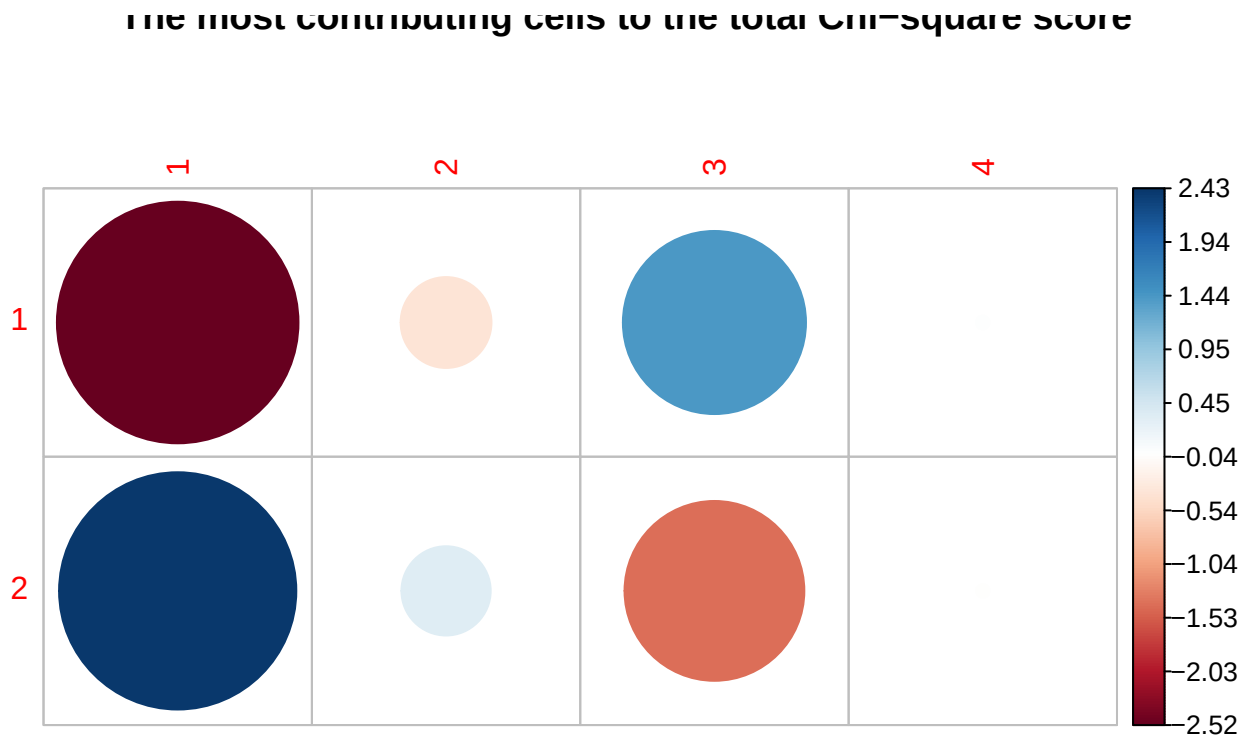
```
# cell contributions:
round(chq1$residuals, 3)
```

```
##            data5$impcntr
## data5$gndr      1      2      3      4
##          1 -2.523 -0.357  1.449  0.008
##          2  2.434  0.344 -1.398 -0.007
```

```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
corrplot(chq1$residuals, is.cor = FALSE, title = "The most contributing cells to the
↪  total Chi-square score")
```



For a given cell, the size of the circle is proportional to the amount of the cell contribution.

