Hypothesis Testing

SISS - Applied Statistics - Chiara Seghieri and Costanza Tortù

2023-11-16

Preliminaries

Recall packages

Import Data

```
rm(list=ls())
value <- read_dta("~/Documents/Sant'Anna/Corso allievi/Data/Value Survey descrittive CI e test/WV6_Data
```

Have a first look at data

```
dim(value)
```

```
## [1] 89565    12
```

```
colnames(value)
```

```
##  [1] "ID"          "cow"         "lifesat"     "age"         "education"
##  [6] "relativism"  "scepticism"  "equality"    "choice"      "voice"
## [11] "trust"       "male"
```

```
head(value)
```

```
## # A tibble: 6 x 12
##       ID cow        lifesat age   education relativism scepticism equality choice
##    <dbl> <dbl+lbl>  <dbl+l> <dbl> <dbl+lbl> <dbl+lbl>  <dbl+lbl>  <dbl+lb> <dbl+>
## 1      1 615 [Alge~ 8 [8]   21    7 [Compl~ 0.333      0.44       0        0.0741
## 2      2 615 [Alge~ 5 [5]   24    7 [Compl~ 0.333      0.22       0.11     0
## 3      3 615 [Alge~ 4 [4]   26    5 [Compl~ 0.333      0.663      0        0.111
## 4      4 615 [Alge~ 8 [8]   28    6 [Incom~ 0.333      0.663      0.387    0
## 5      5 615 [Alge~ 8 [8]   35    3 [Compl~ 0.333      0.55       0.22     0.0741
## 6      6 615 [Alge~ 7 [7]   36    8 [Some ~ 0.333      0.644      0.61     0.111
## # i 3 more variables: voice <dbl+lbl>, trust <dbl+lbl>, male <dbl+lbl>
```

Simplify data

Here we apply some simplifications on data - i) Here we keep only observations with no missings (this is not the right procedure to deal with missings of course :-)) - ii) We focus on a subsample of countries (#360 Romania 255 Germany 380 Sweden 230 Spain)

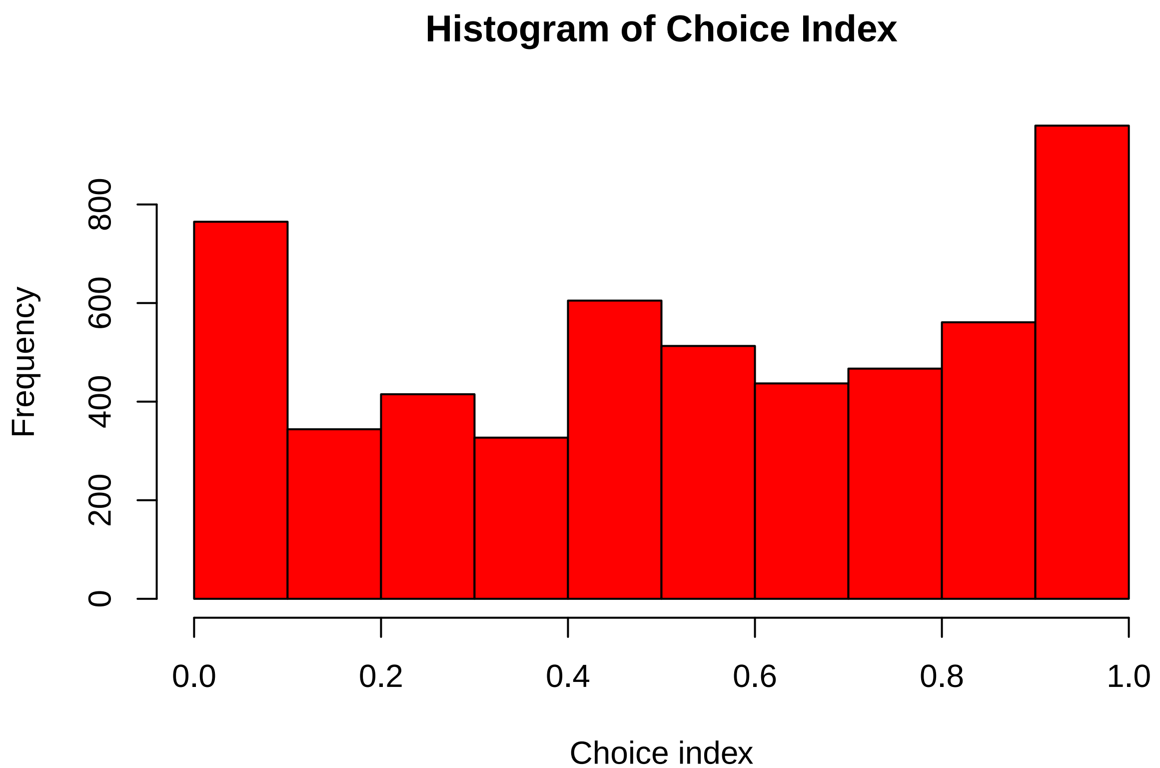
```r
value <- value[complete.cases(value),]

included_countries <- c(360,255,380,230)
value <- value[which(value$cow %in% included_countries),]
value$cow <- as.factor(value$cow)
levels(value$cow) <- c("Spain", "Germany", "Romania", "Sweeden")
```

