Técnicas Multivariantes

Variables aleatorias y muestreo

Índice

Esquema 2

Ideas clave 3

1.1 Introducción y objetivos 3

1.2 Variables 4

1.3 Distribuciones 6

1.4 Muestreo 10

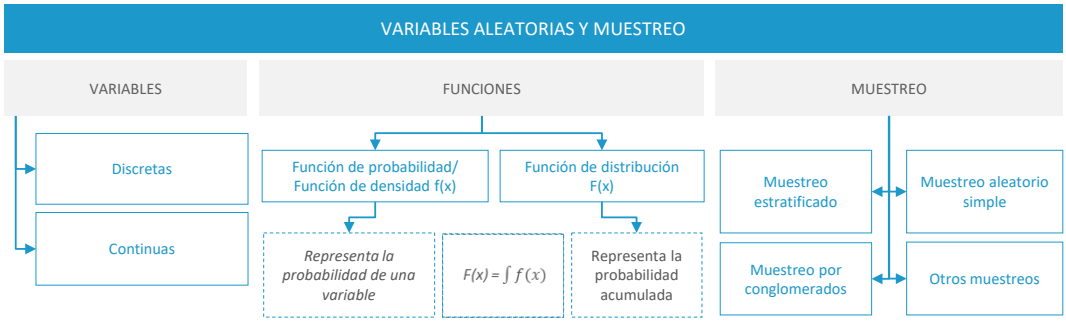
1.5 Inferencia estadística 13

1.6 El entorno de trabajo: Python 15

1.7 Referencias bibliográficas 16

1.8 Ejercicios resueltos 16

Esquema



Ideas clave

1.1 Introducción y objetivos

El muestreo es una herramienta fundamental dentro de la inferencia estadística, que permite realizar estimaciones de una población a través de una proporción, generalmente más pequeña, de la misma (la muestra).

Antes de profundizar en los distintos tipos de muestreo es conveniente repasar algunos aspectos previos.

Los apartados de los que consta este tema son:

- Variables
  - Discretas
  - Continuas
- Distribuciones
  - Función de densidad de probabilidad
  - Función de distribución
- Muestreo
  - Muestreo aleatorio simple (MAS)

- Estratificado
- Conglomerado
- Otros
  - Sistemático
  - No probabilístico

1.2 Variables

Una variable, o variable aleatoria, es una función que asigna un valor, usualmente numérico, al resultado de un experimento aleatorio. Las variables aleatorias, además, permiten construir modelos para multitud de situaciones experimentales que pueden analizarse mediante unas técnicas comunes. El uso de variables es muy útil en numerosas ocasiones, por ejemplo, supóngase que se quiere analizar un conjunto de sucesos aleatorios, correspondientes al resultado de lanzar una moneda al aire. Si realizamos un lanzamiento de la moneda, se obtiene un espacio muestral de 2 elementos: {cara y cruz}. Si realizamos dos lanzamientos de la moneda se obtiene un espacio muestral de 4 elementos: {cara-cara, cara-cruz, cruz-cara, cruz-cruz}. Si realizamos tres lanzamientos de la moneda se obtiene un espacio muestral de 8 elementos: {cara-cara-cara, cara-cara-cruz, cara-cruz-cara, cara-cruz-cruz, cruz-cara-cara, cruz-cara-cruz, cruz-cruz-cara, cruz-cruz-cruz}. Como podemos comprobar, este espacio muestral crece muy rápidamente. Si realizamos veinte lazamientos de la moneda, tendríamos un espacio muestral de $2^{20} = 1048576$ elementos, el cual se antoja bastante inmanejable. En cambio, si definimos el resultado del lanzamiento de las veinte monedas como una variable aleatoria X que cuente el número de caras del problema tendríamos una variable aleatoria que puede tomar como valores posibles $X = \{0, 1, 2, ..., 20\}$ con lo que se habría conseguido reducir el número de elementos de 1048576 a únicamente 21.

Variables discretas

Una variable discreta es una variable aleatoria que pude tomar un conjunto determinado de valores. Dentro de las variables discretas podemos distinguir entre las variables discretas numéricas y las no numéricas, es decir, las categóricas o cualitativas.

De forma general, una variable aleatoria es discreta numérica si el número de valores que puede tomar es finito o infinito numerable. Este hecho quiere decir que una variable discreta se puede definir mediante números naturales: $0, 1, 2, ..., n$. Por ejemplo, el resultado de tirar una moneda al aire, el resultado de lanzar un dado, o el número de ventas de un empleado en un mes son variables discretas numéricas.

Por otro lado, las variables categóricoas o cualitativas (o también llamadas factores), son variables que expresan cualidades o atributos de los agentes o individuos (sexo, nacionalidad, nivel de estudios, tratado/no tratado etc.). Por lo tanto, son variables discretas categóricas, por ejemplo:

- La presencia de una enfermedad que se representa mediante una variable cualitativa.

- El color de un automóvil.

Aunque en muchas ocasiones las variables cualitativas se representan mediante un código numérico, este no tiene sentido ni de orden ni de distancia, por ello no se deben realizar operaciones aritméticas con variables cualitativas.

Variables continuas

Una variable aleatoria es continua cuando puede tomar infinitos valores. Todas las variables continuas son cuantitativas: tienen un sentido numérico por sí mismas. La

altura, el peso o las distancias son ejemplos de variables continuas. En las variables continuas, la probabilidad de cualquier valor es cero. Si definimos a X como la variable aleatoria continua y x un posible valor de la recta real, se puede expresar:

$$P(X = x) = 0$$

Por otro lado, se define que una variable es absolutamente continua si su función de distribución es derivable en prácticamente todo su dominio.

1.3 Distribuciones

Variables discretas

Función de probabilidad f(x)

Para las variables discretas se puede definir la función de probabilidad o función de masa, $f(x)$, que cuantifica la probabilidad de esa variable para cada uno de sus posibles valores.