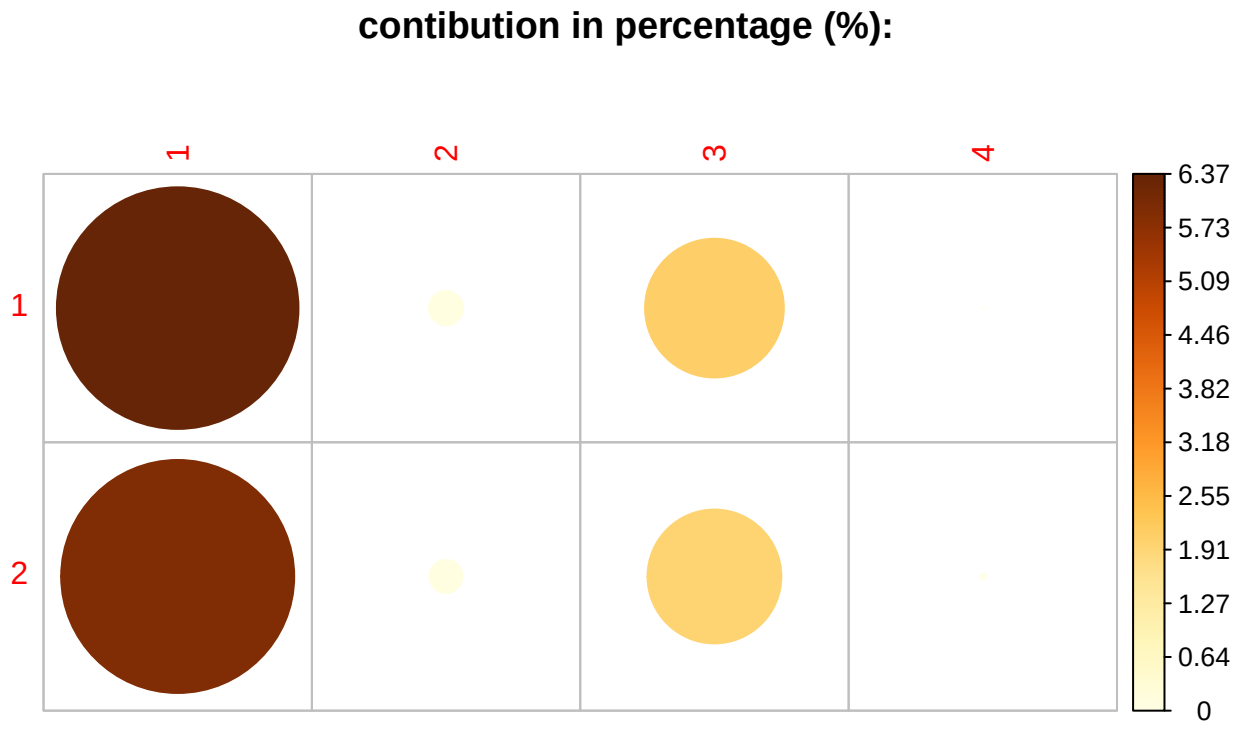
- Positive residuals are in blue. Positive values in cells specify an attraction (positive association) between the corresponding row and column variables. For example, there is no association between male and allow many, but there is a strong asscoaition between male and don‘t allow some.
- Negative residuals are in red. This implies a repulsion (negative association) between the corresponding row and column variables.

The most contributing cells to the total Chi-square score, contibution in percentage (%):

```
cell_contrib <- (chq1$observed - chq1$expected)^2/chq1$expected
round(cell_contrib, 3)
```

```
##           data5$impcntr
## data5$gndr     1     2     3     4
##          1 6.367 0.127 2.100 0.000
##          2 5.924 0.118 1.954 0.000
```

```
corrplot(cell_contrib, is.cor = FALSE, title = "The most contributing cells to the total
↪  Chi-square score,\n contibution in percentage (%): ")
```



The relative contribution of each cell to the total Chi-square score give some indication of the nature of the dependency between rows and columns of the contingency table. The column males is strongly associacted with not allow immigrants at all, while the same association for women is slighly weaker.