Inspect variables

```r
quantitative_variables <- c("lifesat", "age", "relativism", "scepticism", "equality", "choice","voice")
dummies <- c( "male", "trust")
factors <- c("cow", "education")
qualitative_variables <- c(dummies, factors)
```

Look at the distribution of the Choice vari<ble

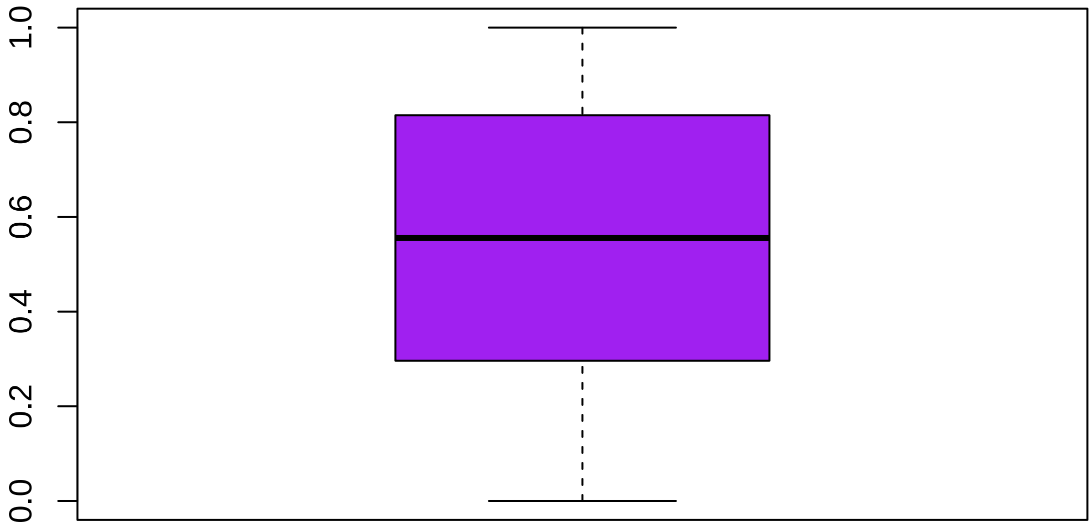
```r
hist(value$choice,
     main ="Histogram of Choice Index",
     col = "red",
     xlab = "Choice index")
```




```r
summary(value$choice)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.2963  0.5556  0.5364  0.8148  1.0000
```

```
boxplot(value$choice,
        col="purple")
```



Hypothesis testing for one sample

Compute sample mean

```
  sample.mean <- mean(value$choice)
  print(sample.mean)
```

```
## [1] 0.5363864
```

Compute sample variance

```
  sample.var <- var(value$choice)
  print(sample.var)
```

```
## [1] 0.1090881
```

Test the hypothesis that the mean of choice is equal to 0.7 (two sided test)

1) Traditional approach

```
alpha = 0.05

tscore <- (sample.mean - 0.7 ) / (sqrt(sample.var / nrow(value)))
tscore
```

```
## [1] -36.38197
```

```
t1 <- qt( p = alpha/2, df = nrow(value) - 1)
accept_region <- c( - t1, t1)
```

The compyted t score is outside the critical region. So, we REJECT HO

2) P-Value Approach

```r
alpha = 0.05

tscore <- (sample.mean - 0.7 ) / (sqrt(sample.var / nrow(value)))
tscore
```

```
## [1] -36.38197
```

```r
pvalue <- pt( q = tscore, df = nrow(value) - 1)
pvalue
```

```
## [1] 1.124885e-259
```