En la Figura 1 se muestra la función de probabilidades de la variable aleatoria obtenida del resultado de lanzar un dado.

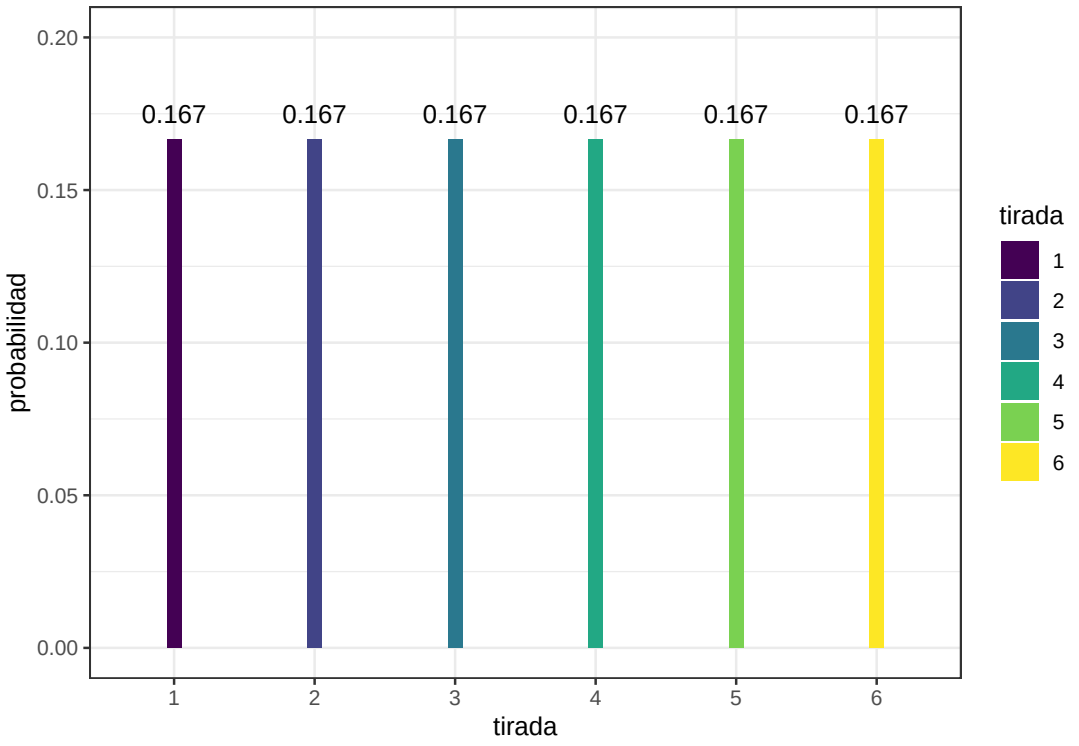


Figura 1: Función de probabilidad

Las propiedades de la función de probabilidad son:

- Es no negativa, es decir:

$$f(x) \geq 0, \forall x \in E_x.$$

- La suma de sus valores es la unidad, es decir:

$$\sum_{x \in E_x} f(x) = 1.$$

- La función de probabilidad de una variable aleatoria discreta determina la distribución de probabilidad de dicha variable, de modo que la función de distribución de una variable aleatoria discreta es una función de salto cuyos puntos de salto son aquellos números reales x_n tales que $P(X = x_n) > 0$. Donde dicha probabilidad es la longitud

del salto.

$$\forall x_n \in X \rightarrow f(x_n) > 0$$

Función de distribución F(x)

Para las variables discretas, también se puede definir la función de distribución de probabilidad o simplemente función de distribución, $F(x)$, en la cual se representa para cada valor toda la probabilidad que se va acumulando por valores inferiores. En la Figura 2 se muestra la función de distribución para los resultados posibles de lanzar un dado.

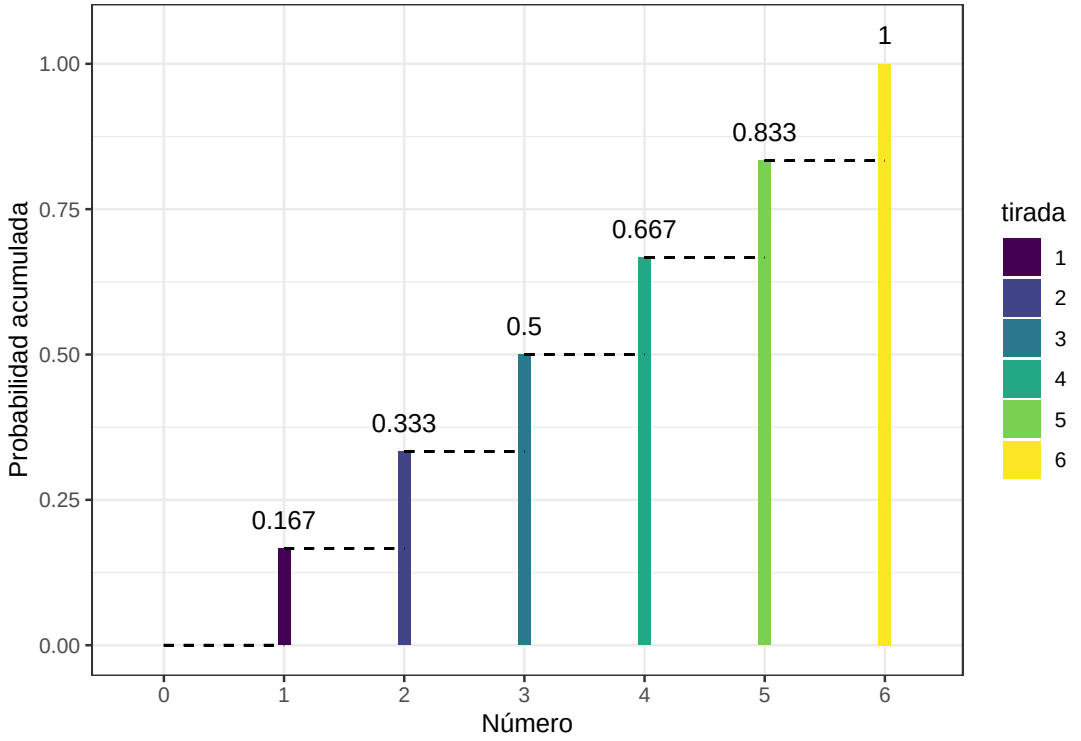


Figura 2: Función de distribución

Las propiedades de la función de distribución de una variable discreta son:

