

# EST-24107: Simulación

**Profesor:** Alfredo Garbuno Iñigo — Primavera, 2022 — Números aleatorios.

**Objetivo:** En esta sección veremos conceptos básicos del curso. Tanto para el trabajo computacional que realizaremos y la entrega de tareas, como las nociones básicas de generación de números aleatorios en una computadora. **Lectura recomendada:** Capítulo 3 de [3]. Capítulo 2 de [1].

## 1. AGENDA

Veremos un poco sobre el ambiente de trabajo. Veremos :

1. nociones básicas `Git` y `GitHub`;
2. familiaridad con `R`.
3. generación de variables aleatorias.

## 2. CONTROL DE VERSIONES

Los *softwares* de control de versiones nos permiten llevar un registro y administración de cambios en archivos. Usualmente para proyectos de programación.

Ayudan a trabajar colaborativamente en ambientes de equipos de trabajo.

Aunque no exploraremos *todo* lo que se puede hacer con `Git` y `GitHub` lo usaremos para llevar un control del desarrollo y de entrega de tareas. Usaremos los principios mas básicos.

## 3. R STATISTICAL PROGRAMMING LANGUAGE

`R` es un lenguaje de programación orientado a cómputo estadístico y generación de gráficos estadísticos. Está escrito para interactuar por medio de ejecución de *scripts* (archivos de texto con instrucciones) o la consola interactiva. Ver Fig. 1.

Es usual utilizar un ambiente de desarrollo para programar e interactuar con el lenguaje. Para `R` el mas común es `Rstudio` el cual tiene además algunas extensiones útiles para el desarrollo de análisis estadístico. Ver Fig. 2.

`Visual Studio Code` es una alternativa multi-lenguaje para desarrollar proyectos de análisis estadístico en `R`. Ver Fig. 3.

Y habemos los que nos *conformamos* con un buen editor de texto como ambiente de desarrollo. Ver Fig. 4.

## 4. NÚMEROS ALEATORIOS

Es posible, que cuando pensamos en generar números aleatorios, idealizamos con lanzar una moneda, un dado, una baraja o una rueda giratoria estilo *Jeopardy!*.

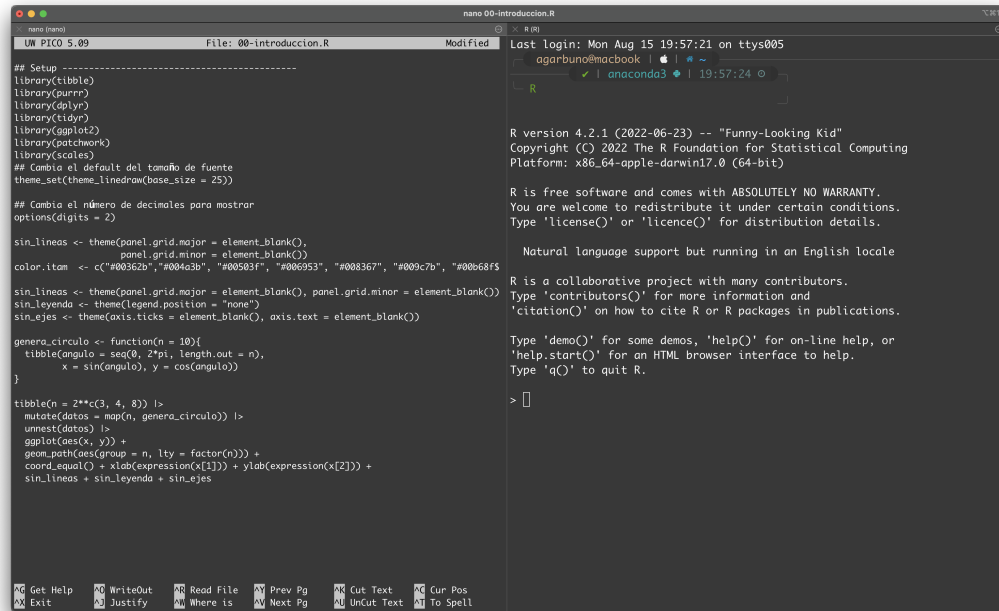


FIGURA 1. Dos ventanas, un editor de texto y una consola de R.

