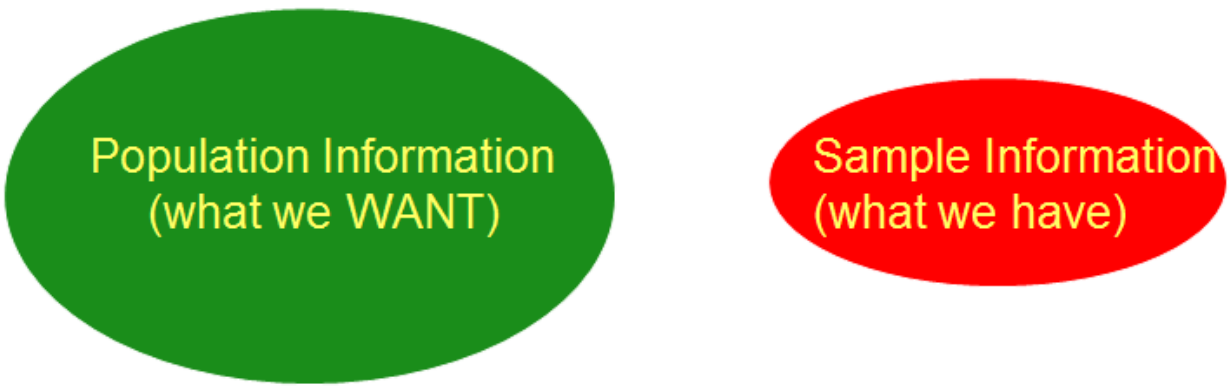
ADVANCED BIOSTATISTICS ABSTAT17

Simulation Modeling Methodologies

Carina Silva
(*carina.silva@estesl.ipl.pt*)
Higher School of Technologies and Health of Lisbon &
Center of Statistics and Applications, University of Lisbon

IGC, April, 10–13, 2017

Much of what we do in statistics involves trying to talk about true characteristics of a process, using an imperfect subset of information from the process.



- Simulations allow doing statistical analysis without making strong parametric assumptions.

All models are wrong, some are useful. George Box

- In general the term simulation applies when reproducing the behavior of a real problem, translated by a physical system, in a controlled environment.
- When this behavior is of a random nature, that is to say, there is no deterministic expression that describes it, this is a problem of stochastic simulation.
- Assuming then that the behavior of the physical system can be described by a probability density function, the simulation method requires the possibility of generating random numbers with this distribution.

- It is possible to simulate any distribution from the simulation of uniform random numbers in $[0, 1]$.
- Thus, the starting point for stochastic simulations is the construction of a random number generator with uniform distribution in $[0, 1]$.

Motivation

- Small samples that are NOT from a normal distribution
- In the old days: non-parametric tests
- More common now: Simulation based statistics:
  - Confidence intervals are much easier to achieve
  - They are much easier to apply in more complicated situations
  - They better reflect today's reality: they are simply now used in many contexts
  - Require: Use of computer - `R` is a super tool for this!

What is a simulation study

Simulation: A numerical technique for conducting experiments on the computer

Monte Carlo simulation: Computer experiment involving random sampling from probability distributions

- Invaluable in statistics
- Usually, when statisticians talk about *simulations*, they mean *Monte Carlo simulations*

What is a simulation?

- (Pseudo) random numbers generated from a computer, since the method is completely deterministic. Thus, this numbers have a similar behaviour from the random numbers.
- A random number generator is an algorithm that can generate x_{i+1} from x_i
- Require a "start" called "seed", i.e., a number that initiates the deterministic/iterative process.
- The "seed" associated to a generator method (algorithm), always produce the same sequence.

What is a simulation?

- This is an important characteristic, because that gives to the user the possibility to reproduce exactly the same results.
- Basically the uniform distribution is simulated in this way.

- Exercise 1: Generate random numbers from an Uniform Distribution in R with different size samples and construct histograms and analyze them.

Simulation of random discrete quantities

A quantity simulated by an uniform distribution (0,1) will be denoted by u and $u_1, u_2, ..., u_n$ will represent a sequence of quantities of that nature. In fact, it ill be a pseudo-random numbers sequence.

- Let's suppose X represents a discrete r.v. taking different values $1, 2, ..., k$ with probabilities $p_1, p_2, ..., p_k$, where:

$$P(X = i) = p_i, i = 1, ..., k, \sum_{i=1}^{k} p_i = 1.$$

Algorithm to generate random discrete quantities

- Algorithm to simulate X where $P(X = i) = p_i, i = 1, ..., k$
  1. Split the interval $[0, 1]$ in k subintervals $I_1, I_2, ..., I_k$ with $I_i = (F_{i-1}, F_i]$, where $F_0 = 0$ and $F_i = p_1 + p_2 + ... + p_i$, $i = 1, ..., k$;
  2. Generate a random number u in [0,1]. It will belong in just one of those subintervals.
  3. If $u \in I_i$, then it will returns $x = i$.
- It's easy to demonstrate that numbers generated like this will have the wanted distribution. In fact,
  1. Accordingly to the algorithm, $P(X = i) = P(u \in I_i)$
  2. How $U(0, 1), P(U \in I_i) = F_i - F_{i-1} = p_i$
  3. $\forall i = 1, ..., k, P(X = x_i) = p_i$

Exercise 2

- There are 4 blood groups, namely A, B, AB and 0. Suppose that in one particular population the alleles A, B and 0 have the following probabilities:
  $p_A^* = 0.21, p_B^* = 0.08, p_0^* = 0.71$
- Accordingly to the genetic model, the probabilities of each of the blood groups will be:
  $p_A = p_A^*(2 - p_A^* - 2p_B^*)$, $p_B = p_B^*(2 - p_B^* - 2p_A^*)$
  $p_{AB} = 2p_A^* p_B^*$, $p_0 = (1 - p_A^* - p_B^*)^2$.
- Consider: $A \Longleftrightarrow 1$, $B \Longleftrightarrow 2$, $AB \Longleftrightarrow 3$, $0 \Longleftrightarrow 4$.

Exercise 2

Simulate $n = 2128$ pseudo-random numbers $U(0, 1)$ using `runinf(n)` and the previous algorithm 5 times. Compare the results with the original data: A=725, B=258, AB=72, 0=1073.

Exercise 3

Generate one sample of size 200 from each of the discrete Binomial and Poisson distributions. Compare the distributions with suitable plots. Change the parameters in each of the models to see how they influence the distributions.

Exercise 4

The Binomial distribution is a possible probability model for the number of stormy days in a season. Do you think that this is a realistic model? If there are 120 days in a winter and the probability of a stormy day in winter is 1/3, write down the parameters, n and p, of the Binomial model. Compute the expectation and variance of the number of stormy days, and compute the probability of a winter having more than 40 stormy days. Generate 100 Binomial variables to represent a sequence of 100 winters and plot the simulated data. What is the average number of stormy days simulated?

Exercise 5

The distribution of the expression values of the ALL patients on the Zyxin gene are distributed according to N(1.6; 0.42).
a) Compute the probability that the expression values are smaller than 1.2?
b) What is the probability that the expression values are between 1.2and 2.0?
c) What is the 15th quantile of that distribution. What does it mean?
d) Use rnorm to draw a sample of size 1000 from the population and compare the sample mean and standard deviation with that of the population.