Confidence intervals

Eduardo Alarcón & Alfonso Pineda

2022-12-03

2 Example 1: Web-page accessing times

We want to construct the confidence intervals for the mean, μ, and for the variance, σ^2, of the distribution of the accessing times to a web page of UC3M from a specific computer at home as well as from a computer in the UC3M campus. The confidence intervals will be constructed by using 55 observations (in seconds). Each observation consists of two accessing times, one measured on a home computer and one on a computer belonging to the university campus (file `TiempoaccesoWeb.xlsx`)

First we read and view the data file. The figure shows the first five observations of this datafile.

```
library(readxl)
SpotifySongs <- read_excel("songstats.xlsx")
```

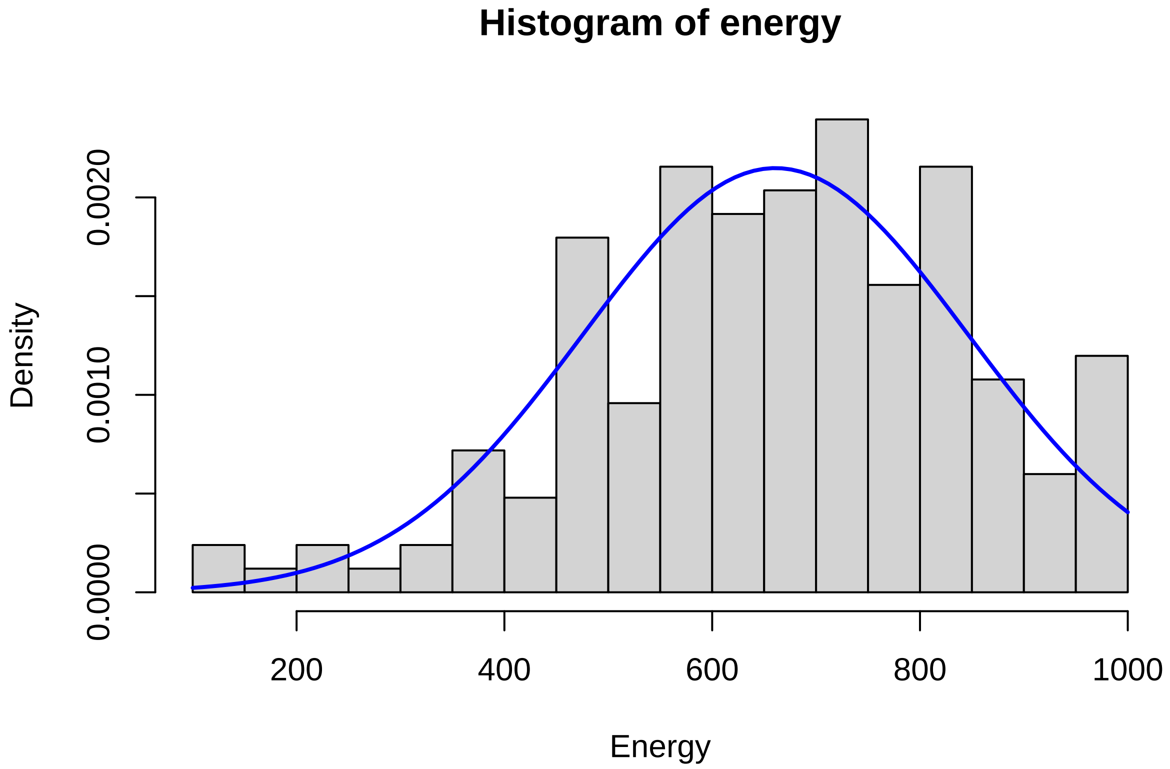
2.1. Univariate analysis of data

Before doing any kind of analysis it is useful to first describe the variables of interest. We start with the access times of the computer at home (variable `energy`). The numerical and graphical analysis can be performed by

```
suppressWarnings(library(summarytools))
energy <- SpotifySongs$energy
descr(energy)
```

```
## Descriptive Statistics
## energy
## N: 167
##
##                       energy
## ----------------- --------
##              Mean   660.92
##           Std.Dev   185.68
##               Min   104.00
##                Q1   551.00
##            Median   667.00
##                Q3   804.00
##               Max   993.00
##               MAD   188.29
##               IQR   249.50
##                CV     0.28
##          Skewness    -0.44
##       SE.Skewness     0.19
##          Kurtosis     0.07
##           N.Valid   167.00
##         Pct.Valid   100.00
```

```
hist(energy, breaks = seq(100, 1000, 50),
     probability = TRUE, # histogram has a total area = 1
     xlab = "Energy")
curve(dnorm(x, mean(energy), sd(energy)),
      col="blue", lwd=2, add=TRUE, yaxt="n")
```



We can notice in the figure that the variable `energy` has a Normal-liked distribution: it is quite symmetric and bell-shaped. The hypothesis of normality is important to compute the confidence intervals. For example to construct a confidence interval for the variance it is *mandatory* to assume the normality since only in that case we know that the estimator is distributed as a Chi-squared distribution.

The summary statistics include measures of central tendency, measures of variability and measures of shape, we can notice that the values of the Skewness and the Kurtosis are quite small confirming the fact that the histogram looks like a Normal distribution.

