Contents

Presentación

Objetivo

El objetivo de estos ejercicios es proporcionar unos materiales de soporte para la asignatura de "Inferencia Estadística" del Máster interuniversitario de Bioiestadística y Bioinformática impartido conjuntamente por la Universitat Oberta de Catalunya (UOC) y la Universidad de Barcelona (UB).

Esta asignatura adolece de las características habituales de las asignaturas de posgrado, y especialmente de un posgrado de estadística (y bioinformática), que muestran algunas de las cosas que no debe de ser esta

asignatura:

Tal como se indica en la introducción a las notas de soporte del curso, este debería:

- Servir para repasar y consolidar los conceptos básicos que la mayoría de estudiantes traerán consigo.
- Además, y sobretodo, debe proporcionar una visión general, lo más completa posible dentro de las limitaciones de tiempo, del campo de la inferencia estadística

Y, naturalmente, una de las formas de consolidar conocimientos, como en cualquier disciplina cuantitatva,es a traves de la resolución de ejercicios que permiten reflexionar, comprender y ver como se aplican los conceptos teóricos introducidos.

Para ello, estos materiales contienen una serie de ejercicios similares a los que se proponen en las actividades y pruebas de evaluación continua de la asignatura.

La mayoría de los ejercicios estan resueltos, pero *es importante intentar resolverlos de forma autónoma antes de consultar la solución.*

En general los ejercicios no presuponen ningún conocimiento especial de matemáticas, más allá de las habilidades básicas que se adquieren durante los estudios de una carrera de ciencias o de ingeniería.

1 Probabilidad y Experimentos aleatorios

1.1 Problema 1

Sean A y B dos sucesos. Suponiendo que $P(A) = 0.3, P(B) = 0.6$, y $P(A \cap B) = 0.1$, calcula las siguientes probabilidades:

a) $P(A \cup B)$

b) $P(A^c)$

c) $P(Ac \cap B)$

d) $P(A \cap B^c)$

e) $P(A^c \cap B^c)$

1.1.1 Solución

a. $P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.3 + 0.6 - 0.1 = 0.8$
b. $P\left(A^c\right) = 1 - P(A) = 1 - 0.3 = 0.7$
c. $P\left(A^c \cap B\right) = P(B) - P(A \cap B) = 0.6 - 0.1 = 0.5$
d. $P\left(A \cap B^c\right) = P(A) - P(A \cap B) = 0.3 - 0.1 = 0.2$
e. $P\left(A^c \cap B^c\right) = 1 - P(A \cup B) = 1 - 0.8 = 0.2$

1.2 Problema 2

Una población está afectada por tres enfermedades diferentes A, B i C. La probabilidad de que una persona sufra A es 0.30 , la probabilidad de que sufra B es 0.20 y la probabilidad de que sufra C es 0.15 . La probabilidad de que una persona sufra A y B es 0.12 , la que sufra A y C es 0.09 y la que sufra B y C es 0.06 . La probabilidad de que una persona sufra las tres enfermedades es 0.03 . Se piden las probabilidades de que una persona escogida al azar:

1. padezca al menos una enfermedad
2. sólo sufra A
3. sufra B o C, pero no sufra A
4. sufra A o no sufra ni B ni C.

1.2.1 Solución

---

a) **¿Cuál es la probabilidad de que una persona padezca al menos una enfermedad?**

Queremos calcular la probabilidad de que una persona sufra al menos una de las tres enfermedades, es decir, $P(A \cup B \cup C)$.

Para calcular $P(A \cup B \cup C)$, usamos la regla de inclusión-exclusión:

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$$

Sustituyendo los valores dados en el enunciado:

$$P(A \cup B \cup C) = 0.30 + 0.20 + 0.15 - 0.12 - 0.09 - 0.06 + 0.03 = 0.41$$

Por lo tanto, la probabilidad de que una persona padezca al menos una enfermedad es **0.41**.