Pvalue is much smaller than alpha. Reject Ho

You can simply use the t.test function

```r
alpha = 0.05

t.test(value$choice,
       mu = 0.7,
       alternative = "two.sided",
       conf.level = 1 - alpha)
```

```
##
##  One Sample t-test
##
## data:  value$choice
## t = -36.382, df = 5393, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0.7
## 95 percent confidence interval:
##  0.5275702 0.5452025
## sample estimates:
## mean of x
## 0.5363864
```

Test the hypothesis that the mean of choice is smaller than 0.7 (one sided test)

```r
alpha = 0.05

t.test(value$choice,
       mu = 0.7,
       alternative = "less",
       conf.level = 1 - alpha)
```

```
##
##  One Sample t-test
##
## data:  value$choice
## t = -36.382, df = 5393, p-value < 2.2e-16
## alternative hypothesis: true mean is less than 0.7
## 95 percent confidence interval:
##       -Inf 0.5437847
## sample estimates:
## mean of x
## 0.5363864
```

Test the hypothesis that the mean of choice is greater than 0.7 (one sided test)

```r
alpha = 0.05

t.test(value$choice,
       mu = 0.7,
       alternative = "greater",
       conf.level = 1 - alpha)
```

```
##
##  One Sample t-test
##
## data:  value$choice
## t = -36.382, df = 5393, p-value = 1
## alternative hypothesis: true mean is greater than 0.7
## 95 percent confidence interval:
##  0.528988      Inf
## sample estimates:
## mean of x
## 0.5363864
```

Test the hypothesis that the mean of choice is eqyal to 0.7 (two sided test), alpha = 0.1

```r
alpha = 0.1

t.test(value$choice,
       mu = 0.7,
       alternative = "two.sided",
       conf.level = 1 - alpha)
```

```
##
##  One Sample t-test
##
## data:  value$choice
## t = -36.382, df = 5393, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0.7
## 90 percent confidence interval:
##  0.5289880 0.5437847
## sample estimates:
## mean of x
## 0.5363864
```

Let's pretend we have information about the population variance of the choice index. Let's say we know this is equal to 0.10. We can test the hypothesis that the mean of the choice index is equal than 0.7 using the z score.

```r
alpha = 0.05

z.test(value$choice,
       mu = 0,
       sigma.x = 0.1,
       alternative = "two.sided",
       conf.level = 1 - alpha)
```

```
##
##  One-sample z-Test
##
## data:  value$choice
```

```
## z = 393.94, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.5337177 0.5390550
## sample estimates:
## mean of x
## 0.5363864
```

Hypothesis testing to compare two samples

Let's focus on two countries: Germany and Romania. We want to test some hypothesis aboiut the difference im the mean choice index in the two countries.

Compute all the quantities you need

```r
n1 <- length(which(value$cow == "Germany"))
xbar1 <- mean(value$choice[which(value$cow == "Germany")])
s1 <- var(value$choice[which(value$cow == "Germany")])

xbar1
```

```
## [1] 0.5322901
```

```r
s1
```

```
## [1] 0.0763298
```

```r
n2 <- length(which(value$cow == "Romania"))
xbar2 <- mean(value$choice[which(value$cow == "Romania")])
s2 <-  var(value$choice[which(value$cow == "Romania")])

xbar2
```

```
## [1] 0.2323472
```

```r
s2
```

```
## [1] 0.06763639
```

Define a subset of the initial dataset that includes only observations from Romania and Germanua

```r
value_gr <- value[which(value$cow %in% c("Romania", "Germany")),]
```

Test the hypothesis that the mean of choice is eqyal in the two countries (two sided test)

```r
alpha = 0.05

t.test(value_gr$choice[which(value_gr$cow == "Germany")],
       value_gr$choice[which(value_gr$cow == "Romania")],
       mu = 0,
       alternative = "two.sided",
       conf.level = 1 - alpha)
```

```
##
##  Welch Two Sample t-test
##
## data:  value_gr$choice[which(value_gr$cow == "Germany")] and value_gr$choice[which(value_gr$cow == "
## t = 31.407, df = 2926.7, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
```

```
## 95 percent confidence interval:
##  0.2812171 0.3186686
## sample estimates:
## mean of x mean of y
## 0.5322901 0.2323472
```

