Universidad Rey Juan Carlos

Departamento de Teoría de la Señal y Comunicaciones
y Sistemas Telemáticos y Computación

Guión de Prácticas: 1 Introducción a la Estadística Descriptiva con Julia
Procesamiento Digital de la Información
Curso 2023-24
Autores: David Casillas Pérez y Francisco J. Valverde Albacete

1. Introducción

En el Tema 1 se define una variable aleatoria X como cualquier función mensurable del espacio de probabilidad previamente definido, que transforma un suceso en una medida del suceso en el conjunto de los número reales, o en uno de sus subconjuntos (p.ej. "ALTURA" de una persona de esta clase). Las variables aleatorias pueden ser continuas o discretas, y dentro de estas últimas pueden ser numerables o finitas.

A nivel de simulación, debemos entender que las variables aleatorias solamente las podemos describir:

a) a partir de su densidad de probabilidad, en el caso de las variables continuas, o a partir de sus funciones de probabilidad, en el caso discreto.
b) a partir de sus funciones de distribución, en todos los casos.

La *Estadística Descriptiva* se encarga de caracterizar estas distribuciones, densidades o funciones de probabilidad.

En esta práctica primero generaremos *muestras (o poblaciones)* de *experimentos i.i.d (independientes e idénticamente distribuidos)*, para poder obtener descripciones muestrales que casen con las teóricas.

2. Generación de muestras aleatorias

Cuando trabajamos con ordenadores, generamos una muestra de valores asociados a una distribución, densidad o probabilidad. ¿Cómo podemos hacer que un ordenador genere números aleatorios que sigan una distribución dada? Es decir, ¿cómo podemos generar observaciones de una variable aleatoria dada su función de probabilidad?

Una pregunta más básica es si es posible generar números aleatorios en un computador. Claramente NO: un computador es una máquina de Turing que

sólo es capaz de ejecutar sentencias ya programadas ('un programa'). Para generar números aleatorias los ordenadores utilizan varias fuentes de datos *que pueden asumirse aleatorios*, y después usan estas fuentes para ejecutar algoritmos pseudo-aleatorios que generan muestras.

Estas fuentes de aleatoriedad pueden ser varias:
• La hora/minuto/segundo/milisengundos en que se ejecutan procesos
• Los clicks de tu ratón.
• Las teclas de tu ordenador.
• El ruido captado en un transistor en lazo abierto

Normalmente, todos los lenguajes de alto nivel ya tienen librerías programadas para que la condición de aleatoriedad se dé. Pero recordad que la aleatoriedad en un PC es una ilusión y así lo vamos a ver en esta sección. Visitad el apéndice "Introducción a Julia para la Estadística" para ver cómo hace Julia esta simulación

Ejercicio 1: Manejar el entorno estadístico básico de Julia

1. Encuentre la manera de generar muestras aleatorias Gausianas en la documentación de Juliastats.
2. Genere 10 muestras aleatorias de una variable aleatoria Gaussiana de media 0 y varianza 1, por ejemplo en el vector "muestra1"
3. Vuelva a generar 10 muestras aleatorias de una variable aleatoria Gaussiana de media 0 y varianza 1, esta vez en el vector "muestra2".
4. Responder: ¿Son iguales las muestras generadas?

Ejercicio 2: Control de aleatoriedad

La *semilla* de un método de generación aleatoria es un número que re-inicializa la fuente aleatoria. Se usa para garantizar la repetitividad de los experimentos estadísticos, un concepto importante en ciencia.

1. Encuentre el mecanismo de fijar la semilla de aleatoriedad de Julia. Fije la semilla a 1.
2. Genere 10 muestras aleatorias como las anteriores.
3. A continuación, vuelva a fijar la semilla a 1 y genere 12 muestras aleatorias de nuevo.

4. Responder: ¿Qué es lo que sucede con las nuevas muestras generadas?

Ejercicio 3: generación de muestras poblacionales

1. Genere ahora 10 muestras de tres variables aleatorias uniformes en los intervalos (0,1), (1,2) y (-1,1) y deles nombres adecuados.
2. Genere también 10 muestras de una variable aleatoria Gaussiana de media 1 y varianza 1, otras tantas de media 0 y varianza 1, y por último, otras tantas de media 1 y varianza 2.

3. Cálculo de momentos muestrales

Los *momentos poblacionales* son una serie de números que se utilizan para describir las distribuciones de las variables aleatorias, muchas veces desconocidas:

- El más famoso de todos los momentos es la *media (poblacional)*, o momento de orden 1, que se define como el valor medio que en general tomará la variable aleatoria.
- El momento central de orden 2, también conocido como *varianza*, mide el grado de dispersión de la distribución frente a su media.
- El momento de orden 3, o *asimetría*, muestra la simetría que guarda respecto a la media la distribución de la variable aleatoria. Si es nula, significa que la función de densidad de probabilidad es par.
- La *curtosis* es el momento central de orden 4 normalizado. Mide cómo son las colas: cuanto más grande la curtosis más sucesos raros tiene la distribución. Típicamente se compara respecto a las de la gausianas, que tienen Kur(G(0,1)) = 3

Frente a los momentos poblacionales, que dependen exclusivamente de las variables aleatorias consideradas, y en definitiva, de sus distribuciones asociadas, existen otros llamados los *momentos muestrales*. Los momentos muestrales dependen, como su propio nombre indica, de las muestras de la variable aleatoria.