- Presenta un perfil escalonado,produciendose un salto en cada uno de los valores

definidos de la variable aleatoria.

- La cuantía de cada salto es precisamente la probabilidad en ese punto, obtenida de la función de probabilidad.

- Entre cada dos puntos (de los definidos en la función de probabilidad) no hay cambios de la probabilidad y por lo tanto no se añade nada a la acumulada.

Variables continuas

Función de densidad de probabilidad f(x)

La función de densidad de probabilidad de una variable continua describe la probabilidad relativa según la cual dicha variable aleatoria tomará determinado valor.

La probabilidad de que la variable aleatoria esté en una región específica del espacio de posibilidades estará dada por la integral de la densidad de esta variable entre uno y otro límite de dicha región. Ello implica que las probabilidades se van acumulando de manera suave, sin saltos bruscos y, por tanto, la derivada de la función de distribución es la función de densidad. La función de densidad de probabilidad (FDP o PDF en inglés) es positiva a lo largo de todo su dominio y su integral sobre todo el espacio es la unidad.

Función de distribución F(x)

La función de densidad de probabilidad, como se ha comentado, no proporciona directamente probabilidades. Las probabilidades son áreas bajo la curva de densidad y se pueden representar mediante la función de distribución. En las figuras 3 y 4 se muestran, respectivamente, las funciones de densidad y de distribución de la normal tipificada.

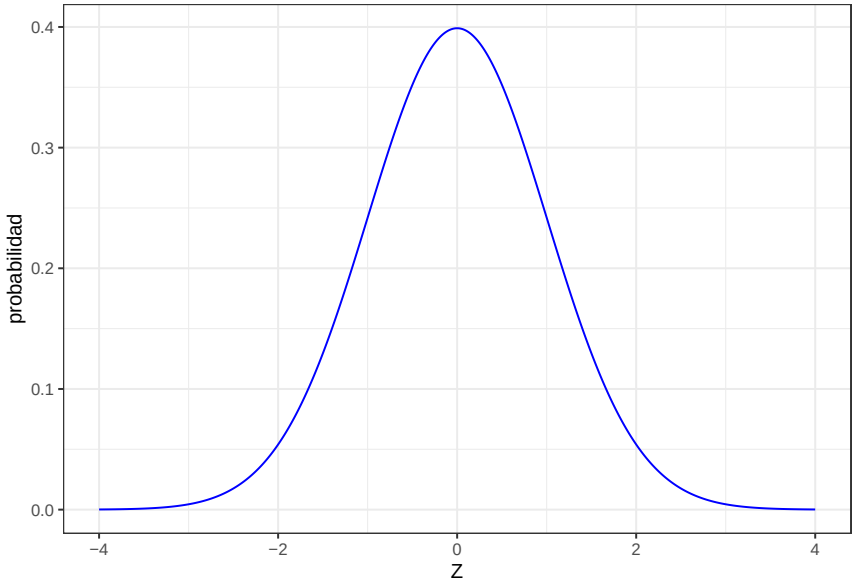


Figura 3: Función de distribución de la normal tipificada

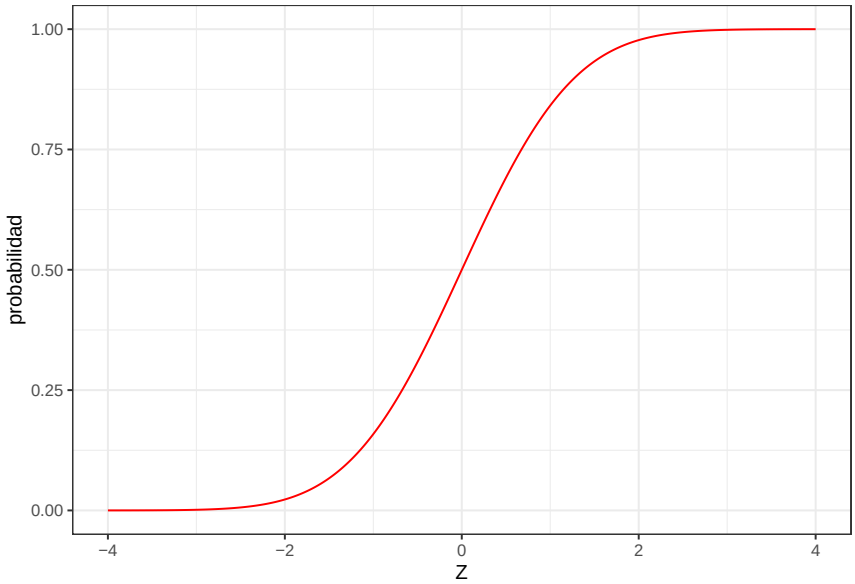


Figura 4: Función de distribución de la normal tipificada

1.4 Muestreo

En la ciencia de datos, y en la estadística en general, se necesita realizar inferencia sobre algunas poblaciones o conjuntos de datos de interés. Por ejemplo, se puede querer estimar la estatura de la población de un país, la contaminación que producen

sus vehículos, etc. Generalmente, no se tiene acceso a los datos de toda la población, para poder calcular el valor de los estadísticos reales, media, mediana, percentilest, etc. por lo que para poder realizar una estimación de los mismos es necesario utilizar el muestreo.