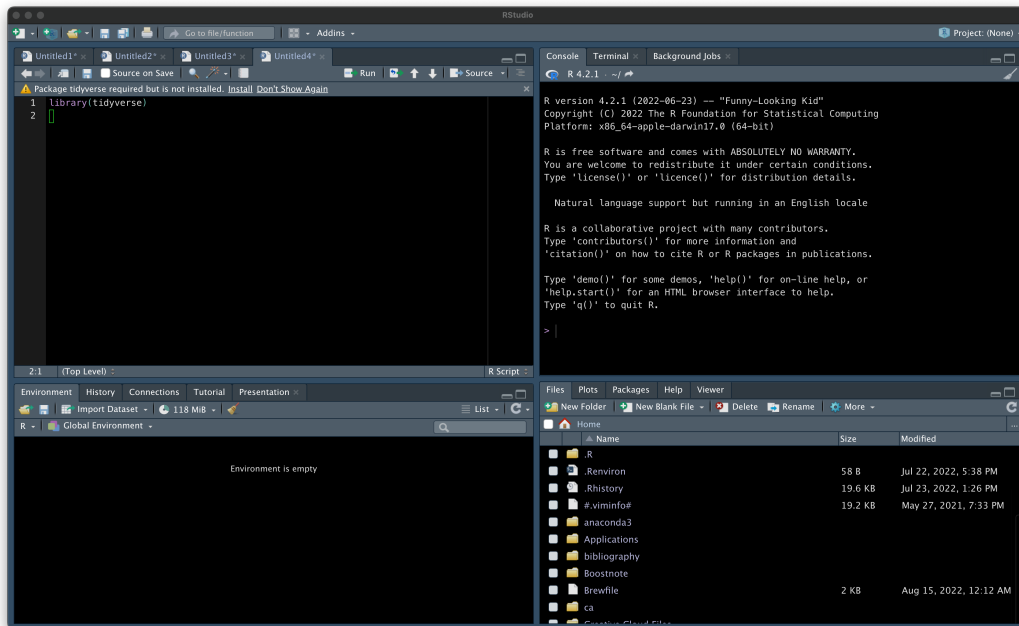


FIGURA 2. Un ambiente de desarrollo, Rstudio.

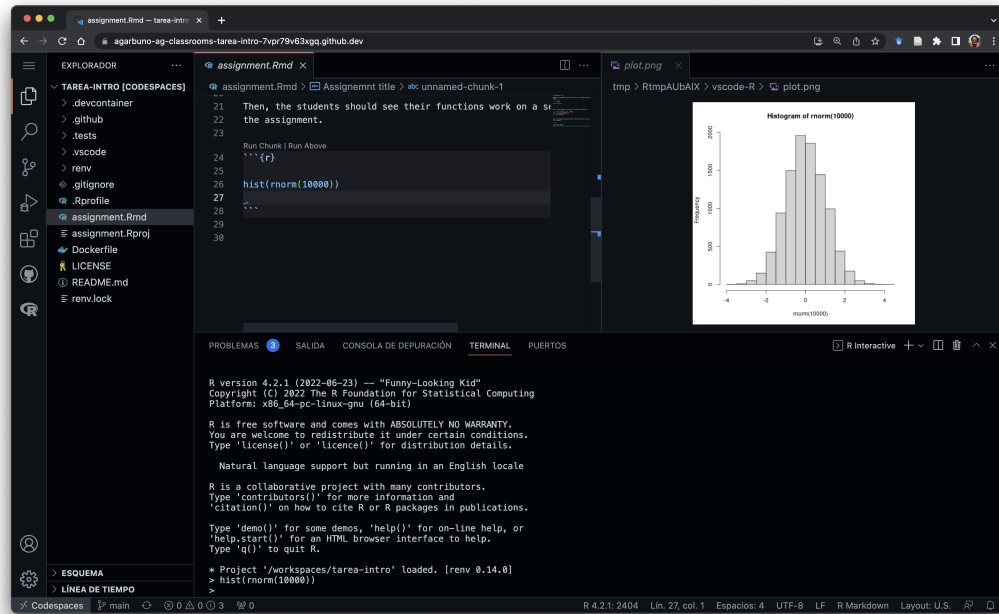


FIGURA 3. Un ambiente de desarrollo general, Visual Code Studio. En la imagen se muestra una sesión en un explorador de internet.