Ejercicio 4: cálculo de momentos muestrales

1. Encuentre cómo hallar momentos ordinarios y centrales de una muestra en Julia.
2. Obtenga una muestra aleatoria de una variable aleatoria Gaussiana de media 2 y varianza 4, de tamaño 10, 100, y 1000 y calcule la media y la varianza muestral de cada muestra por separado.
3. Responder: ¿Cuál de ellas cree que se acerca más a la media y a la varianza poblacional y por qué?
4. Calcule también el momento de orden 3 y 4 de las muestras anteriores, pero sólo para las muestras de tamaño 1000.

Ejercicio 5: Repita el ejercicio 4 para las siguientes distribuciones:

a. Uniforme entre 0 y 1
b. Exponencial de parámetro lambda =1
c. Binomial de parámetro n=100, p=0,89
d. Poisson de parámetro lambda=2

5. Representación de poblaciones mediante histogramas

Un *histograma* no es más que un gráfico que subdivide el rango de valores que puede tomar una determinada variable aleatoria en segmentos disjuntos o *bins*, de tal modo que lleva a cabo un conteo de las muestras que caen en cada uno de los *bins*.

El teorema de Glivenko-Cantelli demuestra que el histograma normalizado converge a la función de distribución de la variable de la que se extraen las muestras.

Ejercicio 6: generación de histogramas

1. Genere una muestra de tamaño 1000 de las siguientes variables aleatoria:
   a. Gaussiana de media 0 y varianza 1
   b. Uniforme de 0 a 1
   c. Exponencial de parámetro lambda =1
   d. Binomial de parámetros 100, p=0,89
   e. Poisson de parámetro lambda=2

2. Localice el paquete y la primitiva que permite construir histogramas en JuliaStats.
3. Dibuje ventanas individuales el histograma de las 5 muestras con la función encontrada.
4. Encuentre y dibuje el histograma normalizado con las funciones:
    a. normalize con la opción de “pdf”
    b. normalize con la opción de “probability”
5. Pinte en cada histograma normalizado la función densidad de probabilidad de cada variable aleatoria. Pista: puede extraer el dominio de los bins con *h.Bins.*
6. Compruebe como coinciden el histograma se acerca a la función de densidad de probabilidad (caso continuo), o la función de probabilidad (discreto) a medida que el número de muestras aumenta.

Apéndices

A Introducción a Julia para la Estadística

Julia es un lenguaje de programación nacido en el seno de la comunidad científica para resolver determinados problemas de la práctica científica del que adolecen otros lenguajes (Python, MatLab, etc.)

En particular, la organización JuliaStats está encargada de proveer una serie de paquetes que doten a Julia de una buena estructura básica para la descripción y el modelado estadísticos.

A.1 Construcción del entorno de trabajo

1. El punto de entrada en Julia (para instalaciones, documentación, filosofía, etc) es la página:
   https://julialang.org
   Si Julia no está instalado en el sistemas o en MyApps, se puede hacer una descarga (que no se mantendrá entre sesiones, probablemente, por la gestión de los laboratorios Windows).
2. Para invocar Julia basta usar la línea de comandos (UN*X) o similar en Windows.
3. Luego hay que bajar al sistema los paquetes necesarios para hacer análisis estadístico
4. Finalmente se activan los paquetes con la expresión:

"Julia> using <Paq1>, <Paq2>,…

5. Una buena *cheat-sheet* (cast. "chuleta") de Julia la mantiene el autor de:

   https://github.com/JuliaDocs/Julia-Cheat-Sheet

   Casi toda (toda?) la práctica se puede resolver consultando sólo esta chuleta.

A.2 Recursos de Julia

- The Julia Programming Language: https://julialang.org
- Estadística con Julia: https://statisticswithjulia.org/tutorials/
- **Statistics** and **Machine Learning** made easy in Julia: https://juliastats.org, en particular los paquetes:
  - StatsBase.jl, basic functionalities for Statistics.
  - Distributions.jl, basic probability distributions.
- The Julia Cheat-sheet: https://github.com/JuliaDocs/Julia-Cheat-Sheet

B. Otros recursos

- Histogramas: https://es.wikipedia.org/wiki/Histograma

- Momentos muestrales

  - https://es.wikipedia.org/wiki/Momento_muestral
  - https://es.wikipedia.org/wiki/Esperanza_(matemática)
  - https://es.wikipedia.org/wiki/Varianza
  - https://es.wikipedia.org/wiki/Asimetría_estadística
  - https://es.wikipedia.org/wiki/Curtosis