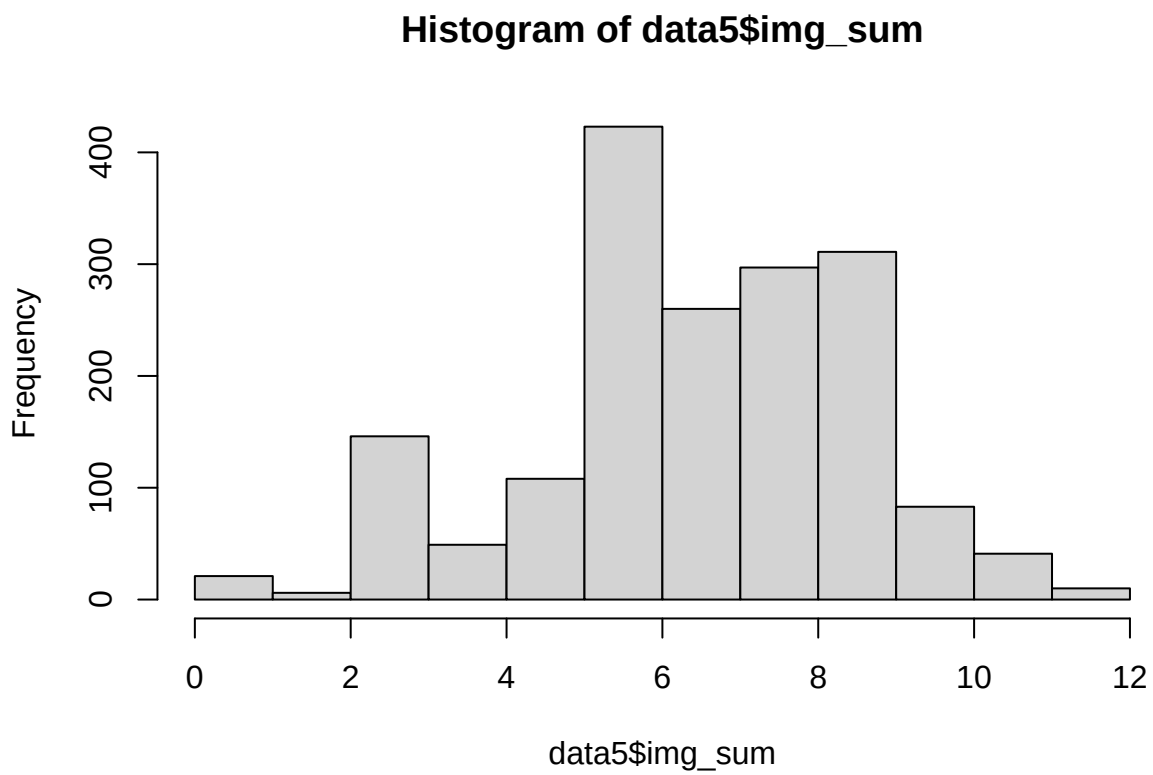
From the figure above, it can be seen that the most contributing cells to the Chi-square are male/allow many (6.367 %), female/allow many (5.924 %), male/do not allow some (2.101 %), females/do not tallow some (1.954 %).

These cells contribute about 16.35% to the total Chi-square score and thus account for most of the difference between expected and observed values.

