Confidence intervals

SISS - Applied Statistics - Chiara Seghieri and Costanza Tortù

2023-11-13

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)

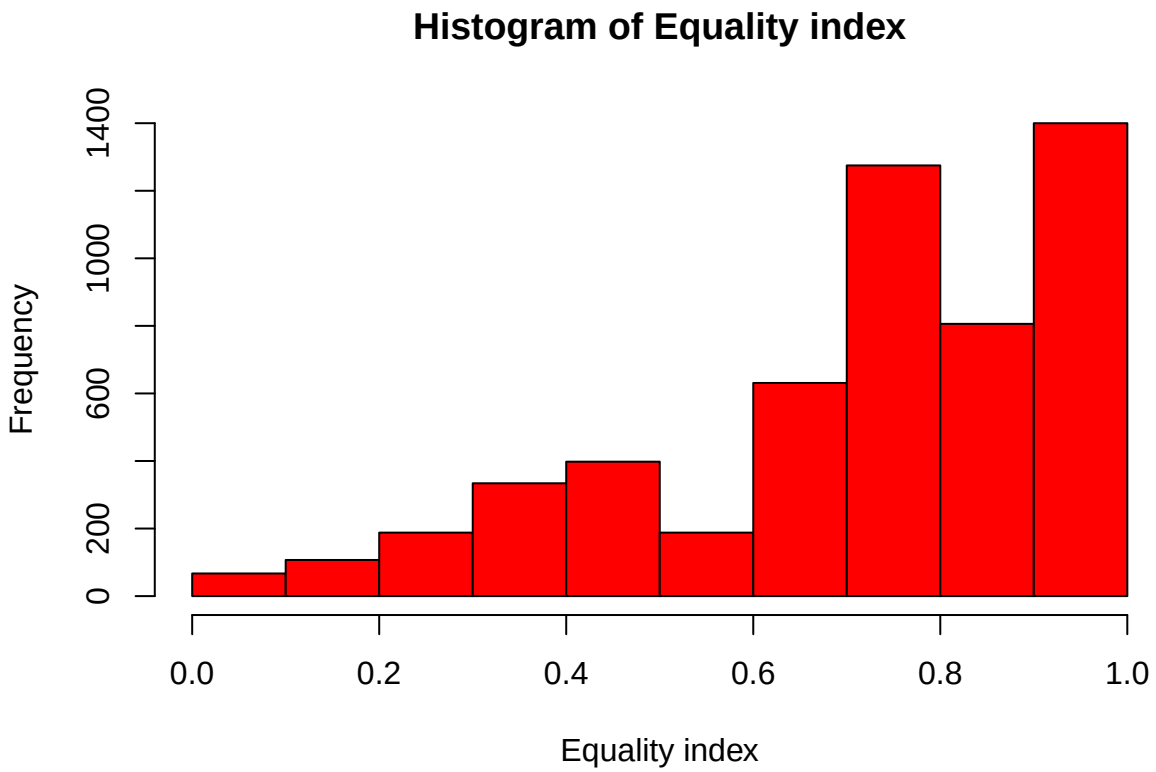
```
value <- value[complete.cases(value),]

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

```
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 equality index

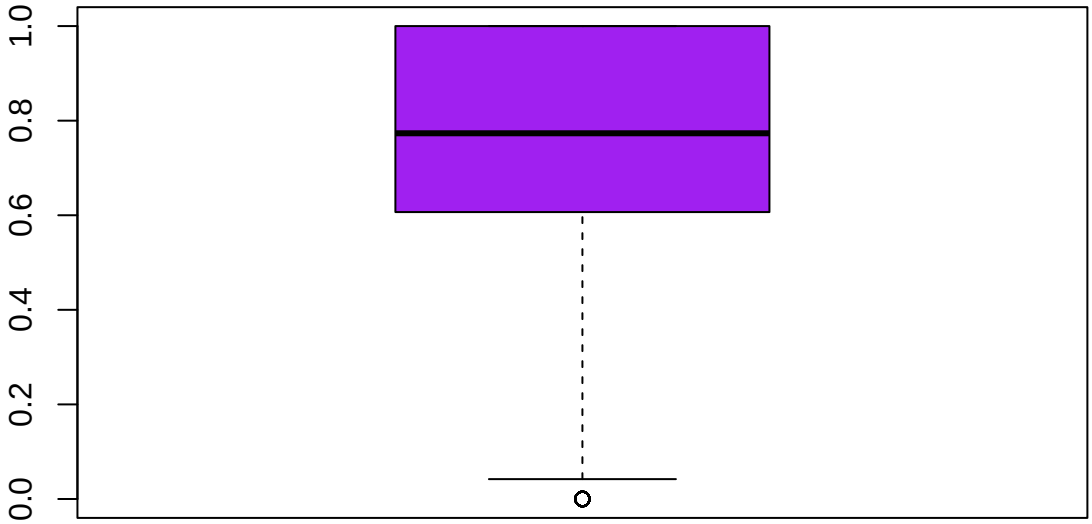
```
hist(value$equality,
     main ="Histogram of Equality index",
     col = "red",
     xlab = "Equality index")
```




```
summary(value$equality)
```