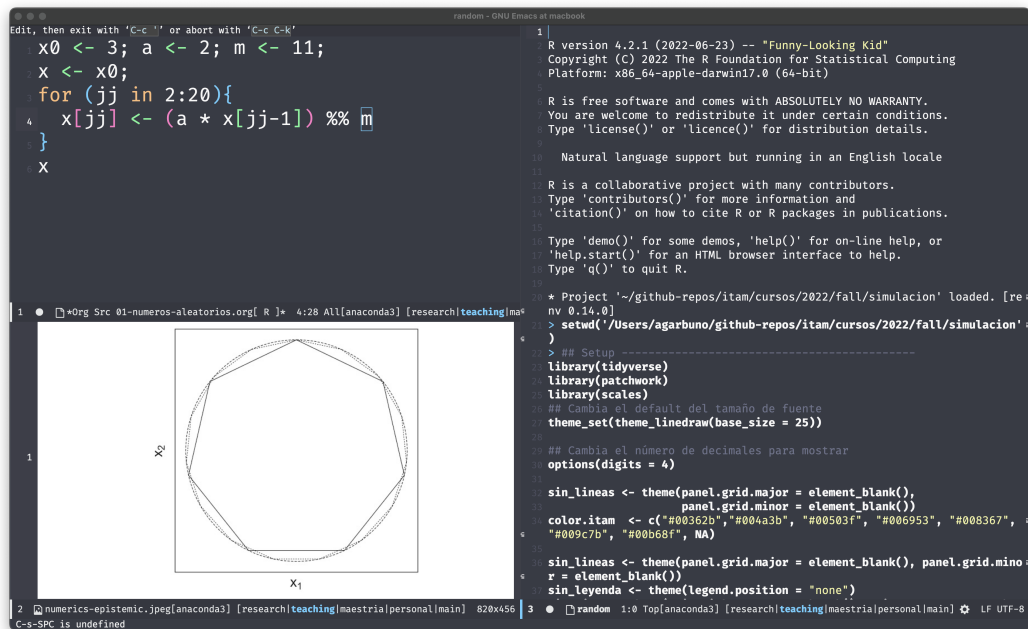


FIGURA 4. Ambiente de desarrollo basado en Emacs.



En nuestra computadora, los **números pseudo-aleatorios** son secuencias generadas de manera determinista de tal forma que *parecen* ser variables aleatorias uniformes independientes. Es decir, parecen ser

$$x_i \stackrel{\text{iid}}{\sim} U(0, 1). \quad (1)$$

El procedimiento mas común es utilizar una semilla  $x_0$  y calcular recursivamente valores  $x_n$  con  $n \geq 1$  por medio de

$$x_n = ax_{n-1} \text{ mód } m, \quad (2)$$

donde  $a$  y  $m$  son enteros positivos.

Nota que  $x_n$  es un valor entre  $0, 1, \dots, m-1$ . Llamamos a la cantidad  $x_n/m$  un número pseudo-aleatorio. Esto nos da un valor en el intervalo  $(0, 1)$ .

Las constantes  $a$  y  $m$  se escogen de tal forma que:

1. Para cualquier punto inicial, la secuencia *parece* ser un secuencia de números aleatorios uniformes.
2. Para cualquier punto inicial, el tiempo estimado para ver una repetición es muy largo.
3. Se puede calcular la secuencia eficientemente.

