# Probabilistic analysis

X. Pouwels

2022-03-07

## Contents



## Install and load packages

First, the required packages need to be installed and loaded.

```
rm(list = ls()) # clear environment

# Install
#install.packages("fitdistrplus")
#install.packages("mvtnorm")

# Load
library(fitdistrplus)
library(mvtnorm)
library(stats)
library(doSNOW)
library(parallel)
library(ggplot2)
```

## Deterministic analysis

Imagine you are interested in calculating the average time a group of student takes to get home from the UT by bicycle, bus, and train. Based on the measurement of travel time (and train delay) in 100 students on the 1st of February 2020, you obtain the following statistics (times are displayed in minutes) :

| Statistic | Time bicycle > bus | Time bus > trainstation | Chance train is delayed | Time in train (no delay) | Time in train (with delay) | Total time |
|---|---|---|---|---|---|---|
| Mean | 15 | 20 | 0.15 | 30 | 60 | 69.5 |
| SD | 2 | Range: 10-30 | 0.035 | 1.73 | 7.73 | |
| Distribution | Normal | Uniform | Beta | Gamma | Gamma | |

The mean travel time is 69.5 minutes [calculated as: Time bicycle > bus + Time bus > trainstation + Chance train is delayed * Time in train (with delay) + (1-Chance train is delayed) * Time in train (no delay)]. However, this number may not represent the true average travel time because none of the times and chances elicited in this survey are fixed. These parameters are random variables which are, by definition, uncertain. This uncertainty (and its consequences) can be assessed though probabilistic analyses (PA).

## Probabilistic analysis

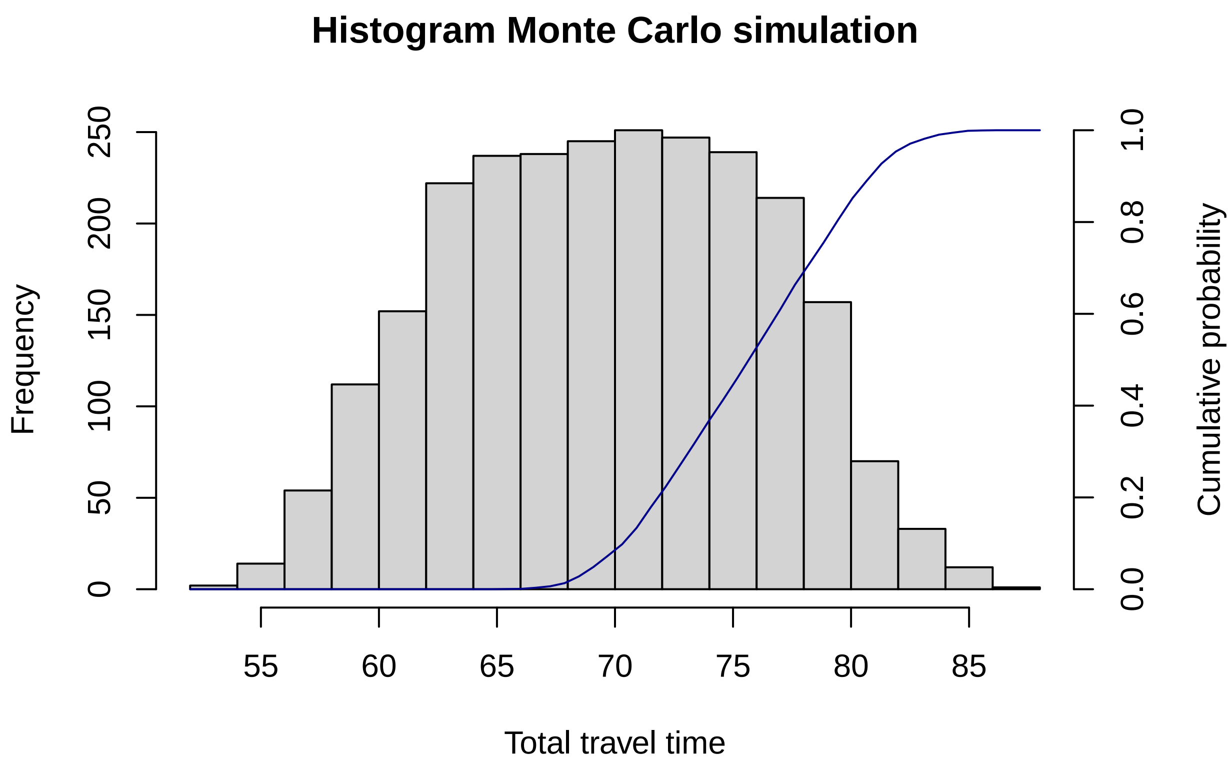
In health economic modelling, PA are mostly performed through Monte Carlo sampling (or simulation). During a Monte Carlo simulation, random values are drawn from the variables used to calculate the (health economic) outcomes. These random values are then used to recalculate the outcomes of the model. This process is repeated a large amount of time (at least 1,000 times) to compute a set of potential outcomes. This set of outcome is used to assess the uncertainty (and its consequences) surrounding the results of the analysis.
Now, let's get back to our 'travel time' example. As you can see, each input parameter is characterised by a mean value (that was used to calculate the deterministic output), a standard deviation (SD), and a type of distribution has been assigned to it. By using the method of moments, these parameters can be used to obtain the parameters of distributions representing each random variable.

```
n_it <- 2500

v_bitobus <- rnorm(n_it, mean = 15, sd = 2) # draw 2,500 random values from a normal distribution for t
v_bustotr <- runif(n_it, min = 10, max = 30) # draw 2,500 random values from a uniform distribution for
v_delay   <- rbeta(n_it, shape1 = 15, shape2 = 85) # draw 2,500 random values from a beta distribution
v_trainnod <- rgamma(n_it, shape = 300, rate = 10) # draw 2,500 random values from a gamma distribution
v_trainwd  <- rgamma(n_it, shape = 60, rate = 1) # draw 2,500 random values from a gamma distribution f
v_total <- v_bitobus + v_bustotr + v_delay * v_trainwd + (1 - v_delay) * v_trainnod # calculate total t
```