Test the hypothesis that the difference in means of choice is greater than 0.2 (one sided test)

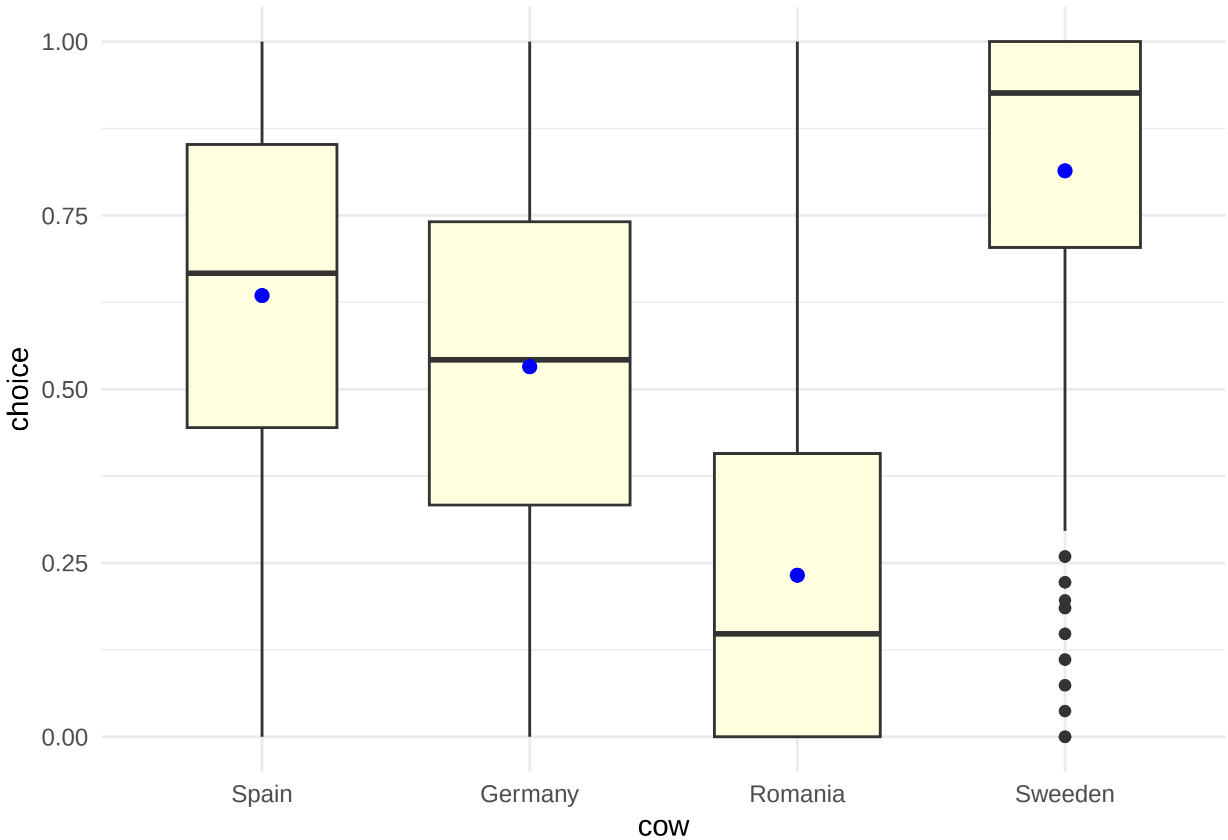
```
alpha = 0.05

t.test(value_gr$choice[which(value_gr$cow == "Germany")],
       value_gr$choice[which(value_gr$cow == "Romania")],
       mu = 0.2,
       alternative = "greater",
       conf.level = 1 - alpha)
```

```
##
##  Welch Two Sample t-test
##
## data:  value_gr$choice[which(value_gr$cow == "Germany")] and value_gr$choice[which(value_gr$cow == "R
## t = 10.465, df = 2926.7, p-value < 2.2e-16
## alternative hypothesis: true difference in means is greater than 0.2
## 95 percent confidence interval:
##  0.2842292       Inf
## sample estimates:
## mean of x mean of y
## 0.5322901 0.2323472
```