La constante  $m$  está asociada al periodo de la secuencia. Por ejemplo, podemos utilizar

$$x_n = 3x_{n-1} \text{ mód } 5, \quad (3)$$

para generar la secuencia a partir de  $x_0 = 3$ ,

```

1 x0 ← 3; a ← 3; m ← 5;
2 x ← x0;
3 for (jj in 2:10){
4   x[jj] ← (a * x[jj-1]) %% m
5 }
6 x
```

```

1 [1] 3 4 2 1 3 4 2 1 3 4
```

Si cambiamos los valores podemos conseguir un periodo mas largo y por lo tanto un mayor colección de números aleatorios.

```

1 x0 ← 3; a ← 2; m ← 11;
2 x ← x0;
3 for (jj in 2:20){
4   x[jj] ← (a * x[jj-1]) %% m
5 }
6 x

```

```

1 [1] 3 6 1 2 4 8 5 10 9 7 3 6 1 2 4 8 5 10 9 7

```

```

1 u ← x[1:(m-1)] / m
2 u

```

```

1 [1] 0.27273 0.54545 0.09091 0.18182 0.36364 0.72727 0.45455 0.90909 0.81818
2 [10] 0.63636

```

Usualmente  $m$  se escoge como un número primo de longitud igual al máximo número representable en una computadora.

Por ejemplo, en una máquina de 32-bits se ha visto que  $m = 2^{31} - 1$  y  $a = 7^5 = 16,807$  funcionan bien.

Esta elección nos permite generar una gran densidad en el intervalo  $(0, 1)$ . ¿Por qué?

#### 4.1. Aleatorios en lenguajes de programación

Los lenguajes de programación tienen funciones para generar números aleatorios. Por ejemplo, en Matlab el enfoque es cómputo numérico por lo tanto el generador de aleatorios uniformes es la opción estándar.

```

1 rand

```

```

1 0.06471600453257242

```

El lenguaje de `python` es multi-propósito. Por lo tanto, no es una opción *natural* y se llaman módulos especializados para generar números aleatorios. El módulo para generar números aleatorios tiene cierto nivel de compatibilidad con otros lenguajes.

```

1 import numpy as np
2 np.random.random()

```

```

1 0.9820617713830841

```

Por último, R es un lenguaje que se originó en la comunidad estadística. Por lo tanto, la generación de números aleatorios requiere de la distribución de interés.

```

1 runif(1)

```

```

1 [1] 0.02314

```

### 4.2. Transformación de uniformes

Es natural considerar la generación de números aleatorios en el intervalo  $[a, b]$ :

```
1 runif(100, min = 7, max = 10)
```

```
1 [1] 8.543 7.699 8.502 8.419 9.686 8.089 9.167 9.401 8.856 8.334 9.452 8.027
2 [13] 8.565 7.140 7.624 9.017 9.749 7.730 7.991 9.252 7.301 9.489 7.817 7.040
3 [25] 9.466 9.520 9.143 8.319 8.430 7.711 8.688 8.821 8.049 8.769 7.837 7.060
4 [37] 9.237 8.734 7.675 9.133 9.048 7.149 7.731 9.946 9.148 7.694 9.217 9.952
5 [49] 8.832 7.513 8.716 8.433 9.376 8.653 9.618 8.998 8.180 7.293 7.034 8.260
6 [61] 8.313 7.555 7.429 8.531 7.501 8.640 7.629 7.031 8.227 8.055 8.306 7.981
7 [73] 9.773 7.482 8.235 9.715 9.138 7.299 7.718 9.217 8.807 8.915 8.260 8.562
8 [85] 9.906 8.889 9.096 8.201 7.178 8.880 7.328 8.209 7.471 7.813 7.492 9.785
9 [97] 8.346 8.881 7.034 7.881
```

#### 4.2.1. Pregunta:

¿Cuál es la relación que existe entre  $X \sim U(0, 1)$  y  $Y \sim U(a, b)$ ?