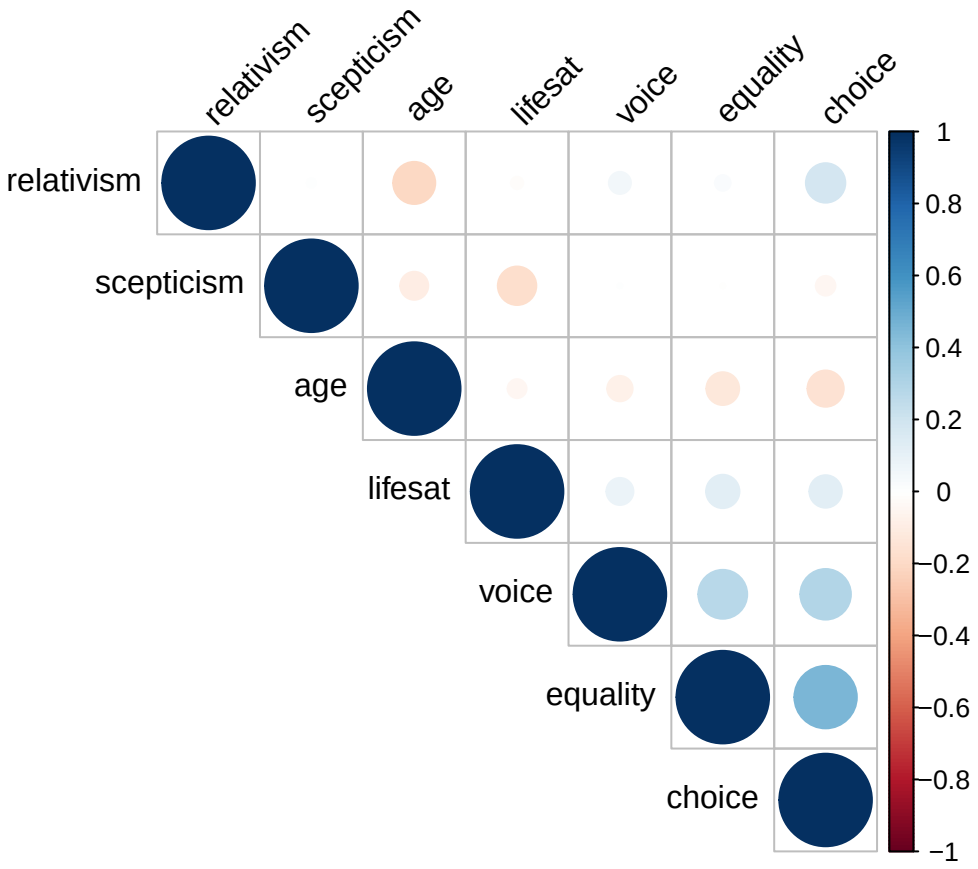
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.6067  0.7733  0.7327  1.0000  1.0000
```

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



Let's have a look at the correlation among continuous variables

```
correlation_matrix <- cor(value[, c(quantitative_variables)])

corrplot(correlation_matrix, type = "upper",
         order = "hclust",
         tl.col = "black", tl.srt = 45)
```



Compute confidence intervals for the average Equality index

Compute sample mean

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

```
## [1] 0.7327239
```

Compute sample variance

```
sample.n <- length(value$equality)
sample.sd <- sd(value$equality)
sample.se <- sample.sd/sqrt(sample.n)
print(sample.se)
```

```
## [1] 0.003383663
```

Find the t-score

```
alpha = 0.05
degrees.freedom = sample.n - 1
t.score = qt(p=alpha/2, df=degrees.freedom,lower.tail=F)
print(t.score)
```