---

b) **¿Cuál es la probabilidad de que una persona solo sufra A?**

Para resolver esto, necesitamos calcular la probabilidad de que la persona sufra A, pero no B ni C, es decir, $P(A \cap B^c \cap C^c)$.

Podemos calcular $P(A \cap B^c \cap C^c)$ restando de $P(A)$ la probabilidad de que la persona sufra A junto con alguna de las otras dos enfermedades:

$$P(A \cap B^c \cap C^c) = P(A) - P(A \cap B) - P(A \cap C) + P(A \cap B \cap C)$$

Sustituyendo los valores:

$$P(A \cap B^c \cap C^c) = 0.30 - 0.12 - 0.09 + 0.03 = 0.12$$

Por lo tanto, la probabilidad de que una persona solo sufra A es **0.12**.

---

c) **¿Cuál es la probabilidad de que una persona sufra B o C, pero no sufra A?**

Aquí buscamos la probabilidad $P(A^c \cap (B \cup C))$, es decir, la probabilidad de que la persona no tenga A, pero tenga B o C.

Primero, calculamos $P(B \cup C)$ utilizando la regla de inclusión-exclusión:

$$P(B \cup C) = P(B) + P(C) - P(B \cap C)$$

Sustituyendo los valores:

$$P(B \cup C) = 0.20 + 0.15 - 0.06 = 0.29$$

Ahora, para calcular $P(A^c \cap (B \cup C))$, restamos de $P(B \cup C)$ la probabilidad de que la persona tenga A y alguna de las enfermedades B o C, es decir, $P(A \cap (B \cup C))$:

$$P(A \cap (B \cup C)) = P(A \cap B) + P(A \cap C) - P(A \cap B \cap C)$$

Sustituyendo los valores:

$$P(A \cap (B \cup C)) = 0.12 + 0.09 - 0.03 = 0.18$$

Finalmente, restamos:

$$P(A^c \cap (B \cup C)) = P(B \cup C) - P(A \cap (B \cup C)) = 0.29 - 0.18 = 0.11$$

Por lo tanto, la probabilidad de que una persona sufra B o C, pero no A, es **0.11**.

---

d) **¿Cuál es la probabilidad de que una persona sufra A o no sufra ni B ni C?**

Aquí buscamos la probabilidad $P(A \cup (B^c \cap C^c))$, es decir, que la persona sufra A o que no sufra ni B ni C.

Primero, calculamos $P(B^c \cap C^c)$, que es la probabilidad de que la persona no sufra ni B ni C. Esto es simplemente $1 - P(B \cup C)$, que ya calculamos previamente:

$$P(B^c \cap C^c) = 1 - P(B \cup C) = 1 - 0.29 = 0.71$$

Ahora, aplicamos la regla de la unión para calcular $P(A \cup (B^c \cap C^c))$:

$$P(A \cup (B^c \cap C^c)) = P(A) + P(B^c \cap C^c) - P(A \cap B^c \cap C^c)$$

Ya calculamos $P(B^c \cap C^c)$, y sabemos que $P(A \cap B^c \cap C^c)$ es la probabilidad de que una persona solo sufra A, que también calculamos previamente:

$$P(A \cap B^c \cap C^c) = 0.12$$

Sustituyendo los valores:

$$P(A \cup (B^c \cap C^c)) = 0.30 + 0.71 - 0.12 = 0.89$$

Por lo tanto, la probabilidad de que una persona sufra A o no sufra ni B ni C es **0.89**.

Resumiendo:

1. La probabilidad de que una persona padezca al menos una enfermedad es **0.41**.
2. La probabilidad de que una persona solo sufra A es **0.12**.
3. La probabilidad de que una persona sufra B o C, pero no A, es **0.11**.
4. La probabilidad de que una persona sufra A o no sufra ni B ni C es **0.89**.