### 4.3. Reproducibilidad

Hemos establecido que la generación de números *pseudo-aleatorios* es un procedimiento determinista. Si sabemos la semilla que generó la secuencia y el algoritmo que la genera, podemos generar dos secuencias idénticas. Por lo tanto, variables aleatorias completamente dependientes.

```
1 runif(5)
2 runif(5)
```

```
1 [1] 0.7008 0.7849 0.2975 0.3811 0.9899
2 [1] 0.89497 0.52216 0.04905 0.35857 0.78911
```

```
1 set.seed(108); runif(5)
2 set.seed(108); runif(5)
```

```
1 [1] 0.4551 0.4040 0.3513 0.6643 0.4635
2 [1] 0.4551 0.4040 0.3513 0.6643 0.4635
```

### 4.4. Aleatoriedad o pseudo-aleatoriedad

Consideremos una secuencia generada  $X_1, \dots, X_n$ . Entonces el conocimiento de  $X_n$  entonces no debería de dar información sobre  $X_{n+1}$  **si no conocemos el generador**.

La pseudo-aleatoriedad de nuestra secuencia es **limitada**. Pues dos muestras  $(X_1, \dots, X_n)$  y  $(Y_1, \dots, Y_n)$  que sean producidas por el mismo algoritmo no son independientes, ni idénticamente distribuidas o comparables en algún sentido probabilístico.

La validez de un generador se basa en una secuencia  $X_1, \dots, X_n$  con  $n \rightarrow \infty$ . No en una colección infinita de réplicas con longitud fija.

La distribución de esta colección de tuplas depende únicamente de la distribución de las semillas iniciales. Ver Capítulo 2 de [2].

4.4.1. *Definición [Generador pseudo-aleatorio]*: Decimos que un algoritmo es un generador de números uniformes pseudo-aleatorios si para algún valor inicial  $u_0$  y la aplicación de una transformación  $D : \mathbb{R} \rightarrow \mathbb{R}$  produce una secuencia

$$u_n = D^n(u_0) = \underbrace{(D \circ \dots \circ D)}_{n \text{ veces}}(u_0), \quad (4)$$

de valores en el intervalo  $(0, 1)$ . Además, el comportamiento de los valores  $(u_1, \dots, u_n)$  se comportan como si fueran una muestra iid de variables uniformes  $(V_1, \dots, V_n)$ .

#### 4.5. Comportamiento uniforme

Para validar que el generador de pseudo-aleatorios es válido tendremos que comparar las muestras generadas contra una distribución de probabilidad uniforme. ¿Qué podemos hacer?

#### 4.6. Ideas

Podemos comparar contra la distribución teórica que estamos generando. Esto es para tratar de garantizar estadísticamente que nuestra muestra se *ve* como una realización de números aleatorios uniformes.

### 5. PRUEBA DE KOLMOGOROV-SMIRNOV

Para comparar una muestra de números aleatorios podemos utilizar la prueba Kolmogorov-Smirnov (KS). La idea es sencilla: contrastar la **función de acumulación empírica** contra la **función de acumulación de una uniforme**.

5.0.1. *Definición [Función de acumulación de una variable uniforme]*:

5.0.2. *Definición [Función de acumulación empírica]*: Dada una muestra aleatoria  $X_1, \dots, X_n$  de variables con función de distribución  $\mathbb{P}$ , definimos

$$\hat{\mathbb{P}}_n(x) = \frac{\text{muestras menores o iguales a } x}{n}. \quad (5)$$

5.0.3. *Nota*: La función de acumulación empírica (EDF) la podemos definir a través de los **estadísticos de orden**.

5.0.4. *Definición [Estadísticos de orden]*: Dada una muestra aleatoria  $X_1, \dots, X_n$  los estadísticos de orden se definen como el reordenamiento  $X_{(1)} \leq \dots \leq X_{(n)}$ , donde

$$X_{(1)} = \min\{X_1, \dots, X_n\}, \quad \dots, \quad X_{(n)} = \max\{X_1, \dots, X_n\}. \quad (6)$$

### 6. CONCLUSIONES

#### REFERENCIAS

- [1] C. Robert and G. Casella. *Introducing Monte Carlo Methods with R*. Springer New York, New York, NY, 2010. [1](#)
- [2] C. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer Science & Business Media, mar 2013. [6](#)
- [3] S. M. Ross. *Simulation*. Academic Press, Amsterdam, 5th edition edition, 2013. [1](#)