ANOVA

```
value %>%
  ggplot(aes(cow, choice)) +
  geom_boxplot(fill = "lightyellow",
               varwidth = T) +              # proporzionale alla variabilità
  stat_summary(fun.data = mean_cl_boot,     # punto che indica le medie
               geom = "point",
               size = 2, col = "blue")+
  theme(legend.position = "null") +
  theme_minimal()
```

```
## Don't know how to automatically pick scale for object of type
## <haven_labelled/vctrs_vctr/double>. Defaulting to continuous.
```




```
aov.res <- aov(data = value,
               choice ~ cow)

summary(aov.res)
```

```
##                Df Sum Sq Mean Sq F value Pr(>F)
## cow             3  215.6   71.86    1039 <2e-16 ***
## Residuals    5390  372.7    0.07
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
coef(aov.res)
```

```
## (Intercept)  cowGermany  cowRomania  cowSweeden
##   0.6346301  -0.1023401  -0.4022829   0.1794577
```

```
summary.aov(aov.res,
            split = list(cow = list("Germany" = 1,
            "Romania" = 2, "Sweeden" = 3)))
```

```
##                   Df Sum Sq Mean Sq  F value Pr(>F)
## cow                3  215.6   71.86 1039.078 <2e-16 ***
##   cow: Germany     1    0.1    0.05    0.725  0.395
##   cow: Romania     1  198.1  198.15 2865.277 <2e-16 ***
##   cow: Sweeden     1   17.4   17.37  251.231 <2e-16 ***
## Residuals       5390  372.7    0.07
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Paired Data

What happens with paired data (just a toy example)

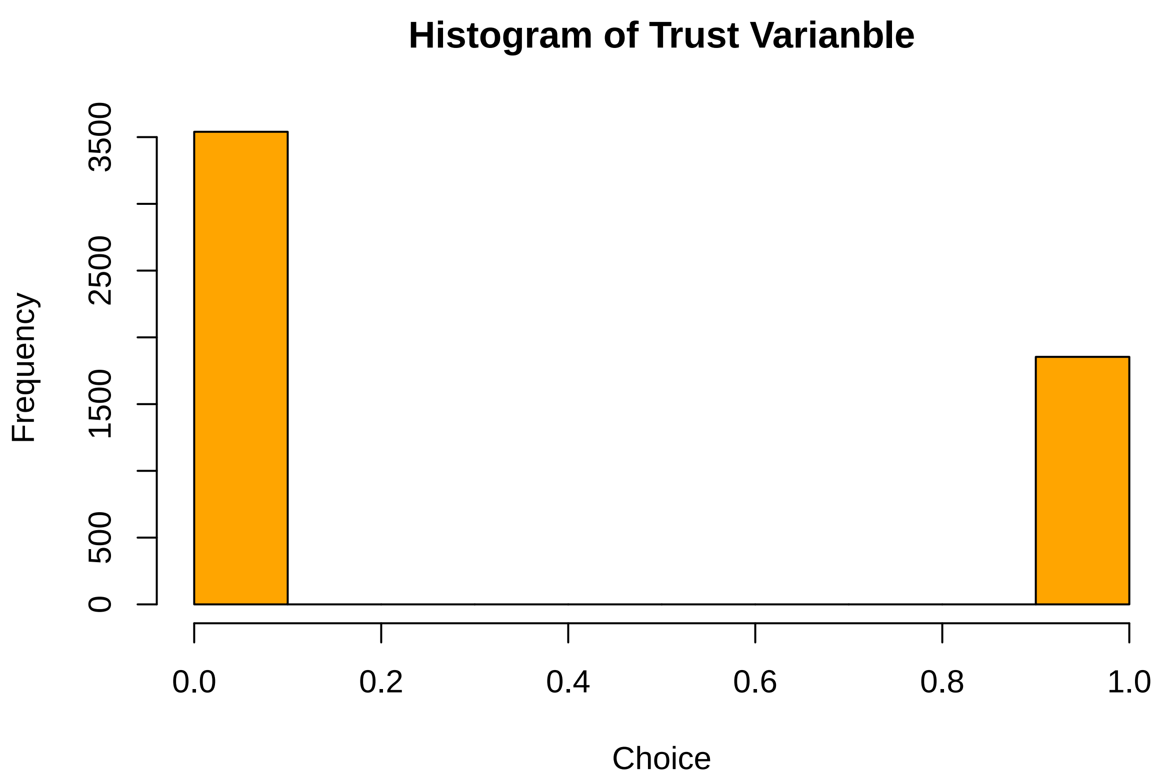
```
before <- c(122, 124, 120, 119, 119, 120, 122, 125, 124, 123, 122, 121)
after <- c(123, 125, 120, 124, 118, 122, 123, 128, 124, 125, 124, 120)