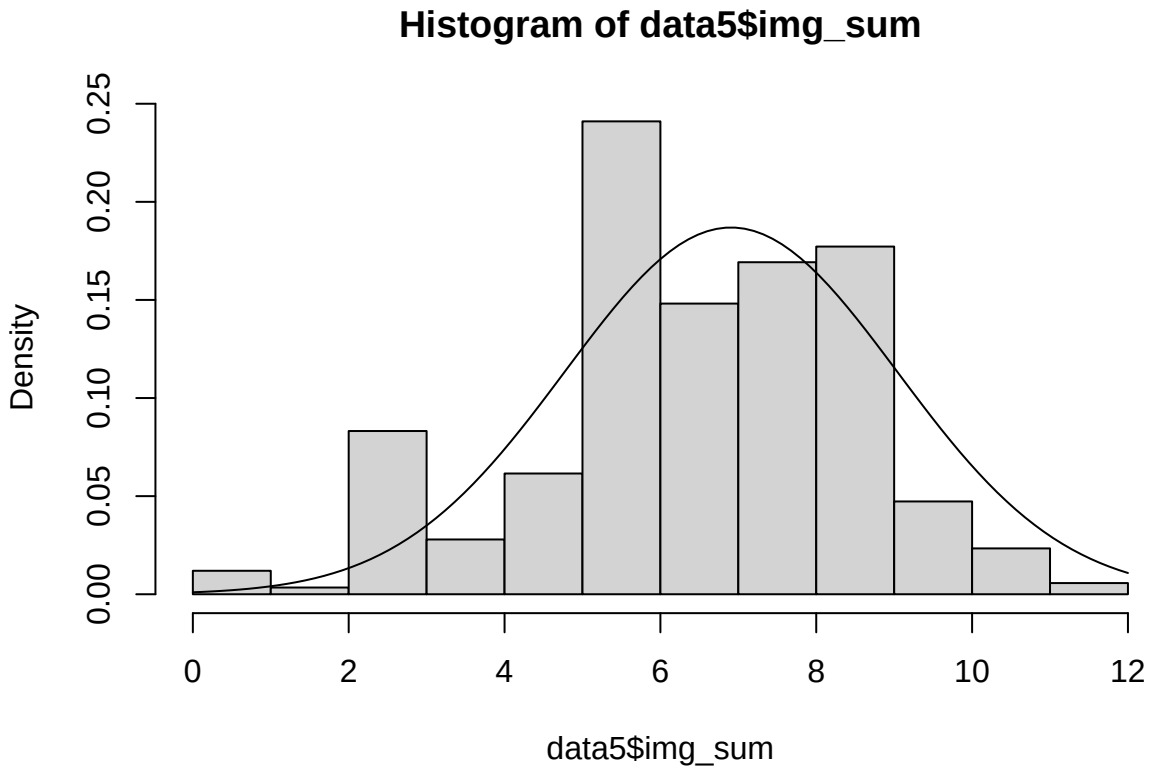
Testing sum variable

I will use sum variable from the previous exercise and gender variable.

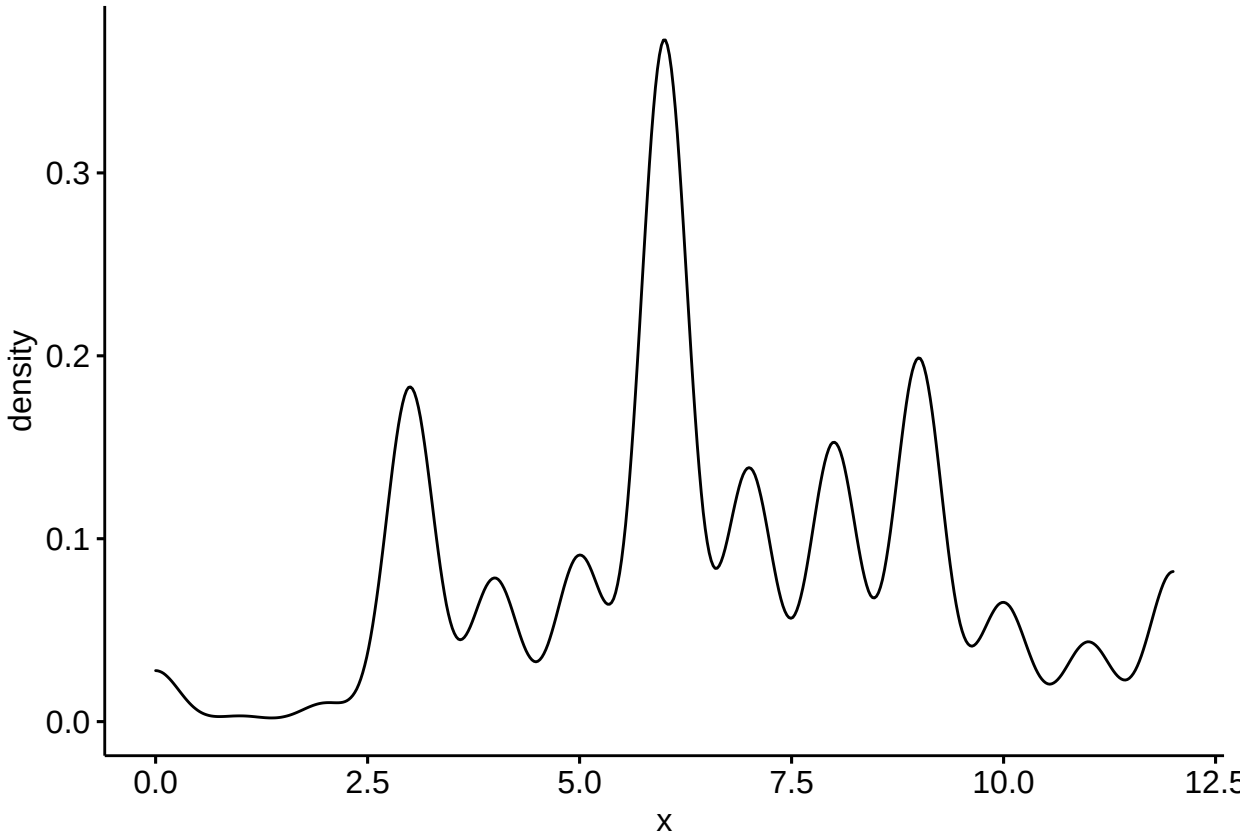
```
hist(data5$img_sum)
```




```
# histogram with normal curve
x <- data5$img_sum; s <- sd(x, na.rm = TRUE); m <- mean(x, na.rm = TRUE)
hist(data5$img_sum, probability=TRUE)
curve(dnorm(x, mean=m, sd=s), add=TRUE)
```




```
library(ggpubr)
ggdensity(data$img_sum,  na.rm = T)
```



