

Práctica 2

Introducción a la Estadística - Estadística descriptiva

Aclaración: Los ejercicios marcados con (*) son optativos.

Ejercicio 1. El desarrollador Web de un e-commerce desea estimar el tiempo medio de permanencia por sesión en su página web. Para ello, selecciona al azar una muestra de 50 visitantes de la página y registra el tiempo de permanencia de cada uno de ellos durante una sesión. Suponga que el desvío estándar (poblacional) del tiempo de permanencia en la página es de 10 segundos.

- Defina al parámetro de interés y proponga un estimador para él.
- Calcule aproximadamente la probabilidad de que el estimador propuesto difiera del verdadero valor del parámetro en menos de 2 segundos.
- ¿Cuántos visitantes se deberían seleccionar si se quiere que la probabilidad del ítem anterior sea por lo menos 0,9?

Ejercicio 2. Se quiere estimar el colesterol LDL promedio de todos los pacientes de un hospital. Para ello, se extrae una muestra al azar de 75 pacientes y se registra el colesterol LDL de cada uno de ellos.

- ¿Cuál es la población y cuál es el parámetro de interés?
- Proponga un estimador para el parámetro de interés.
- Definamos la v.a. X = “colesterol LDL (en mg/dL) de un paciente del hospital elegido al azar”. Suponga que X tiene distribución normal de varianza 625. Calcule (exacta o aproximadamente, según lo que sea posible), la probabilidad de que el estimador propuesto difiera del parámetro de interés en más de 3 mg/dL. ¿El cálculo fue exacto o aproximado? Justifique.

Ejercicio 3. Se quiere estimar la proporción de usuarios de determinado navegador de internet que personalizan su home-page para recibir noticias de temas y sitios que les resulten de interés. Para ello, se realiza un estudio en el que se recolectan datos de $n = 1675$ usuarios de dicho navegador. En éste, se revela que 496 de ellos personalizaron su home-page.

- ¿Cuál es la población y cuál es el parámetro de interés?

- b) Utilizando la información de la muestra calcule su mejor estimación del valor del parámetro.
- c) ¿Que habría que hacer para calcular el valor del parámetro exactamente?

Ejercicio 4. El pueblo Costa Esperanza tiene 100 habitantes de los cuales 15 trabajan 4 hs diarias, 30 trabajan 6 hs diarias, 20 trabajan 7 hs diarias y 35 trabajan 8 hs diarias.

- a) ¿Cuál es la proporción (poblacional) de habitantes que trabajan 4 hs diarias?
- b) Si definimos X = “cantidad de horas diarias de trabajo de un habitante elegido al azar del pueblo Costa Esperanza”, calcular la $P(X = 4)$. ¿Qué relación tiene este valor con el hallado en el ítem anterior?
- c) ¿Cuál es el promedio (poblacional) de la cantidad de horas de trabajo diario de los habitantes de Costa Esperanza?
- d) Calcule $\mathbb{E}(X)$. ¿Qué relación tiene este valor con el calculado en el ítem anterior?

Ejercicio 5. Cierta población tiene N habitantes de los cuales N_1 trabajan t_1 hs diarias, N_2 trabajan t_2 hs diarias, N_3 trabajan t_3 hs diarias y N_4 trabajan t_4 hs diarias. Sea

X = cantidad de horas de trabajo diario de un habitante elegido al azar de dicho pueblo.

- a) Pruebe que $\mathbb{P}(X = t_1)$ es igual a la proporción de habitantes de dicho pueblo que trabajan t_1 horas diarias. Es decir, que hay una equivalencia entre *probabilidad* y *proporción poblacional*.
- b) Pruebe que $\mathbb{E}(X)$ es igual a la cantidad de horas promedio de trabajo diario de los habitantes de dicho pueblo. Es decir, que hay una equivalencia entre *esperanza* y *media (o promedio) poblacional*.

Ejercicio 6. Un equipo de investigadores del Laboratorio de Neurociencia Cognitiva y Comportamiento desea estudiar el efecto de la fatiga cognitiva en el procesamiento del lenguaje. Para ello, administraron una prueba estandarizada de Tiempo de Reacción en milisegundos (ms) ante un estímulo léxico complejo a una muestra representativa de 30 personas jóvenes.

Se obtuvieron los siguientes resultados:

422.9 598.9 2024.4 835.6 308.4 647.7 779.0 560.1 658.4 691.8 2.0 498.9 134.9
536.4 501.7 30.3 216.3 706.7 116.9 835.6 208.4 106.2 531.9 26.8 181.4 424.3
187.4 338.9 84.9 394.5

- a) Definir la variable de estudio.
- b) Proponer un estimador para el tiempo medio de reacción y obtener una estimación.
- c) Proponer un estimador para la probabilidad de que el tiempo de reacción sea superior a los 500 ms y calcular una estimación.

- d) ¿Cuánto valen el mínimo, el máximo, el primer cuartil, el tercer cuartil y la mediana muestrales?
- e) Representar los datos en la recta numérica y decidir si la distribución es simétrica o asimétrica?
- f) Calcular la varianza y el desvío estándar muestral, como así también la distancia intercuartil.
- g) La muestra obtenida, tiene outliers?
- h) Realizar un boxplot.
- i) Representar los datos en un histograma de densidad.

Ejercicio 7. Considerar el conjunto de datos `airquality`, de la librería `datasets` de R. Ejecute el comando `data(airquality)` para cargar el conjunto en el panel del environment, ábralo para visualizarlo y pida ayuda con el comando `?airquality` para obtener información sobre el significado de sus variables.

- a) ¿Cuántas observaciones y cuántas variables tiene el conjunto de datos? (Hint: Explore los comandos `dim`, `nrow` y `ncol`.)
- b) ¿Cuáles son los nombres de las variables? (Hint: explore el comando `names`.)
- c) ¿De qué tipo es cada variable? Utilice el comando `str` (aplicado al conjunto de datos) y el comando `class` aplicado a cada variable.
- d) ¿Qué variables tienen datos faltantes y cuántos son? (Hint: explore el comando `summary` aplicado al conjunto de datos. Recuerde que los datos faltantes en R se codifican con `NA`).
- e) Realizar un histograma de densidad de las variables `Wind` y `Temp`.
 - I. ¿Diría que cada distribución es simétrica? Si hay asimetrías, ¿hacia qué lado son?

En base el histograma de `Temp`, calcule aproximadamente...

- II. qué proporción de datos tienen valor de `Temp` entre 65 y 70 .
- III. cuántos datos tienen valor de `Temp` entre 65 y 70.
- f) Realizar un boxplot para las variables `Wind` y `Temp`. ¿Identifica outliers? En caso afirmativo, ¿a qué valores corresponden? Explore el comando `sort`.
- g) A partir del boxplot de `Wind` calcule aproximadamente las 5 medidas resumen. Luego, calcúlelas con R y compare.
- h) Calcule media y mediana de la variable `Temp`, ¿cuál es mayor? ¿a qué cree que se debe?

- i) Calcule dos medidas de dispersión de la variable **Wind**

Ejercicio 8. (*) Sea $x_1, \dots, x_n$ un conjunto de datos.

a) Sea $f(a) = \sum_{i=1}^n (x_i - a)^2$. Probar que $f(a)$ alcanza su mínimo en $a = \bar{x}$.

b) Probar que $\sum_{i=1}^n (x_i - \bar{x}) = 0$

c) Probar que $\sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - n\bar{x}^2$.

Respuestas numéricas

Ejercicio 1

b) 0,8427

c) $n \geq 68$

Ejercicio 2

c) 0,2984

Ejercicio 3

b) Estimación: 0,2961