1.3 Problema 3

Por los síntomas observados en un enfermo, y según la experiencia acumulada en un gran número de situaciones similares, se deduce que ha podido coger la enfermedad A con probabilidad 1/3, o la enfermedad B con probabilidad 2/3. Con el fin de precisar el diagnóstico, se hace un análisis clínico al enfermo con dos resultados posibles, positivo o negativo. Se sabe, también por experiencia, que en los pacientes que tienen la enfermedad En el análisis es positiva con probabilidad 0.99 , y en los que padecen la enfermedad B lo es con probabilidad 0.06

a) ¿Cuál es la probabilidad de que el análisis dé un resultado negativo?

b) Si el resultado ha sido positivo, ¿cuál es la probabilidad de que el paciente sufra la enfermedad A? ¿Y la probabilidad de que padezca la enfermedad B?

1.3.1 Solución

a.

$$P(Neg) = P(Neg|A) \cdot P(A) + P(Neg|B) \cdot P(B) = \\ = 0.01 \cdot 1/3 + 0.94 \cdot 2/3 = 0.63$$

b.

$$\mathrm{P(A}|Pos) = \frac{P(\text{ Pos }|A)P(A)}{P(\text{ Pos})} = 0.8919, \quad \text{para A,}$$
$$\mathrm{P(B}|Pos) = 1 - \mathrm{P(A}/Positiu) = 0.1081, \quad \text{para } B.$$

Las probabilidades las hemos calculado con R a partir de la información del enunciado:

```
pA<-1/3
pB<-2/3
ppA<-0.99
ppB<-0.06
pn<-(1-ppA)*pA+(1-ppB)*pB
pn
```

```
## [1] 0.63
```

1.4 Problema 4

El embolismo pulmonar es una condición relativamente común que necesita hospitalización y que a menudo ocurre en pacientes hospitalizados. La presión arterial menor de 90 mm HG es uno de los criterios importantes para diagnosticar esta condición. Supongamos que la sensibilidad del test es del 95% y la especificidad del test es del 75% y la prevalencia es del 20%.

a) Calcula el valor predictivo positivo del test.

b) Calcula el valor predictivo negativo del test.

c) Responde a las preguntas anteriores si la prevalencia fuera del 80%.

1.4.1 Solución

a) Calcula el valor predictivo positivo del test

$$VP+ = P(\text{ Embolismo / Test }+) = \frac{\text{Sens} \times \text{Prev}}{\text{Sens} \times \text{Prev } + (1 - \text{ Esp })(1 - \text{ Prev })}$$

```
sens<-0.95
esp<-0.75
prev<-0.20
vpp<-(sens*prev)/(sens*prev+(1-esp)*(1-prev))
vpp
```

```
## [1] 0.4871795
```

b) Calcula el valor predictivo negativo del test

$$VP- = \frac{\text{Esp}(1 - \text{Prev})}{\text{Esp}(1 - \text{Prev}) + (1 - \text{ Sens })\,\text{Prev}}$$

```
vpn<-(esp*(1-prev))/(esp*(1-prev)+(1-sens)*prev)
vpn
```

```
## [1] 0.9836066
```

Como se observa al tratarse de una prueba muy sensible y poco específica hay pocos falsos negativos y cuando el test da negativo hay una probabilidad muy alta (0.98) de que el individuo sea sano. No así cuando da positivo. Sólo el 48% serán verdaderos enfermos.

c) Responde a las preguntas anteriores si la prevalencia fuera del 80%

```
prev<-0.80
vpp<-(sens*prev)/(sens*prev+(1-esp)*(1-prev))
vpp
```

```
## [1] 0.9382716
```

```
vpn<-(esp*(1-prev))/(esp*(1-prev)+(1-sens)*prev)
vpn
```

```
## [1] 0.7894737
```

Si la prevalencia es más alta, el VP- sigue siendo alto, aunque no tanto pero hemos aumentado el VP+ hasta el 93% y no habrá tantos falsos positivos. Lo que está claro es el VPN y el VPP dependen de la prevalencia de la enfermedad.

