Chapter 6 - Confidence Intervals

Dr. Alessandro Ruggieri

Contents

Intro

Consider data of GDP per capita across countries:

```r
# import data on GDP per capita across countries
data_gdp <- read.csv("GDP_205U_NOC_NB_A-filtered-2020-12-22.csv",  header = TRUE)
# rename columns
names(data_gdp) <- c("country", "label", "source", "year", "gdp")
## Select 2019 data
data_gdp_2019<-data_gdp[data_gdp$year == "2019",]
## Final data
x <- data_gdp_2019$gdp
```

We want to find a 90% confidence interval for the mean of GDP per capita.

Case 1 - Known distribution

Assume that the GDP per capita is normally distributed with unknown mean and unknown standard deviation.

Since we don't know the value for the mean and standard deviation in the population, we replace them with an estimates:

```r
# Sample mean
xbar<-mean(x)
```

and

```r
# Sample standard deviation
sigma<-sd(x)
```

In this case, a (1-α)*100% confidence interval is defined as $[\bar{x} - t_{\frac{\alpha}{2},d_f}\frac{\sigma}{\sqrt{n}}, \bar{x} + t_{\frac{\alpha}{2},d_f}\frac{\sigma}{\sqrt{n}}]$

The value for α is 0.10.

```r
# confidence level
alfa<-0.10
```

and the sample size n can be constructed as follows:

```r
# sample size
n<-length(x)
print(n)
```

```
## [1] 277
```

All we need now is to find a t-value. To do so, we need to compute the degrees of freedom.

```r
# degrees of freedom
d_f<-n-1
print(d_f)
```

```
## [1] 276
```

Therefore, we can get $t_{\frac{\alpha}{2},d_f}$ as follows:

```r
# t-value
t <- qt( (1-alfa)/2, d_f)
print(t)
```

```
## [1] -0.125777
```

So the margin of error is equal to

```r
# margin of error
me <- t*sigma/sqrt(n)
print(me)
```

```
## [1] -279.8765
```

Therefore the lower and the upper bound of the confidence interval at 90% is are equal to

```r
# lower  bound
lb<-xbar - me
print(lb)
```

```
## [1] 31588.85
```

```r
# upper bound
ub<-xbar + me
print(ub)
```

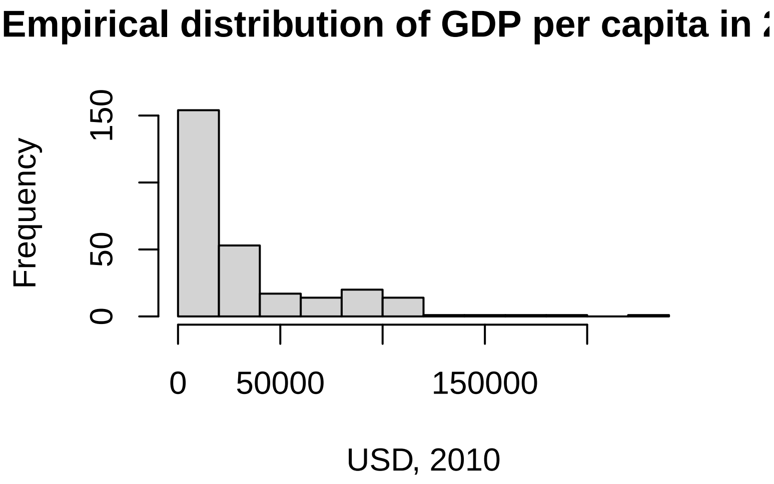
```
## [1] 31029.1
```

Case 2 - Unknown Distribution

Assume that distribution of theGDP per capita is unknown

Now we no longer assume the data are normal. Note that a look at the histogram plot show that this was not such a great assumption to begin with.

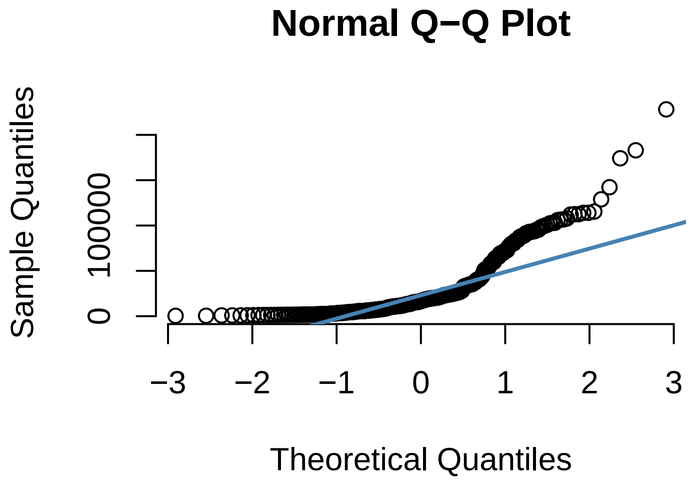
```r
# histogram of data
hist(x, xlab="USD, 2010", main="Empirical distribution of GDP per capita in 2019")
```



Another way to check the normality of the data is to produce a quantile-to-quantile (QQ) plot. The Q–Qplot is allows to compare the sample distributions x to the normal distribution, by plotting their quantiles against each other.

The function qqnorm() produces a normal QQ plot, which draws the correlation between the sample x and the normal distribution. Whereas the function qqline() adds a 45 degree reference line. If normally distributed, the sample quantile of x would lie on the 45 degree line.

```
# Sample quantile-to-quantile plot
qqnorm(x, pch = 1, frame = FALSE)
qqline(x, col = "steelblue", lwd = 2)
```



The QQ plot confirms the sample data is not normally distributed. So at best, the confidence intervals from above are approximate. The approximation, however, might not be very good. A bootstrapped confidence interval might be helpful. Here are the steps involved.

- From the sample, draw a new sample, WITH replacement, of size n
- Calculate the sample average, called the bootstrap estimate.
- Store it
- Repeat steps 1-3 3 many times. (We'll do 1000).

```
maxi<-1000
bstrap <- c()
for (i in 1:maxi){
# First take the sample
bsample <- sample(x,n,replace=TRUE)
# now calculate the bootstrap estimate
bestimate <- mean(bsample)
bstrap <- c(bstrap,bestimate)
}
```

- To construct a 90% confidence interval, we will use the 5% sample quantile as the lower bound, and the 95% sample quantile as the upper bound.

```
# lower bound
lb<-quantile(bstrap,(alfa)/2)
print(lb)
```

```
##       5%
## 27788.79
```

```
# upper bound
ub<-quantile(bstrap,(1-alfa)/2)
print(ub)
```

```
##      45%
## 31082.33
```

This method of using the sample quantiles to find the bootstrap confidence interval is called the Percentile Method.