Among these values, the sample mean and variance are the "point" estimations of the population mean and variance. That is, we have that in this sample, the "point" estimation of the parameters of interest are $\widehat{\mu} = 660.92$ and $\widehat{\sigma}^2 = 185.68^2$.

Our objective is to make an "interval" estimation of these parameters.

2.2 Analysis of the Normality of the variable

We perform a goodness-of-fit test to check if the variable can be assumed normal. For simplicity, we will use the normality tests provided by package `nortest`

```
library(nortest)
ad.test(energy)
```

```
##
##  Anderson-Darling normality test
##
```

```
## data:  energy
## A = 0.46126, p-value = 0.2563
```

```
cvm.test(energy)
```

```
##
##  Cramer-von Mises normality test
##
## data:  energy
## W = 0.053358, p-value = 0.4607
```

```
lillie.test(energy)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  energy
## D = 0.040216, p-value = 0.7325
```

```
pearson.test(energy)
```

```
##
##  Pearson chi-square normality test
##
## data:  energy
## P = 10.725, p-value = 0.6339
```

```
sf.test(energy)
```

```
##
##  Shapiro-Francia normality test
##
## data:  energy
## W = 0.98224, p-value = 0.03169
```

2.3 Confidence intervals

To obtain the confidence intervals for the mean, μ, and the variance, σ^2, we evaluate the following expressions:

- Confidence interval for μ with known variance: $\bar{x} \mp z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$, where n is the sample size, σ is the known standard deviation, $z_{\alpha/2}$ is the $(\alpha/2)$-percentil of the standard normal distribution (`qnorm(alpha/2)` in R) and $\bar{x}$ is the sample mean (`mean(x)` in R).

In the example

```
alpha = 0.05
n = length(energy)
xmean = mean(energy)
xsd = sd(energy)
z = qnorm(alpha/2, lower.tail = FALSE)
LowerLimit = xmean - z * xsd / sqrt(n)
UpperLimit = xmean + z * xsd / sqrt(n)
LowerLimit
```

```
## [1] 632.7611
```

```
UpperLimit
```

```
## [1] 689.0832
```

where we assume that σ is known and equals to `sd(TiempoAccesoWeb$Ordenador_Casa)`. $\mu \in (632.7611, 689.0832)$

The above confidence intervals can be obtained as output of some functions implementing hyphotesis testing: * Confidence interval for μ with known variance using `z.test`:

```
suppressWarnings(library(BSDA))
z.test(energy, sigma.x = xsd)$conf.int
```

```
## [1] 632.7611 689.0832
## attr(,"conf.level")
## [1] 0.95
```