1.5 Problema 5

Un índice que evalúa el síndrome de la muerte súbita (SMS) tiene una sensibilidad del 68% y una especificidad del 82%. ¿Cuáles son los valores predictivos positivo y negativo del índice si se aplica a una población donde se producen un $0,21\%$ de muertes súbitas sobre el total de nacimientos?

1.5.1 Solución

La prevalencia del síndrome de la muerte súbita en la población es del 0.21%, es decir 0.0021.

Nos piden que calculemos respectivamente los valores predictivos positivo y negativo del test. Es decir, que tan bien funciona el test para detectar la enfermedad (SMS) cuando da un resultado positivo ($T+$) y para indicar su ausencia (SMS^c), mediante un resultado negativo ($T-$).

$$VP+ = P[SMS|T+], \qquad VP- = P[SMS^c|T-],$$

Puede hacerse el cálculo directamente a partir de las probabilidades condicionadas.

$$\begin{aligned} VP+ = P[SMS|T+] &= \frac{P[T+|SMS] \times P[SMS]}{P[T+]} = \\ &= \frac{P[T+|SMS] \times P[SMS]}{P[T+|SMS] \times P[SMS] + P[T+|SMS^c] \times P[SMS^c]} = \\ &= \frac{\text{Sensibilidad} \times \text{Prevalencia}}{\text{Sensibilidad} \times \text{Prevalencia} + \text{1-Especificidad} \times \text{1-Prevalencia}} \end{aligned}$$

De forma análoga:

$$VP- = P[SMS^c|T-] = \frac{P[T-|SMS^c] \times P[SMS^c]}{P[T-]} =$$
$$= \frac{P[T-|SMS^c] \times P[SMS^c]}{P[T-|SMS^c] \times P[SMS^c] + P[T-|SMS] \times P[SMS]} =$$
$$= \frac{\text{Especificidad} \times \text{1-Prevalencia}}{\text{Especificidad} \times \text{1-Prevalencia} + \text{1-Sensibilidad} \times \text{Prevalencia}}$$

Estos cálculos se reañlizan de forma imediata usando R:

```
sensi <- 0.68
espec <- 0.82
prev <- 0.0021
vp.pos <- (sensi * prev )/ (sensi * prev + (1-espec)* (1-prev))
cat ("El valor predictivo positivo es: ", vp.pos)
```

```
## El valor predictivo positivo es:  0.007887324
```

```
vp.neg <- (espec * (1-prev) )/ (espec * (1-prev) + (1-sensi)* (prev))
cat ("El valor predictivo negativo es: ", vp.neg)
```

```
## El valor predictivo negativo es:  0.9991794
```

Como en el caso anterior, podemos ver que. al ser la prevalencia muy baja, el valor predicpositivo del test también lo es puesto que un test + tan solo indica en un 0,79% de veces la presencia del síndrome, correctamente.

2 Variables aleatorias y Distribuciones de probabilidad

2.1 Ejercicio 2.1

Se sabe que la presencia de algunas mutaciones en una región genómica puede influir en la sobreexpresión ("Up") o la inhibición ("Down") de dos genes distintos. Se conocen 6 variantes de dicha mutación y, dado que los efectos de la sobreexpresión de los dos genes son muy similares se ha optado por contar únicamente cuántos genes se sobre-expresan en presencia de cada una de ellas (un individuo puede presentar una única variante). Un estudio realizado sobre 300 pacientes ha permitido estimar las siguientes probabilidades de aparición de cada mutación así como el número de genes sobre-expresados asociados a las mismas. Los resultados se encuentran disponibles en la tabla siguiente:

Mutación	Probabilidad	N° de genes
e_1	0.15	0
e_2	0.13	1
e_3	0.07	1
e_4	0.30	2
e_5	0.20	2
e_6	0.15	0

Consideremos la variable aleatoria: $X =$ "Número de genes sobre expresados"

1. Obtener su distribución de probabilidad y representarla gráficamente
2. Calcular la esperanza y la varianza de dicha variable

SOLUCIÓN

La variable aleatoria que nos interesa es $X =$ "Número de genes sobre-expresados".