t.test(x = before,
       y = after,
       paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  before and after
## t = -2.5289, df = 11, p-value = 0.02803
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  -2.3379151 -0.1620849
## sample estimates:
## mean difference
##           -1.25
```

Test for proportions (one sample)

Look at the distribution of the Trust vari<ble

```
hist(value$trust,
     main ="Histogram of Trust Varianble",
     col = "orange",
     xlab = "Choice")
```




```
summary(value$trust)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  0.0000  0.3437  1.0000  1.0000
```

Compute sample proportion

```
sample.prop <- mean(value$trust)
print(sample.prop)
```

```
## [1] 0.3437152
```

Compute sample variance

```
sample.var <-   sample.prop * (1 -  sample.prop)
print(sample.var)
```

```
## [1] 0.2255751
```

Count the number of successes

```
success <-  length(which(value$trust == 1))
n = nrow(value)
```

Test the hypothesis that the proportion of trust is equal to 0.5

```r
alpha = 0.05

prop.test(success, n,
       p = 0.5,
       alternative = "two.sided",
       conf.level = 1 - alpha)
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  success out of n, null probability 0.5
## X-squared = 526.37, df = 1, p-value < 2.2e-16
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.3310640 0.3565906
## sample estimates:
##         p
## 0.3437152
```

Test the hypothesis that the proportion of trust is smaller than 0.5 (one sided test)

```r
alpha = 0.05

prop.test(success, n,
       p = 0.5,
       alternative = "less",
       conf.level = 1 - alpha)
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  success out of n, null probability 0.5
## X-squared = 526.37, df = 1, p-value < 2.2e-16
## alternative hypothesis: true p is less than 0.5
## 95 percent confidence interval:
##  0.0000000 0.3545215
## sample estimates:
##         p
## 0.3437152
```

Test the hypothesis that the proportion of trust is smaller than 0.3 (one sided test)

```r
alpha = 0.05

prop.test(success, n,
       p = 0.3,
       alternative = "less",
       conf.level = 1 - alpha,
         correct = FALSE)
```

```
##
##  1-sample proportions test without continuity correction
##
## data:  success out of n, null probability 0.3
## X-squared = 49.086, df = 1, p-value = 1
## alternative hypothesis: true p is less than 0.3
```

```
## 95 percent confidence interval:
##  0.0000000 0.3544282
## sample estimates:
##         p
## 0.3437152
```

Test the hypothesis that the proportion of trust is eqyal to 0.5 (two sided test), alpha = 0.1

```
alpha = 0.1

prop.test(success, n,
       p = 0.5,
       alternative = "two.sided",
       conf.level = 1 - alpha,
         correct = FALSE)
```

```
##
##  1-sample proportions test without continuity correction
##
## data:  success out of n, null probability 0.5
## X-squared = 526.99, df = 1, p-value < 2.2e-16
## alternative hypothesis: true p is not equal to 0.5
## 90 percent confidence interval:
##  0.3331590 0.3544282
## sample estimates:
##         p
## 0.3437152
```

Test for proportions (two samples)

Let's focus on two countries: Germany and Romania. We want to test some hypothesis aboiut the difference in the proportion of people who trust in others in the two countries.

Compute all the quantities you need

```
n1 <- length(which(value$cow == "Germany"))
p1 <- mean(value$trust[which(value$cow == "Germany")])
s1 <- p1 * (1-p1)
success1 <- length(which(value$trust == 1 & value$cow == "Germany"))

success1
```

```
## [1] 828
```

```
p1
```

```
## [1] 0.4305772
```

```
s1
```

```
## [1] 0.2451805
```

```
n2 <- length(which(value$cow == "Romania"))
p2 <- mean(value$trust[which(value$cow == "Romania")])
s2 <-  p2 * (1-p2)
success2 <- length(which(value$trust == 1 & value$cow == "Romania"))

success2
```

```
## [1] 92
```

```r
p2
```

```
## [1] 0.07006855
```

```r
s2
```

```
## [1] 0.06515894
```

Test the hypothesis that the proportion of trust is eqyal in the two countries (two sided test)