To investigate whether the random values represent adequately the random variable, we can compare their means and standard deviation.

| Statistic | Time bicycle > bus | Time bus > trainstation | Chance train is delayed | Time in train (no delay) | Time in train (with delay) | Total time |
|---|---|---|---|---|---|---|
| Mean | 15 | 20 | 0.15 | 30 | 60 | 69.5 |
| SD | 2 | 5.78 | 0.035 | 1.73 | 7.73 | |
| Distribution | Normal | Uniform | Beta | Gamma | Gamma | |
| PA Mean | 14.93 | 20.21 | 0.15 | 30.06 | 59.94 | 69.69 |
| PA SD | 1.989 | 5.804 | 0.036 | 1.745 | 7.795 | 6.489 |

```r
histo <- hist(v_total, breaks = seq(from = 30, to = 90, by = 1), plot = FALSE) # register histogram val

#create plot with histogram and cumulative probability
par(mar = c(5, 4, 4, 4) + 0.3)
hist(v_total, main = "Histogram Monte Carlo simulation", xlab = "Total travel time")
par(new = TRUE)
plot(x = histo$mids, y = cumsum(histo$counts) / sum(histo$counts), type = 'l', col = 'darkblue',
     axes = FALSE, xlab = "", ylab = "")# Create second plot without axes
axis(side = 4, at = pretty(c(0:1)))      # Add second axis
mtext("Cumulative probability", side = 4, line = 3)
```




We can also investigate the distribution of the total travel time by visualising the histogram and cumulative probability of each outcome. Based on this information, we are able to calculate the probability that a student would be home within or after a certain travel time, for instance. The table below describes the probability that the travel time of students is below 60 minutes, between 64 and 72 minutes, and above 76 minutes. These probabilities can be calculated for each potential value of the travel time.

| Interval | Probability |
|---|---|
| t < 60 | 7% |
| 64 > t < 72 | 54% |
| t > 76 | 19% |