El muestreo es una herramienta por la cual se estudia una proporción más pequeña de una población para realizar inferencia estadística. Habitualmente, nos referiremos al tamaño de la población con N, y al de la muestra, con n. Además, nos referiremos con letras griegas a las distintas características estadísticas de la población (media: μ, desviación típica: σ) y con el alfabeto latino cuando nos refiramos a la muestra (media: $\bar{x}$, desviación típica: s).

Muestreo aleatorio simple (MAS)

El muestreo aleatorios simple es una técnica de muestreo en la cual todos los elementos que forman parte de la población tienen las mismas probabilidades de ser seleccionados para formar parte de la muestra.

Se dice que un conjunto de variables aleatorias $X_1, X_2, ..., X_n$ son una muestra aleatoria si se cumple que estas variables son independientes y están idénticamente distribuidas. No es necesario que los datos sean una muestra aleatoria, pero cuando lo son, se pueden aplicar distintas simplificaciones.

Estratificado

En algunas ocasiones, es preferible utilizar otros métodos de muestreo en vez del muestreo aleatorio simple (MAS). Por ejemplo, cuando la popblación es grande el MAS resulta muy costoso. En esos casos se han desarrollado otros tipos de muestreos. Uno de los más empleados es el muestreo estratificado, que se usa cuando se da la particularidad de que la variable de interés de estudio tiene distintos valores entre

diferentes subpoblaciones dentro de la población. Por ejemplo, si se quiere caracterizar la estatura de los mayores de edad de un determinado país, se puede dividir a la población total en diferentes subpoblaciones estratificando por sexo y edad. Una posible estratificación sería: hombres menores de 50, hombres mayores de 50, mujeres mayores de 50 y mujeres menores de 50 años de edad. De esta forma, si dividimos la población en esas subpoblaciones o estratos la muestra va a representar mejor al conjunto de la población para tamaños de muestra más reducidos que si emplearamos el MAS.

El método para realizar el muestreo estratificado es el siguiente:

- En primer lugar, se generan los L estratos en los que se divide la población de N elementos. En cada uno de los L estratos hay N_j elementos, con $j = 1, 2, ..., L$. La generación de los estratos no es trivial, sino que persigue dos objetivos:

  - Varianza intra-estrato mínima. Los elementos dentro de cada estrato tienen características muy parecidas de forma que existe una determinada homogeneidad dentro de cada estrato.

  - Varianza inter-estrato máxima. Al comparar los elementos entre diferentes estratos, estos deben tener características muy diferentes.

Al obtener los estratos de este modo, cada elemento del estrato puede representarlo por completo.

- Una vez se han generado los estratos, se aplica muestreo aleatorio simple sobre cada uno de ellos, obteniendo n_j muestras, con $j = 1, 2, ..., L$.

Conglomerado

El muestreo por conglomerados persigue la misma idea de subdividir a la población en subgrupos como hacía el muestreo estratificado. Sin embargo, la generación de los conglomerados tiene un enfoque distinto. En el ejemplo anterior, de lo que se trataría es de formar conglomerados en los que en cada uno de los conglomerados existiesen individuos de todas las subpoblaciones, de modo que cada conglomerado se parecería a la población general. EL modo de realizar este análisis es:

- En primer lugar, se generan los C conglomerados en los que se divide la población de N elementos. La generación de los conglomerados trata de cumplir los siguientes dos objetivos:
  - Varianza intra-estrato máxima. Los elementos dentro de cada conglomerado tienen características muy diferentes, de forma que existe una determinada heterogeneidad dentro de cada conglomerado.
  - Varianza inter-estrato mínima. Al comparar los elementos entre diferentes conglomerados, estos deben tener características muy parecidas.
- Al obtener los conglomerados de este modo, cada uno de ellos puede representar a la población por completo.

1.5 Inferencia estadística

Una vez se ha realizado el muestreo, se procede a realizar inferencia estadística. Para obtener estimadores de, por ejemplo, la media (μ) y la desviación típica de la población (σ) se van a emplear la media muestral y la desviación típica (o la cuasidesviación típica) muestral. En este punto, es necesario introducir algunos conceptos: los estadísticos y la distribución en el muestreo.

- Un estadístico es una función medible de los datos. Antes de observar los datos (de obtener la muestra), todo estadístico es una variable aleatoria.

- Distribución en el muestreo: Es la distribución de un estadístico. Si obtuviésemos distintas muestras del mismo tamaño, obtendríamos distintos valores del estadístico. La distribución de estos valores es lo que se conoce como distribución en el muestreo. Aunque en la mayoría de los casos sólo vamos a tener una muestra, y por lo tanto un único valor del estadístico, la consideración de la distribución en el muestreo es una construcción muy útil para poder realizar inferencia estadística.

Por ejemplo, supongamos que los datos de la población de interés son una muestra aleatoria de una distribución $N(\mu, \sigma)$ con ambos parámetros desconocidos. En esas condiciones, $\bar{x}$ y s^2, media y varianza muestrales, son estadísticos suficientes. Esto es debido a que la distribución en el muestreo de $\bar{x}$ es $N(\mu, \frac{\sigma}{\sqrt{n}})$ y a que los estadísticos $\bar{x}$ y s^2 son independientes. Mediante el pivote $\frac{(n-1)s^2}{\sigma^2}$, que se distribuye como una distribución χ^2_{n-1}, se puede obtener una estimación de la desviación típica poblacional. Mediante el pivote $\sqrt{n}\frac{\bar{x}-\mu}{s}$, que se distribuye como una distribución t_{n-1} se puede obtener una estimación de la media poblacional.