2.1.1 Distribución de probabilidad

Para obtener la distribución de probabilidad de X, necesitamos sumar las probabilidades de las mutaciones que tienen el mismo número de genes sobre-expresados.

Los posibles valores de X son 0, 1 y 2. A continuación calculamos la probabilidad de cada uno:

- Para $X = 0$, las mutaciones son e_1 y e_6:

$$P(X = 0) = P(e_1) + P(e_6) = 0.15 + 0.15 = 0.30$$

- Para $X = 1$, las mutaciones son e_2 y e_3:

$$P(X = 1) = P(e_2) + P(e_3) = 0.13 + 0.07 = 0.20$$

- Para $X = 2$, las mutaciones son e_4 y e_5:

$$P(X = 2) = P(e_4) + P(e_5) = 0.30 + 0.20 = 0.50$$

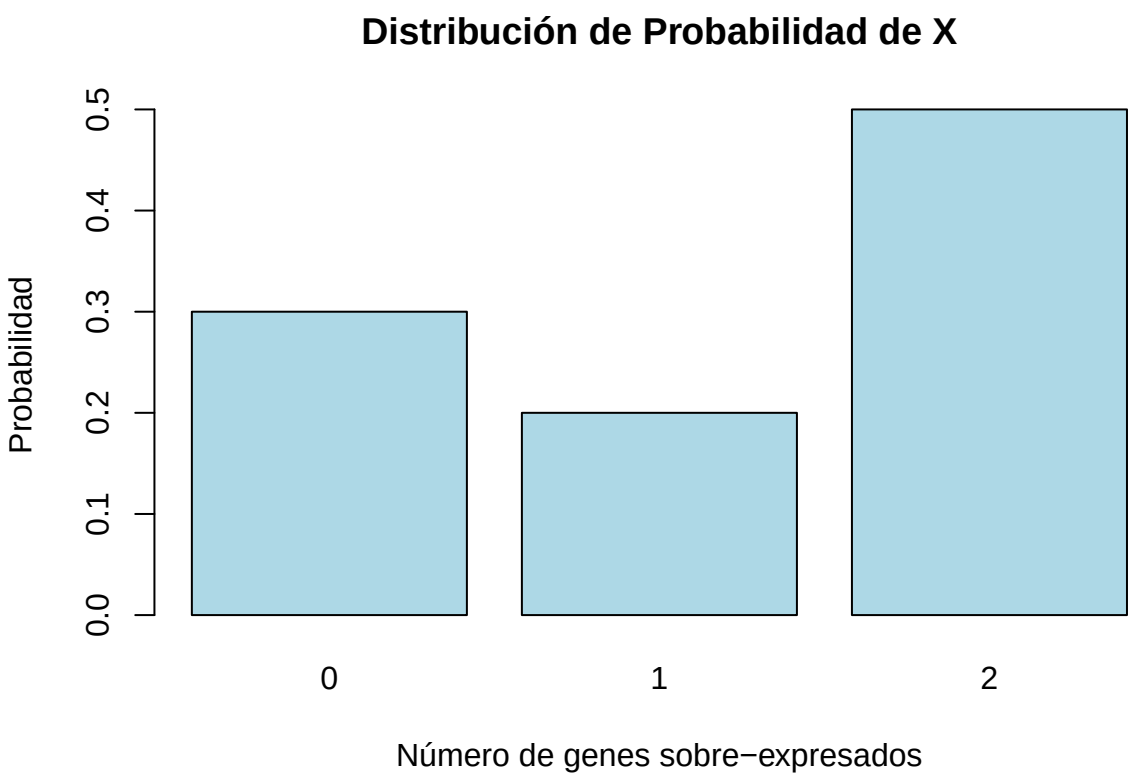
La distribución de probabilidad de X es la siguiente:

$$P(X = x) = \begin{cases} 0.30 & \text{si } x = 0, \\ 0.20 & \text{si } x = 1, \\ 0.50 & \text{si } x = 2. \end{cases}$$