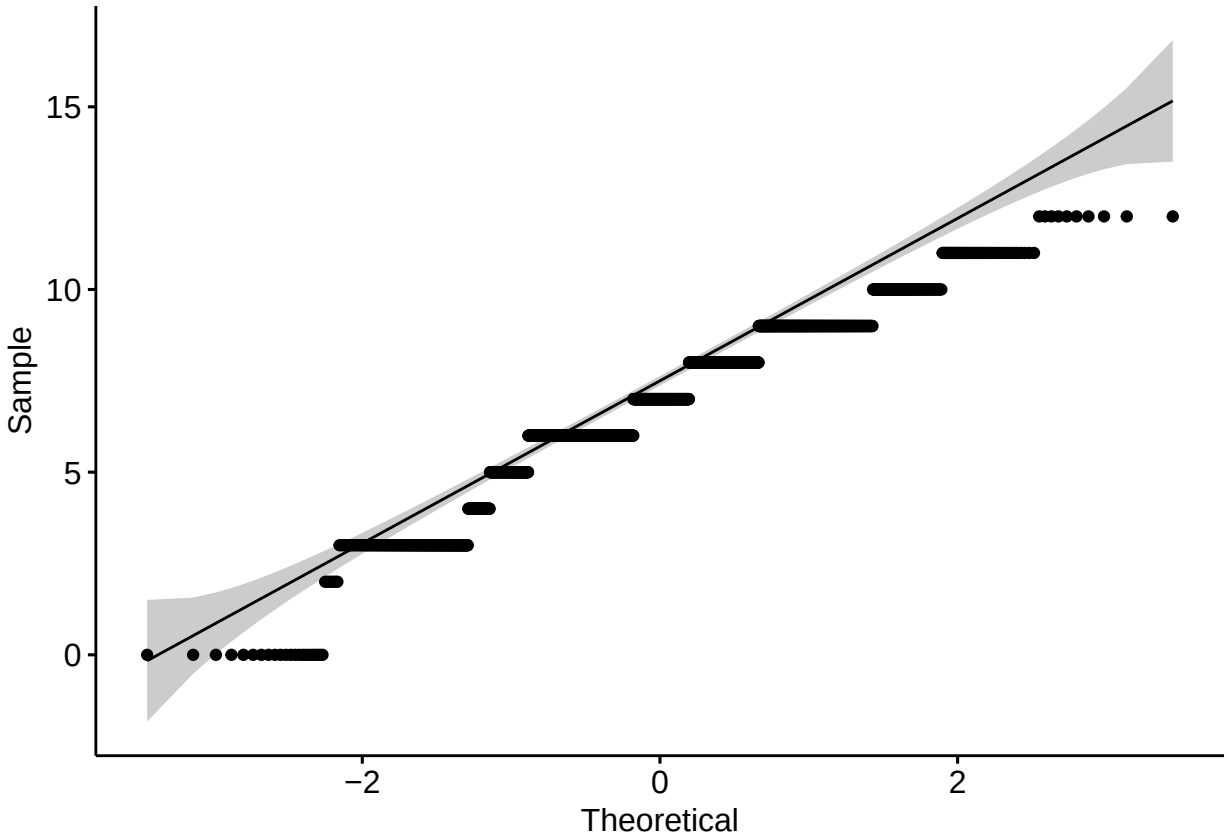

```
#find skewness
#install.packages("moments")
library(moments)
#?skewness

print(skewness(data5$img_avg, na.rm = T))
```

```
## [1] -0.2512158
```

The curve looks normal, but the distribution of responses does not look normal for me. I don‘t know for with reason curve does not show skewness. However, skewness is -0.25, that indicantes that the data is fairly simmetrical according to https://www.spcforexcel.com/knowledge/basic-statistics/are-skewness-and-kurtosis-useful-statistics#:~:text=So%2C%20when%20is%20the%20skewness,the%20data%20are%20highly%20skewed.