Más allá de la estimación de los parámetros de una distribución, para hacer cualquier afirmación en estadística (este medicamento es eficaz, somos más altos que nuestros padres, etc.) no basta con utilizar la estadística descriptiva, sino que es necesario realizar un contraste de hipótesis. Aun así corremos el riesgo de equivocarnos (cometer errores de tipo I y II).

- Contrastes diseñados para fijar el error I (proteger la hipótesis nula).

- Contrastes paramétricos: además minimizan el error II (uniformemente de máxima potencia).

Algunos de los principales contrastes de hipóteis son: la comparación de dos poblaciones utilizandp el test F de Snedecor para comparar dos varianzas y el test t de Student

para comparar dos medias a partir de muestras independientes.

1.6 El entorno de trabajo: Python

En esta asignatura se van a poner en práctica los conocimientos que se van adquiriendo. Para realizar este fin se propone emplear el entorno de programación Python, el cual es uno de los lenguajes de programación más utilizados en la actualidad y uno de los preferidos tanto en la industria como para los científicos de datos. Python fue creado en 1991 por Guido Van Rossom y desde entonces no ha parado de crecer: contiene miles de paquetes, dispone de una red de apoyo y de recursos disponible inmensa y se encuentra en muchas de las aplicaciones actuales.

Para poder trabajar con Python se recomienda encarecidamente realizar el curso que tienes disponible en UNIR por cursar este máster. Sin embargo, a lo largo de la asignatura se explicarán las librerías utilizadas y se desgranarán las líneas de código empleadas, tanto para la teoría, como para los ejercicios resueltos.

Para el desarrollo de los temas de esta asignatura, así como para los ejercicios resueltos, necesitarás instalar las siguientes librerías:

- numpy
- pandas
- array
- plotnine
- scipy
- matplotlib
- sklearn
- math
- statsmodels

Material audiovisual

Accede al vídeo: Teoremas aplicados estadística

1.7 Referencias bibliográficas

Evans, M. and Rosenthal, J. S. (2015). *Probabilidad y estadística : la ciencia de la incertidumbre*. Reverté, Barcelona.

Palmer, A. L. (1999). *Análisis de Datos, Etapa Exploratoria*. 01 edition.

Pat Fernández, L. A. (2013). *Introducción a los modelos de regresión*. Plaza y Valdés, México, D.F.

Peña, D. (1987). *Regresión y diseño de experimentos*.

1.8 Ejercicios resueltos

Ejercicio 1.

Obtén con python las muestras de las siguientes distribuciones:

a) 10 valores de la uniforme [0, 1)

b) 5 días de la semana, con repetición.

c) 8 valores de tirar 2 dados.

Solución

a) Se emplea la función *random.rand* de la librería **numpy** para simular una distribución uniforme.

```
# se va a necesitar la libreria numpy----------------------------
import numpy as np
# definir semilla para que los resultados sean los mismos--------
np.random.seed(seed = 11)
# apartado a)
muestra_uniforme = np.random.rand(5)
print("a)", muestra_uniforme)
```

```
## a) [0.18026969 0.01947524 0.46321853 0.72493393 0.4202036 ]
```

b) Se emplea la función *random.randint* de la librería **numpy** para simular números enteros.

```
# apartado b)-----------------------------------------------------
np.random.seed(seed = 11)
muestra_semana = np.random.randint(1, 7, size = 10)
# definir dias de la semana
dias_semana = ["lunes", "martes", "miercoles", "jueves",
"viernes", "sabado", "domingo"]
# representar los dias de la semana (ojo con los indices)--------
print("b)", dias_semana[muestra_semana[0]-1],
      dias_semana[muestra_semana[1]-1],
      dias_semana[muestra_semana[2]-1],
```

```
      dias_semana[muestra_semana[3]-1],
      dias_semana[muestra_semana[4]-1]
      )
```

```
## b) martes lunes jueves martes sabado
```

c) Se emplea la función *random.randint* de la librería **numpy** para simular números enteros. Hay que simular cada dado por separado para tener en cuenta correctamente las probabilidades de la suma de tiradas de los mismos.

```
# apartado c)-------------------------------------------------
np.random.seed(seed = 11)
muestra_dado1 = np.random.randint(1, 6, size = 8)
muestra_dado2 = np.random.randint(1, 6, size = 8)
muestra_dados = muestra_dado1 + muestra_dado2
print("c)", muestra_dados)
```

```
## c) [ 7  2  5  7 10  5  8  3]
```

Ejercicio 2

La altura de una población se distribuye como una normal de media 170 y desviación típica 5. Se pide:

a) Obtener las siguientes muestras de la población y representar las distribuciones obtenidas gráficamente:

- 10 valores de la normal (μ = 170, σ = 5)
- 100 valores de la normal (μ = 170, σ = 5)
- 10000 valores de la normal (μ = 170, σ = 5)

b) Obtener el valor teórico de la distribución normal (μ = 170, σ = 5)

c) Calcular la probabilidad de que una persona mida menos de 160 cm o más de 195 cm.