Podemos representarla gráficamente usando R:

```r
# Valores de X y sus probabilidades
X_values <- c(0, 1, 2)
probabilities <- c(0.30, 0.20, 0.50)

# Crear el gráfico
barplot(probabilities, names.arg = X_values, col = "lightblue",
        main = "Distribución de Probabilidad de X",
        xlab = "Número de genes sobre-expresados", ylab = "Probabilidad")
```



2.1.2 Esperanza y varianza

La **esperanza** (o valor esperado) de una variable aleatoria discreta X se calcula como:

$$E(X) = \sum_x x \cdot P(X = x)$$

Sustituyendo los valores:

$$E(X) = 0 \cdot 0.30 + 1 \cdot 0.20 + 2 \cdot 0.50 = 0 + 0.20 + 1.00 = 1.20$$

La **varianza** de X se calcula como:

$$\text{Var}(X) = E(X^2) - [E(X)]^2$$

Primero calculamos $E(X^2)$:

$$E(X^2) = \sum_x x^2 \cdot P(X = x)$$

$$E(X^2) = 0^2 \cdot 0.30 + 1^2 \cdot 0.20 + 2^2 \cdot 0.50 = 0 + 0.20 + 2.00 = 2.20$$

Entonces, la varianza es:

$$\text{Var}(X) = 2.20 - (1.20)^2 = 2.20 - 1.44 = 0.76$$

Verificamos los cálculos con R:

```
# Calcular esperanza y varianza
esperanza <- sum(X_values * probabilities)
esperanza_cuadrado <- sum(X_values^2 * probabilities)

varianza <- esperanza_cuadrado - esperanza^2

esperanza
```

```
## [1] 1.2
```

```
varianza
```

```
## [1] 0.76
```

2.2 Ejercicio 2.2

Para describir el número de mutaciones presentes en un volumen estándar de un tumor unos investigadores han propuesto el modelo siguiente

$$p(x) = \frac{K}{2+x}, x = 0, 1, 2, 3, 4, 5$$

1. Determinar qué valor debe de tener K para que $p(x)$ sea una función de masa de probabilidad
2. Calcular su esperanza y su varianza
3. Calcular las probabilidades de los sucesos:
   - 1 Un tumor presenta exactamente tres mutaciones
   - 2 Un tumor presenta al menos una mutación
   - 3 Un tumor presenta como máximo dos mutaciones.

SOLUCIÓN

Se considera el modelo para la distribución de probabilidades de mutaciones en un tumor dado por:

$$p(x) = \frac{K}{2+x}, x = 0, 1, 2, 3, 4, 5$$

2.2.1 Valor de K

Para que $p(x)$ sea una función de masa de probabilidad, la suma de todas las probabilidades debe ser igual a 1. Es decir:

$$\sum_{x=0}^{5} p(x) = 1$$

Sustituyendo la fórmula de $p(x)$:

$$\sum_{x=0}^{5} \frac{K}{2+x} = 1$$

Simplificamos la suma:

$$K \sum_{x=0}^{5} \frac{1}{2+x} = 1$$

La suma es:

$$\sum_{x=0}^{5} \frac{1}{2+x} = \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5} + \frac{1}{6} + \frac{1}{7}$$

Podemos calcular esta suma numéricamente en R:

```
# Valores de la suma
suma <- sum(1 / (2 + 0:5))

# Calcular el valor de K
K <- 1 / suma
K
```

```
## [1] 0.6278027
```

2.2.2 Esperanza y la varianza

La **esperanza** de X se calcula como:

$$E(X) = \sum_{x=0}^{5} x \cdot p(x) = \sum_{x=0}^{5} x \cdot \frac{K}{2+x}$$

La **varianza** se calcula usando:

$$\text{Var}(X) = E(X^2) - [E(X)]^2$$

Para esto, primero calculamos $E(X^2)$:

$$E(X^2) = \sum_{x=0}^{5} x^2 \cdot p(x) = \sum_{x=0}^{5} x^2 \cdot \frac{K}{2+x}$$