```
## [1] 1.960404
```

Compute margin of error

```
margin.error <- t.score * sample.se
print(margin.error)
```

```
## [1] 0.006633347
```

Now we are ready to compute the confidence interval

```
lower.bound <- sample.mean - margin.error
upper.bound <- sample.mean + margin.error
print(c(lower.bound,upper.bound))
```

```
## [1] 0.7260906 0.7393573
```

Compute confidence intervals for the average Equality index in a subsample

Compute sample mean

```
  N_subsample <- 250
  sampled_units <- sample(value$ID, N_subsample, replace = F)
  value_subsample <- value[which(value$ID %in% sampled_units),]
```

```
  sample.mean <- mean(value_subsample$equality)
  print(sample.mean)
```

```
## [1] 0.7527149
```

Compute sample variance

```
sample.n <- length(value_subsample$equality)
sample.sd <- sd(value_subsample$equality)
sample.se <- sample.sd/sqrt(sample.n)
print(sample.se)
```

```
## [1] 0.01387489
```

Find the t-score

```
alpha = 0.05
degrees.freedom = sample.n - 1
t.score = qt(p=alpha/2, df=degrees.freedom,lower.tail=F)
print(t.score)
```

```
## [1] 1.969537
```

Compute margin of error

```
margin.error <- t.score * sample.se
print(margin.error)
```

```
## [1] 0.02732711
```

Now we are ready to compute the confidence interval

```
lower.bound <- sample.mean - margin.error
upper.bound <- sample.mean + margin.error
print(c(lower.bound,upper.bound))
```

```
## [1] 0.7253878 0.7800420
```

Compute confidence intervals for the variance of Equality index

Compute sample variance

```
sample.n <- length(value$equality)
sample.var <- var(value$equality)
print(sample.var)
```

```
## [1] 0.06175685
```

Find the chi-scores

```
alpha = 0.05
degrees.freedom = sample.n - 1
chi.scores = qchisq(c(1-alpha/2, alpha/2), df =  degrees.freedom)
print(chi.scores)
```

```
## [1] 5598.441 5191.348
```

Now we are ready to compute the confidence interval

```
lower.bound <- degrees.freedom*sample.var/chi.scores[1]
upper.bound <- degrees.freedom*sample.var/chi.scores[2]

print(c(lower.bound,upper.bound))
```

```
## [1] 0.05949062 0.06415574
```

Confidence intervals for the difference in means

Let's focus on two countries: Germany and Romania. We want to build up a confidence interval for the difference im the mean equality index in the two countries.

1) Compute all the quantities you need

```
n1 <- length(which(value$cow == "Germany"))
xbar1 <- mean(value$equality[which(value$cow == "Germany")])
s1 <- var(value$equality[which(value$cow == "Germany")])
n2 <- length(which(value$cow == "Romania"))
xbar2 <- mean(value$equality[which(value$cow == "Romania")])
s2 <-  var(value$equality[which(value$cow == "Romania")])
```

2) Compute pooled variance

```
sp = ((n1-1)*s1^2+(n2-1)*s2^2)/(n1+n2-2)
sp
```

```
## [1] 0.004205391
```

3) Compute the margin of error

```
margin <- qt(1-alpha/2,df=n1+n2-1)*sqrt(sp/n1 + sp/n2)
margin
```

```
## [1] 0.004551935
```

4) Compute the confidence interval

```
lowerinterval <- (xbar1-xbar2) - margin
lowerinterval
```

```
## [1] 0.1726434
```

```
upperinterval <- (xbar1-xbar2) + margin
upperinterval
```

```
## [1] 0.1817473
```

Confidence intervals for the difference in proportions

Let's focus on two countries: Germany and Romania. We want to build up a confidence interval for the difference im the proportion of people who trust in people in the two countries.

1) Compute all the quantities you need

```
n1 <- length(which(value$cow == "Germany"))
p1 <- mean(value$trust[which(value$cow == "Germany")])

n2 <- length(which(value$cow == "Romania"))
p2 <- mean(value$trust[which(value$cow == "Romania")])
```

2) Compute the margin of error

```
margin <- qnorm(1-alpha/2)*sqrt(p1*(1-p1)/n1 + p2*(1-p2)/n2)
margin
```

```
## [1] 0.02608483
```

3) Compute the confidence interval

```
lowerinterval <- (p1-p2) - margin
lowerinterval
```

```
## [1] 0.3344239
```

```
upperinterval <- (p1-p2) + margin
upperinterval
```

```
## [1] 0.3865935
```