Solución:

a) se cargan las librerías que se van a necesitar para la resolución del problema.

```
# cargar librerias-----------------------------------------------
import pandas as pd
import numpy as np
from pandas.core.common import flatten
from plotnine import *
from array import *
import scipy.stats as stats
import math
import matplotlib as mpl
```

Mediante la función *random.normal* de la librería **numpy** se obtienen los valores de la normal.

```
# definir semilla para que los resultados sean los mismos--------
np.random.seed(seed = 3)
# apartado a)
normal_a1 = np.random.normal(170, 5, 10)
# apartado b)
normal_a2 = np.random.normal(170, 5, 100)
# apartado c)
normal_a3 = np.random.normal(170, 5, 10000)
```

Estos valores se organizan en un dataframe con la ayuda de la librería **pandas**.

```
# crear dataframe-------------------------------------------------
# lista de los valores muestreo
l_a1 = normal_a1.tolist()
l_a2 = normal_a2.tolist()
l_a3 = normal_a3.tolist()
l_comb = l_a1 + l_a2 + l_a3
# lista de la muestra a la que corresponde cada valor
clase = [["n = 10"] * 10, ["n = 100"] * 100,
["n = 10000"] * 10000]
# una sola lista
flattened_clase = list(flatten(clase))
# print(flattened_clase)
# crear dataframe
d = {"Altura":l_comb, "Muestra":flattened_clase}
df = pd.DataFrame(d)
```

Se representan gráficamente con la librería **plotnine** que permite representar los gráficos utilizando la sintáxis de *ggplot*.

```
(
    ggplot(df, aes(x = "Altura", color = "Muestra",
    fill = "Muestra")) +
    geom_density(alpha = 0.1) +
    ylab("densidad de probabilidad") +
    xlab("altura (cm)") +
    theme_bw() +
    theme(legend_position = "right",
    subplots_adjust={'right': 0.8})
)
```

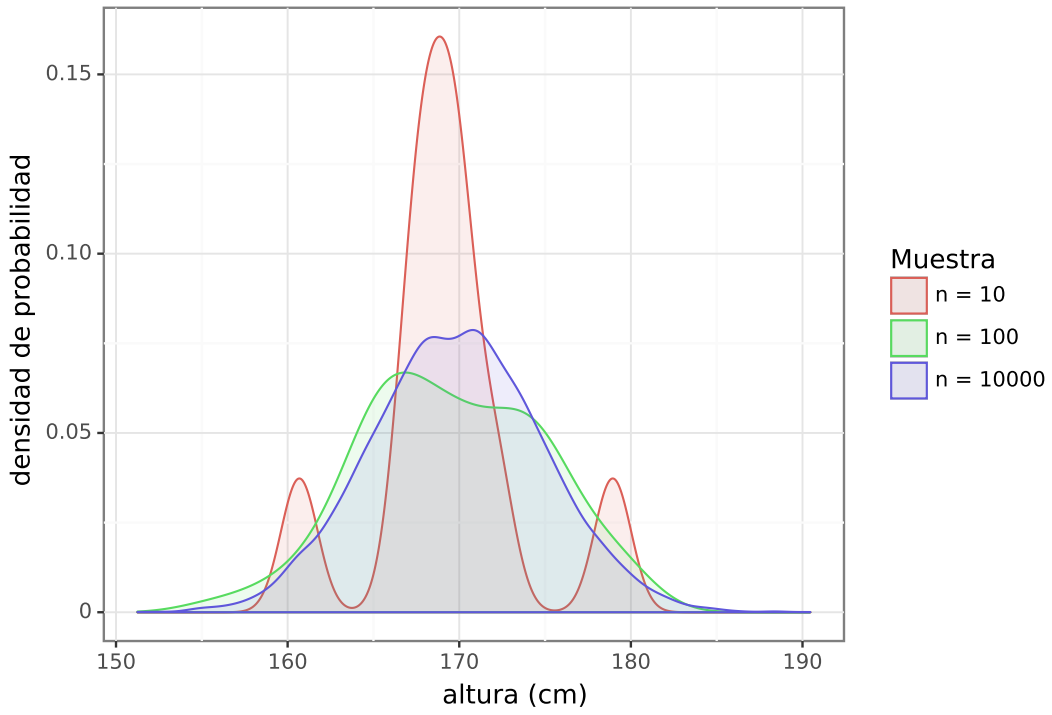


Figura 5: funciones de densidad obtenidas mediante muestreo

Se observa, que a mayor tamaño de muestra, la función de densidad obtenida se asemeja a la función de densidad de la distribución normal.

b) Se representa la distfunción de densidad real de la normal (μ = 170, σ = 5). Para ello se utiliza la librería **scipy**, que mediante la función *stats.norm.pdf* permite obtener la función de densidad de una distribución normal. Se ratifica que se parece mucho a la función de densidad obtenida en el apartado a) con n = 10000.

```
mu = 170
variance = 25
sigma = math.sqrt(variance)
x = np.linspace(mu - 3*sigma, mu + 3*sigma, 100)
y = stats.norm.pdf(x, mu, sigma)

(
ggplot(aes(x = x, y = y)) +
```

```
geom_line(color = "purple") +
ylim(0, 0.16) +
ylab("densidad de probabilidad") +
xlab("altura (cm)") +
theme_bw() +
theme(legend_position = "right",
subplots_adjust={'right': 0.8})
)
```