The density plots (curves) look very different and I cound‘t understand what is the difference and wich one I need to use...

```
# qqplot
library(ggpubr)
ggqqplot(data5$img_sum)
```



The dots are following the line, with some shift at the upper end. The plot does not look good since there are some frequent categories of responses. The website that helped me to recall this matherial: https://desktop.arcgis.com/en/arcmap/latest/extensions/geostatistical-analyst/normal-qq-plot-and-general-qq-plot.htm#:~:text=Points%20on%20the%20Normal%20QQ,deviate%20from%20the%20reference%20line.. It seems like this plot shows that the variable is nornmally distributed.

The Shapiro-Wilk's test or Shapiro test is a normality test in frequentist statistics. The null hypothesis of Shapiro's test is that the population is distributed normally. It is among the three tests for normality designed for detecting all kinds of departure from normality. If the value of p is equal to or less than 0.05, then the hypothesis of normality will be rejected by the Shapiro test. On failing, the test can state that the data will not fit the distribution normally with 95% confidence.

```
shapiro.test(data5$img_sum)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data5$img_sum
## W = 0.95221, p-value < 2.2e-16
```

According to https://www.geeksforgeeks.org/shapiro-wilk-test-in-r-programming/#:~:text=The%20Shapiro%2DWilk's%20test%20or,normality%20test%20in%20frequentist%20statistics.&text=If%20the%20value%20of%20p,distribution%20normally%20with%2095%25%20confidence., we can not assume normality. The p-value is smaller than 0.05. Hence, the distribution of the given data is significantly different from normal distribution.
This result made me confused :(. The calculated skewness indicated that the distribution is normal, while the shapiro test rejected the H0: variable is normally distributed.

```r
ks.test(x = data5$img_sum, y = pnorm)
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  data5$img_sum
## D = 0.98327, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

The test on the mean of the sample compares the hypothesis H0 with H1. The p-value is 2.2e-16. At risk 5%, we reject H0: https://toltex.imag.fr/OneTestSPLS. I can not conduct t-test since my distribution in not normal. Thereofre, I need to condct non parametric means test, Mann-Whitney test.

means test

Before testing means, I need to create two subsets.

```r
# Test some fixed value
# H0: mu1 = mu2, H1: mu1 ne mu2
# males and females separately:
img_sum_males <- data  %>% filter(gndr == 1) %>% select(img_sum)
img_sum_females <- data  %>% filter(gndr == 2) %>% select(img_sum)


img_sum_males <- as.numeric(unlist(img_sum_males))
img_sum_females <- as.numeric(unlist(img_sum_females))


# test first the equality of variances
var.test(img_sum ~ gndr, data = data)
```

```
##
##  F test to compare two variances
##
## data:  img_sum by gndr
## F = 0.93345, num df = 23019, denom df = 26498, p-value = 6.679e-08
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.9104351 0.9570657
## sample estimates:
## ratio of variances
##          0.9334489
```

F value is quite close to 1, indicating that the variances are similar, p-value < 0.05. There is not significant difference between the variances of the two sets of data.
Since the variation is not significantly different, I can assume that bot groups are not normally distributed. Therefore, I can not use t-test.

Using the Mann-Whitney-Wilcoxon Test, we can decide whether the population distributions are identical without assuming them to follow the normal distribution.

```
library(rstatix) #for second type of Mann-Whitney-Wilcoxon Test
```

```
##
## Attaching package: 'rstatix'

## The following object is masked from 'package:janitor':
##
##     make_clean_names

## The following object is masked from 'package:stats':
##
##     filter
```

```
#install.packages("rstatix")
#install.packages("coin")
library(coin)
```

```
## Loading required package: survival

##
## Attaching package: 'coin'

## The following objects are masked from 'package:rstatix':
##
##     chisq_test, friedman_test, kruskal_test, sign_test, wilcox_test
```