```r
alpha = 0.05

prop.test(x = c(success1, success2),
          n = c(n1,n2),
          conf.level = 1 - alpha,
            correct = FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity correction
##
## data:  c(success1, success2) out of c(n1, n2)
## X-squared = 498.38, df = 1, p-value < 2.2e-16
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.3344239 0.3865935
## sample estimates:
##     prop 1     prop 2
## 0.43057722 0.07006855
```

Test the hypothesis that the difference in proportion of trust is greater in Germany (one sided test)

```r
alpha = 0.05


prop.test(x = c(success1, success2),
          n = c(n1,n2),
          alternative = "greater",
          conf.level = 1 - alpha,
          correct = FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity correction
##
## data:  c(success1, success2) out of c(n1, n2)
## X-squared = 498.38, df = 1, p-value < 2.2e-16
## alternative hypothesis: greater
## 95 percent confidence interval:
##  0.3386176 1.0000000
## sample estimates:
##     prop 1     prop 2
## 0.43057722 0.07006855
```

Test for variance (one sample)

We now focus on variances and we analyze the variance of the equality index. We test whether it is equal to 0.1 by using a chi-square test

```
varTest(as.numeric(value$equality),
        alternative = "two.sided",
        conf.level = 0.95,
        sigma.squared = 0.1
        )
```

```
##
## Results of Hypothesis Test
## --------------------------
##
## Null Hypothesis:                 variance = 0.1
##
## Alternative Hypothesis:          True variance is not equal to 0.1
##
## Test Name:                       Chi-Squared Test on Variance
##
## Estimated Parameter(s):          variance = 0.06175685
##
## Data:                            as.numeric(value$equality)
##
## Test Statistic:                  Chi-Squared = 3330.547
##
## Test Statistic Parameter:        df = 5393
##
## P-value:                         1.101022e-118
##
## 95% Confidence Interval:         LCL = 0.05949062
##                                  UCL = 0.06415574
```

Now we test whether it is greater than 0.1

```
varTest(as.numeric(value$equality),
        alternative = "greater",
        conf.level = 0.95,
        sigma.squared = 0.1
        )
```

```
##
## Results of Hypothesis Test
## --------------------------
##
## Null Hypothesis:                 variance = 0.1
##
## Alternative Hypothesis:          True variance is greater than 0.1
##
## Test Name:                       Chi-Squared Test on Variance
##
## Estimated Parameter(s):          variance = 0.06175685
##
## Data:                            as.numeric(value$equality)
##
## Test Statistic:                  Chi-Squared = 3330.547
##
## Test Statistic Parameter:        df = 5393
##
```

```
## P-value:                          1
##
## 95% Confidence Interval:          LCL = 0.05984857
##                                   UCL =        Inf
```

Test for variance (two samples

We want to compare the variance of the equality index in Germany and the variance of the equality test in Romania. We use an F test.

First, we test whether the variance of the equality in Germany is equal than the one estimated in Romania.

```
var.test(x = as.numeric(value$equality[which(value$cow == "Germany")]),
         y = as.numeric(value$equality[which(value$cow == "Romania")]),
         conf.level = 0.95,
         ratio = 1
        )
```

```
##
##  F test to compare two variances
##
## data:  as.numeric(value$equality[which(value$cow == "Germany")]) and as.numeric(value$equality[which
## F = 0.78274, num df = 1922, denom df = 1312, p-value = 1.112e-06
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.7084131 0.8640485
## sample estimates:
## ratio of variances
##          0.7827391
```

Now, we test whether it is smaller

```
var.test(x = as.numeric(value$equality[which(value$cow == "Germany")]),
         y = as.numeric(value$equality[which(value$cow == "Romania")]),
         conf.level = 0.95,
         ratio = 1,
         alternative = "less"
        )
```

```
##
##  F test to compare two variances
##
## data:  as.numeric(value$equality[which(value$cow == "Germany")]) and as.numeric(value$equality[which
## F = 0.78274, num df = 1922, denom df = 1312, p-value = 5.562e-07
## alternative hypothesis: true ratio of variances is less than 1
## 95 percent confidence interval:
##  0.0000000 0.8504372
## sample estimates:
## ratio of variances
##          0.7827391
```