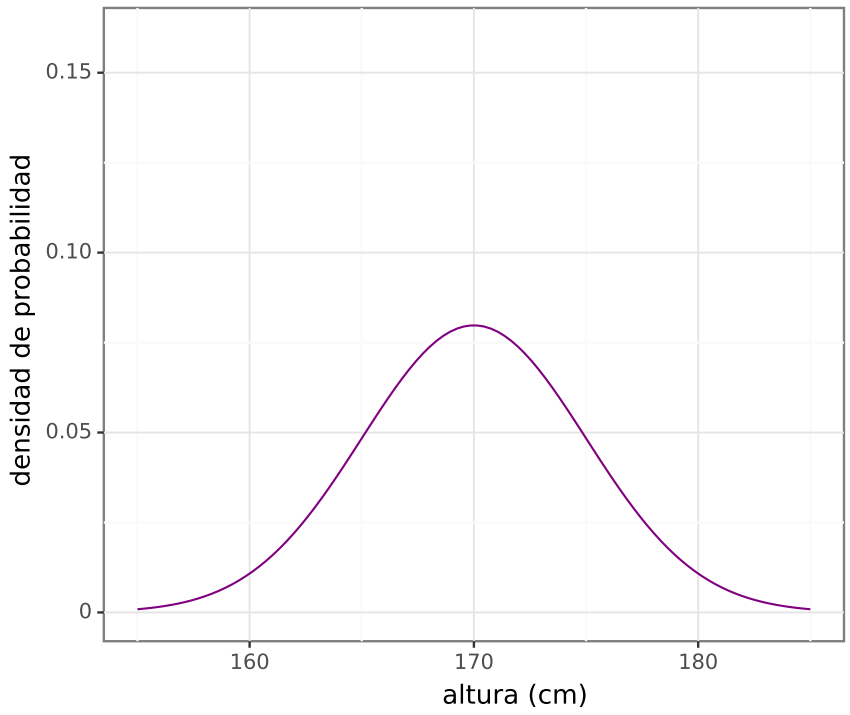


Figura 6: función de densidad real

c) La probabilidad será la suma de la probabilidad de que una persona mida menos de 160 cm + la probabilidad de que mida más de 195 cm. Además, la probabilidad de que mida más de 195 cm se puede obtener como 1 - probabilidad de que mida menos de 195 cm. Por lo tanto:

$$p = p_1 + p_2 = p(X < 160) + 1 - p(X < 195)$$

En python, p1 y p2 se pueden calcular mediante la función *stats.norm.cdf* de la librería **scipy**.

```
p1 = stats.norm.cdf(160, mu, sigma)
p2 = 1 - stats.norm.cdf(195, mu, sigma)
print("p =" , round(p1 + p2, 4))
```

```
## p = 0.0228
```

Se observa que la probabilidad es muy pequeña. A continuación, se presenta la función de distribución de probabilidad.

```
# representar la funcion de distribucion de probabilidad
x = np.linspace(mu - 6*sigma, mu + 6*sigma, 100)
y = stats.norm.cdf(x, mu, sigma)

(
ggplot(aes(x = x, y = y)) +
geom_line(color = "purple") +
geom_vline(xintercept = [160, 195], linetype = "dashed") +
ylim(0, 1) +
ylab("probabilidad acumulada") +
xlab("altura (cm)") +
theme_bw() +
theme(legend_position = "right",
subplots_adjust={'right': 0.8})
)
```

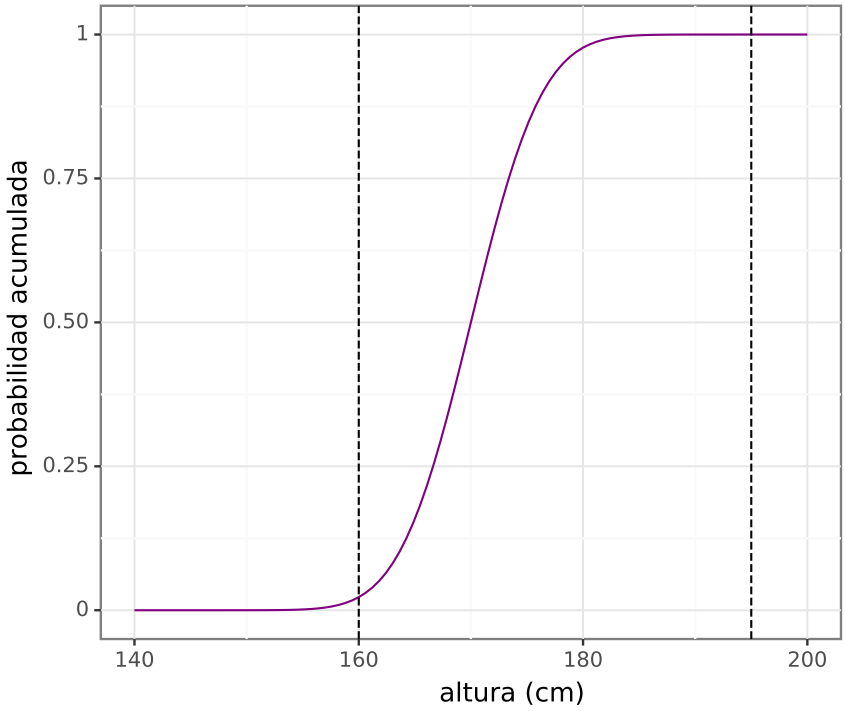


Figura 7: función de distribución real

Ejercicio 3

Latveria y Symkaria son dos poblaciones vecinas. Los hombres y mujeres de Latveria miden de media 175 cm y 169 cm respectivamente y tienen para ámbas distribuciones una desviación típica de 6 cm. En Symkaria, en cambio, las mujeres miden de media 173 cm y los hombres 171. La desviación típica en la altura de las mujeres de Symkaria es de 7 cm y en la de los hombres de 3 cm.

a) Con estos datos, conforma un dataframe que contenga las medidas de 2000 hombres y de 2000 mujeres de cada población. Calcula la media y la desviación típica de la población.

b) Realiza un muestreo estratificado con 4 estratos y n = 100/estrato. Calcula la media y desviación típica de cada estrato.

c) Realiza un muestreo por conglomerados con 2 conglomerados y n = 200/conglomerado. Calcula la media y desviación típica de cada conglomerado.