```
data5$gndr <- as.data.frame(data5$gndr)
data5$gndr <- as.numeric(unlist(data5$gndr))
wilcox.test(img_sum_males, img_sum_females) #it was hard to interpret
```

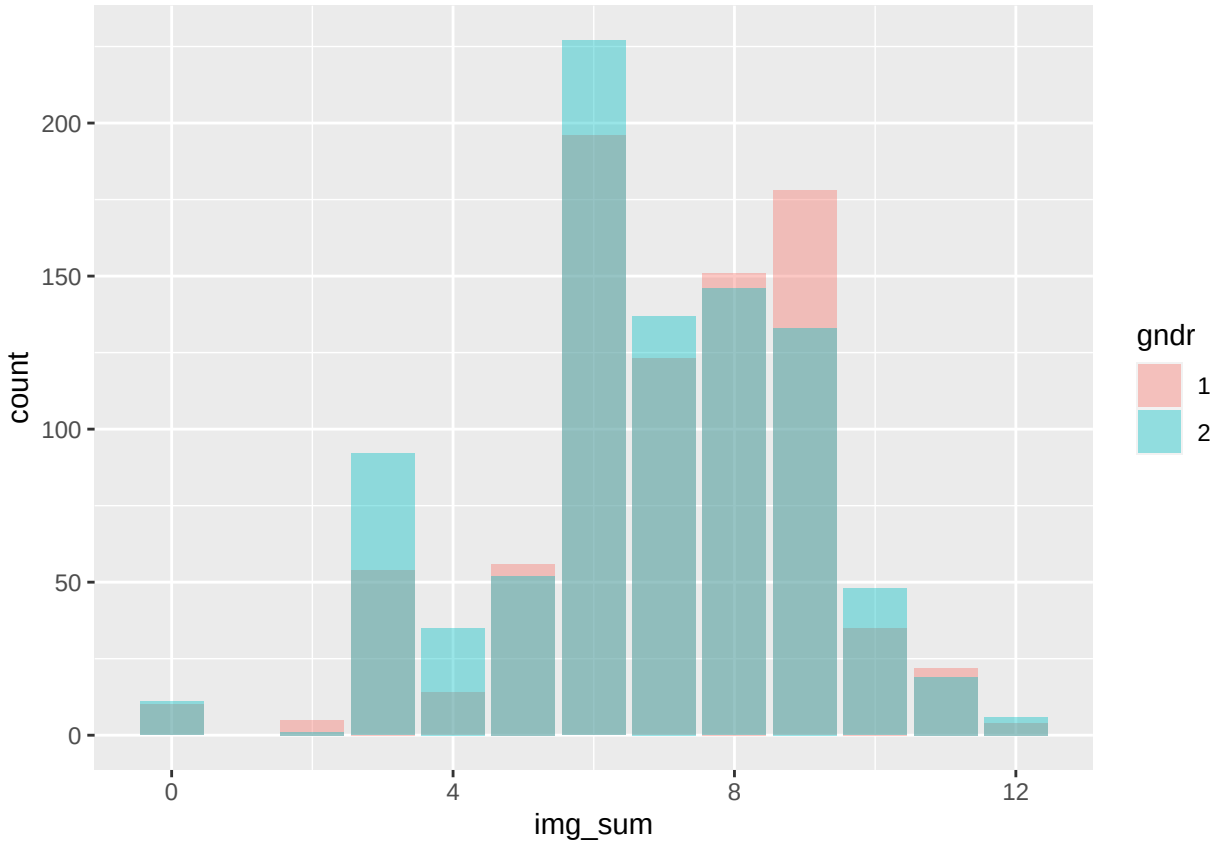
```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  img_sum_males and img_sum_females
## W = 311107639, p-value = 9.804e-05
## alternative hypothesis: true location shift is not equal to 0
```

```
data5 %>%wilcox_effsize(img_sum ~ gndr)#tried diffrent way
```

```
## # A tibble: 1 x 7
##   .y.     group1 group2 effsize    n1    n2 magnitude
## * <chr>   <chr>  <chr>    <dbl> <int> <int> <ord>
## 1 img_sum 1      2       0.0773   848   907 small
```

The Mann-Whitney U test results in a two-sided test p-value < 0.05. This indicates that we should reject the null hypothesis that distributions are equal and conclude that there is a significant difference between genders.
https://www.rdocumentation.org/packages/rstatix/versions/0.7.0/topics/wilcox_effsize helped to find an effect size.
Here the results show a small effect size (0.077), therefore there is no statistical significance for the difference between two vectors. It can be vizualized as fillowing:

```r
data5$gndr <- as.factor(data5$gndr)
ggplot(data5, aes(img_sum, fill = gndr)) +
  geom_bar(position = 'identity', alpha = 0.4)
```



To sum up, there is no big difference in attitued to immigrants between male and female respondents.