DASS: Confidence intervals

Alla Tambovtseva

Confidence intervals

First, let's install `DescTools` library that is used for calculating descriptive statistics and confidence intervals:

```
install.packages("DescTools")
```

Load this library:

```
library(DescTools)
```

Confidence interval for a proportion

Now consider the following example. We asked 100 people and found that 30 people supported the capital punishment, so $n = 100$, $p = 0.3$. We have to calculate 90% confidence interval for the proportion of people that approve of the capital punishment.

```
# first goes the number of people we are interested in,
# then goes the total number of people
# and then we specify the confidence level

ci90 <- BinomCI(30, 100, conf.level = 0.90)
ci90
```

```
##      est   lwr.ci    upr.ci
## [1,] 0.3 0.230705 0.3798321
```

The function `BinomCI()` returns three numbers: a sample proportion, a lower of a confidence interval and an upper one. Let's interpret the results obtained.

- We are 90% confident that the true proportion of people supporting capital punishment lies between 0.23 and 0.38. If we repeat the same research on the samples of size 100 many times, 90% of condidence intervals will include the true proportion of people supporting capital punishment.

Now let's calculate 95% and 99% confidence intervals:

```
ci95 <- BinomCI(30, 100, conf.level = 0.95)
ci95
```

```
##      est    lwr.ci    upr.ci
## [1,] 0.3 0.2189489 0.3958485
```

```
ci99 <- BinomCI(30, 100, conf.level = 0.99)
ci99
```

```
##      est    lwr.ci    upr.ci
## [1,] 0.3 0.1974607 0.4274276
```

Calculate lengths of each interval and compare:

```
l90 <- ci90[3] - ci90[2]
l90
```

```
## [1] 0.1491272
```

```
l95 <- ci95[3] - ci95[2]
l95
```

```
## [1] 0.1768997
```

```
l99 <- ci99[3] - ci99[2]
l99
```

```
## [1] 0.229967
```

As expected, the higher is the sample size, the narrower a confidence interval is (so, we get more precise, less dispersed results).

Imagine that now we asked 200 people and found that 60 people approved of the capital punishment, so $n = 200$, $p = 0.3$. Let's calculate 90% confidence interval:

```
ci90n <- BinomCI(60, 200, conf.level = 0.90)
ci90n
```

```
##      est    lwr.ci    upr.ci
## [1,] 0.3 0.2496597 0.3556791
```

Calculate its length and compare with 90% confidence inteval for $n = 100$:

```
l90n <- ci90n[3] - ci90n[2]
l90n
```

```
## [1] 0.1060194
```

```
l90
```

```
## [1] 0.1491272
```

Now let's work with real data. We will work with data on the Chilean plebiscite we discussed before.

```
# load data and delete rows with NA's
df <- read.csv("http://math-info.hse.ru/f/2017-18/ps-ms/Chile.csv")
df <- na.omit(df)
```

Look at unique values of vote:

```
table(df$vote)
```

```
##
##   A   N   U   Y
## 177 867 551 836
```

868 people intended to vote for Pinochet being in rule, 889 people – against him.

```
yes <- 868
no <- 889
```

Now let's calculate 95% confidence interval for the proportion of people who supported Pinochet.

```
BinomCI(yes, yes + no, conf.level = 0.95)
```

```
##            est    lwr.ci    upr.ci
## [1,] 0.4940239 0.4706848 0.5173891
```

- We are 95% confident that the proportion of people who supported Pinochet is between and 0.47 and 0.52. If we repeat the research many times on the samples of the same size, 95% of confidence intervals will cover the true proportion of Pinochet's supporters.

Confidence interval for a mean

Choose people who voted for Pinochet staying in rule:

```
forP <- df[df$vote == "Y", ]
```

Calculate 95% confidence interval for the mean age of these people:

```
MeanCI(forP$age)
```

```
##     mean   lwr.ci   upr.ci
## 40.20335 39.17485 41.23185
```

Interpretation: We are 95% confident that the average age of people that supported Pinochet is between 39 and 41. If we repeat the research many times on the samples of the same size, 95% of calculated confidence intervals will include the true mean age of Pinochet's supporters.

Note: if the confidence level is 0.95, we can skip the option `conf.level` as this value is set by default in R.

Confidence interval and testing hypotheses

Although we have not discussed hypotheses testing, we can overview the idea of making conclusions based on confidence intervals. Suppose that two 95% confidence intervals for population means intersect. What does it mean? It means that population means are likely to coincide! Thus, we cannot say that population means are significantly different!

Let's compare two 95% confidence intervals: the first interval is for the mean age of people supporting Pinochet and the second one is for the mean age of people against Pinochet.

```
forP <- df[df$vote == "Y", ]
agP <- df[df$vote == "N", ]

MeanCI(forP$age)
```

```
##     mean   lwr.ci   upr.ci
## 40.20335 39.17485 41.23185
```

```
MeanCI(agP$age)
```

```
##     mean   lwr.ci   upr.ci
## 35.99885 35.04379 36.95390
```

Two confidence intervals do not intesect, so the true mean age of people who supported Pinochet is not likely to coincide with the true mean age of people who did not. Thus, we can be 95% confident that on average people who tend to vote for and against Pinochet were of the different age. From the confidence intervals, we see that on average supporters of Pinochet are older.