UTADEO
UNIVERSIDAD DE BOGOTÁ JORGE TADEO LOZANO

Simulación Estocástica 2025-1

Taller 2: Análisis estadístico de datos simulados, intervalos de confianza, bootstrap, modelos y simulación.

Javier Riascos Ochoa, PhD

Indicaciones: Por grupo, subir al link en AVATA **un (1)** archivo Rmarkdown o Notebook de Python con el procedimiento analítico, códigos (de los puntos que lo requieran), resultados y análisis, y **un (1)** archivo pdf con la salida impresa.

Fecha de entrega: Jueves 10 de Abril de 2025

1. **Stopping generating new simulation data (1 point)** *(Ross, Simulation)*

   Write a program to generate standard normal random variables until you have generated n of them, where $n \geq 100$ is such that $S/\sqrt{n} < 0.01$, where S is the sample standard deviation of the n data values. Note that this is the "Method for Determining When to Stop Generating New Data". Also, answer the following questions:

   a) How many normals do you think will be generated? Give an analytic estimate.

   b) How many normals did you generate?

   c) What is the sample mean of all the normals generated?

   d) What is the sample variance?

   e) Comment on the results of (c) and (d). Were they surprising?

2. **Gaining confidence with confidence intervals (1 point)** *(Jones, et al.)*

   We know that the $U(-1, 1)$ r.v. has mean 0. Use a sample of size 1000 to estimate the mean and give a 95 % confidence interval (CI). Does the CI contain 0? Repeat the above a large number of times (≥ 100). What percentage of time does the CI contain 0? Write your code so that it produces output similar to the following:

```
Number of trials: 10

Sample mean  lower bound  upper bound  contains mean?

    -0.0733      -0.1888       0.0422               1
    -0.0267      -0.1335       0.0801               1
    -0.0063      -0.1143       0.1017               1
    -0.0820      -0.1869       0.0230               1
    -0.0354      -0.1478       0.0771               1
    -0.0751      -0.1863       0.0362               1
    -0.0742      -0.1923       0.0440               1
     0.0071      -0.1011       0.1153               1
     0.0772      -0.0322       0.1867               1
    -0.0243      -0.1370       0.0885               1

100 percent of CI's contained the mean
```

3. **Standard deviation of a proportion (1 points)** *(Jones, et al.; Chapter 18)*

   Assume a manager is using the sample proportion $\hat{p}$ to estimate the proportion p of a new shipment of computer chips that are defective. He doesn't know p for this shipment, but in previous shipments it has been close to 0.01, that is 1 % of chips have been defective.

   a) If the manager wants the standard deviation of $\hat{p}$ to be about 0.02, how large a sample should she take based on the assumption that the rate of defectives has not changed dramatically?

   b) Now suppose something went wrong with the production run and the actual proportion of defectives in the shipment is 0.3, that is 30 % are defective. Now what would be the actual standard deviation of $\hat{p}$ for the sample size you choose in *a)*?

4. **Bets (2 points)**

   You pay 10.000 pesos to participate in a bet game, which consists in tossing two coins together. If two heads fall, you earn 15.000 pesos. If one head and one tail fall you earn 5.000 pesos. In any other case you earn nothing. Let X the random variable of your profit.

   a) Analytically find the probability mass function p_X, the mean $E[X]$, and variance $Var(X)$ of X.

   b) Write a code that simulates the r.v. X using the command *sample*. Generate an *iid* sample $\{X_i\}$ of size $n = 10^5$.

   c) Modify your code to calculate: (i) the estimated mean profit $\overline{X}_j$ for each sample subsequence: $\{X_1, X_2, \ldots, X_j\}, j = 2, 3, \ldots, n$, (ii) the 95 % CI's of each estimated mean profit $\overline{X}_j$.

   d) Plot $\overline{X}_j$ and their 95 % CI's in terms of $j = 2, \ldots, 10^5$. Add an horizontal line corresponding to the actual value $E[X]$.

   e) Repeat *c)* and *d)* to estimate the probabilities $p_X(x)$, their 95 % CI's, and their plots for each $j = 2, \ldots, 10^5$, adding the actual values.

5. **Bootstrap (2 points)**

   This exercise is based on the article Introduction to Bootstrapping in Statistics with an Example and the dataset *body_fat.csv* that contains the body fat percentages of 92 adolescent girls. Generate a program that gives:

   a) An histogram of the sample data.

   b) The 95 % confidence interval of the mean of the data from the traditional method (i.e., via the Central Limit Theorem).

   c) A number of 500 bootstrapped samples from the original dataset, with 92 observations each.

   d) An histogram of the means of each bootstrapped sample.

   e) A 95 % bootstrapped confidence interval of the mean.

   f) A comparison of both confidence intervals.

6. **Reliability of a system (2 points)** *(Ross, Simulation; Ross, Introduction to Probability Models)*

   Suppose a 3-out-of-4 system where each component is functioning with probabilities $\mathbf{p} = \{p_i\}$:

   a) Write the structure function of the system $\phi(\mathbf{x})$.

   b) Deduce analytically the reliability function $R(\mathbf{p})$. Evaluate it when $\{p_i\} = \{0.9, 0.5, 0.2, 0.1\}$.

   c) Estimate via simulation, for each $n = 10^2, 10^3, 10^4, 10^5$ realizations, the reliability of the system $R(\mathbf{p})$. Take $\{p_i\} = \{0.9, 0.5, 0.2, 0.1\}$.

   d) Find the standard deviation of the reliability estimation for each n.

   e) Find the 95 % CI's for each n. Do the confidence intervals contain the actual reliability value from item *b)*?

7. **Reading assignment (1 point)**

   Read *Chapter 2: Data and sampling distributions* of the book Practical statistics for data scientists (available as E-book in Utadeo). Include in your submission an exploration of the dataset *loans_income.csv* by replicating some of the codes of the sections: *Sampling Distribution of a Statistic*, *Bootstrap*, or *Confidence Intervals*. The dataset and R codes and notebooks are located in the github repository of the book.