Solución

a) Mediante la función *random.normal* de la librería **numpy** se obtienen los valores de la normal para los distintos grupos.

```
# definir semilla para que los resultados sean los mismos--------
np.random.seed(seed = 3)
latveria_h = np.random.normal(175, 6, 2000)
latveria_m = np.random.normal(169, 6, 2000)
symkaria_h = np.random.normal(171, 3, 2000)
symkaria_m = np.random.normal(173, 7, 2000)
```

De manera similar a lo realizado en el ejercicio 2 se conforma el data frame.

```
# conformar data.frame-----------------------------------------
# lista de los valores muestreo
l_lh = latveria_h.tolist()
l_lm = latveria_m.tolist()
l_sh = symkaria_h.tolist()
l_sm = symkaria_m.tolist()
l_comb = l_lh + l_lm + l_sh + l_sm
# lista del sexo a la que corresponde cada valor
sexo = [["hombre"] * 2000, ["mujer"] * 2000] * 2
# una sola lista
flattened_sexo = list(flatten(sexo))
# len(flattened_clase)
# lista del pais a la que corresponde cada valor
pais = [["Latveria"] * 4000, ["Symkaria"] * 4000]
# una sola lista
```

```
flattened_pais = list(flatten(pais))
# len(flattened_clase)

# crear dataframe-------------------------------------------------
d = {"Altura":l_comb,"Sexo":flattened_sexo,"Pais":flattened_pais}
df = pd.DataFrame(d)
# mostrar previa del df-------------------------------------------
print(df)
```

```
##           Altura    Sexo      Pais
## 0     185.731771  hombre  Latveria
## 1     177.619059  hombre  Latveria
## 2     175.578985  hombre  Latveria
## 3     163.819044  hombre  Latveria
## 4     173.335671  hombre  Latveria
## ...          ...     ...       ...
## 7995  178.735655   mujer  Symkaria
## 7996  172.388692   mujer  Symkaria
## 7997  169.920577   mujer  Symkaria
## 7998  167.041882   mujer  Symkaria
## 7999  164.487935   mujer  Symkaria
##
## [8000 rows x 3 columns]
```

Se otbienen los resultados de la media y desviación típica poblacional empleando las funciones *mean* y *std* de la librería **numpy**.

```
media = np.mean(df["Altura"])
desv = np.std(df["Altura"])
print("La media es ", round(media, 2), " y la desv. tip. es ",
round(desv, 2))
```

```
## La media es  171.85  y la desv. tip. es  6.14
```

b) Para seleccionar los estratos debemos seleccionar las 4 combinaciones distintas de sexo y país. POsteriormente, se realiza una estracción aleatoria de 100 individuos por estrato con la función *sample* sobre el dataframe. Por último, se calcula la media y la desviación estándar para cada estrato.

```
np.random.seed(seed = 11)
# generar filtros-----------------------------------------------
e_h = df["Sexo"] == "hombre"
e_m = df["Sexo"] == "mujer"
e_L = df["Pais"] == "Latveria"
e_S = df["Pais"] == "Symkaria"
# generar estratos----------------------------------------------
estrato1 = df[e_h & e_L] # hombres Latveria
estrato2 = df[e_m & e_L] # mujeres Latveria
estrato3 = df[e_h & e_S] # hombres Symkaria
estrato4 = df[e_m & e_S] # mujeres Symkaria
# muestrear en cada estrato-------------------------------------
sample_1 = estrato1.sample(frac = 0.05)
sample_2 = estrato2.sample(frac = 0.05)
sample_3 = estrato3.sample(frac = 0.05)
sample_4 = estrato4.sample(frac = 0.05)
# calcular media y desv. tip.
print("La media del estrato 1 es ",
round(np.mean(sample_1["Altura"]), 2),
" y la desv. tip. es ", round(np.std(sample_1["Altura"]), 2))
```

```
## La media del estrato 1 es  174.61  y la desv. tip. es  6.14
```

```
print("La media del estrato 2 es ",
round(np.mean(sample_2["Altura"]), 2),
" y la desv. tip. es ", round(np.std(sample_2["Altura"]), 2))
```

```
## La media del estrato 2 es  168.28  y la desv. tip. es  6.06
```

```
print("La media del estrato 3 es ",
round(np.mean(sample_3["Altura"]), 2),
" y la desv. tip. es ", round(np.std(sample_3["Altura"]), 2))
```

```
## La media del estrato 3 es  170.99  y la desv. tip. es  2.78
```

```
print("La media del estrato 4 es ",
round(np.mean(sample_4["Altura"]), 2),
" y la desv. tip. es ", round(np.std(sample_4["Altura"]), 2))
```

```
## La media del estrato 4 es  173.94  y la desv. tip. es  6.04
```

Se observa que las medias y desviaciones típicas muestrales son similares a las poblacionales.

c) Aprovechando los estratos definidos del apartado anterior se conforman los conglomerados. En concreto, cada conglomerado estará formado por 50 observaciones de cada uno de los estratos.

```
np.random.seed(seed = 11)
# muestrear en cada estrato para el conglomerado 1---------------
sample_1 = estrato1.sample(frac = 0.025)
sample_2 = estrato2.sample(frac = 0.025)
sample_3 = estrato3.sample(frac = 0.025)
sample_4 = estrato4.sample(frac = 0.025)
# conformar conglomerado 1---------------------------------------
con_1=sample_1.append(sample_2).append(sample_3).append(sample_4)
# cong1

# muestrear en cada estrato para el conglomerado 2---------------
sample_1 = estrato1.sample(frac = 0.025)
sample_2 = estrato2.sample(frac = 0.025)
```

```
sample_3 = estrato3.sample(frac = 0.025)
sample_4 = estrato4.sample(frac = 0.025)
# conformar conglomerado 2----------------------------------------
con_2=sample_1.append(sample_2).append(sample_3).append(sample_4)

# calcular media y desv. tip.
print("La media del conglomerado 1 es ",
round(np.mean(con_1["Altura"]), 2),
" y la desv. tip. es ", round(np.std(con_1["Altura"]), 2))
```

```
## La media del conglomerado 1 es  172.31  y la desv. tip. es  6.07
```

```
print("La media del conglomerado 2 es ",
round(np.mean(con_2["Altura"]), 2),
" y la desv. tip. es ", round(np.std(con_1["Altura"]), 2))
```

```
## La media del conglomerado 2 es  172.64  y la desv. tip. es  6.07
```

Se observa que las medias y desviaciones típicas muestrales de los conglomerados son similares a la media poblacional y desviación típica general.