Podemos calcular la esperanza y la varianza en R de la siguiente forma:

```
# Calcular la esperanza
esperanza <- sum((0:5) * K / (2 + 0:5))

# Calcular la esperanza al cuadrado
esperanza_cuadrado <- sum((0:5)^2 * K / (2 + 0:5))

# Calcular la varianza
varianza <- esperanza_cuadrado - esperanza^2

esperanza
```

```
## [1] 1.766816
```

```
varianza
```

```
## [1] 2.761769
```

2.2.3 Probabilidades

Probabilidad de que un tumor presente exactamente tres mutaciones

La probabilidad de que $X = 3$ es:

$$P(X = 3) = p(3) = \frac{K}{2 + 3}$$

Podemos calcularlo en R:

```
# Probabilidad de X = 3
P_X_3 <- K / (2 + 3)
P_X_3
```

```
## [1] 0.1255605
```

Probabilidad de que un tumor presente al menos una mutación

La probabilidad de que $X \geq 1$ es:

$$P(X \geq 1) = 1 - P(X = 0)$$

Podemos calcularlo en R:

```
# Probabilidad de X >= 1
P_X_1 <- 1 - K / (2 + 0)
P_X_1
```

```
## [1] 0.6860987
```

Probabilidad de que un tumor presente como máximo dos mutaciones

La probabilidad de que $X \leq 2$ es:

$$P(X \leq 2) = P(X = 0) + P(X = 1) + P(X = 2)$$

Podemos calcularlo en R:

```
# Probabilidad de X <= 2
P_X_2 <- sum(K / (2 + 0:2))
P_X_2
```

```
## [1] 0.6801196
```

2.3 Ejercicio 2.3

Un modelo simplificado del tiempo de supervivencia, en años, tras un diagnóstico de una variante de leucemia es el siguiente:

$$f_x(x) = -0.5 \cdot x + 1, \quad \text{donde} \quad 0 \leq x \leq 2$$

1. Comprobar que f_X es una densidad. Representarla gráficamente.
2. Calcular F_Xy representarla gráficamente.
3. Calcular $P(X \geq 1), P(X > 1), P(X = 1), f_x(1)$.
4. Calcular la probabilidad de que un individuo diagnosticado con leucemia sobreviva :

(i) menos de seis meses, (ii) entre seis meses y un año, (iii) más de dos años.

5. Calcular $E(X)$ i $\text{Var}(X)$.
6. En vista que el modelo anterior no ha resultado satisfactorio una bioestadística ha propuesto un modelo alternativo consistente en modelizar la variable como:

$$g_X(x) = \exp(-kx), \text{ dondex } \geq 0$$

Calcular la constante k para que g_x sea una función de densidad de probabilidad. Repetir los cálculos de los apartados b), c), d) y e) con el nuevo modelo. Discutir adecuación de ambos modelos a una situación real.

SOLUCIÓN

2.3.1 $f_X(x)$ es una densidad

Para comprobar que $f_X(x)$ es una función de densidad, necesitamos verificar que cumple las dos condiciones básicas:

1. $f_X(x) \geq 0$ para todo x en su dominio.
2. La integral de $f_X(x)$ sobre todo su dominio debe ser 1, es decir:

$$\int_0^2 f_X(x)\,dx = 1$$

La función de densidad dada es $f_X(x) = -0.5 \cdot x + 1$ con $0 \leq x \leq 2$.

Primero, comprobamos que $f_X(x) \geq 0$ para $x \in [0, 2]$. Evaluamos los valores extremos:

- $f_X(0) = -0.5 \cdot 0 + 1 = 1$
- $f_X(2) = -0.5 \cdot 2 + 1 = 0$

La función es no negativa en el intervalo dado.

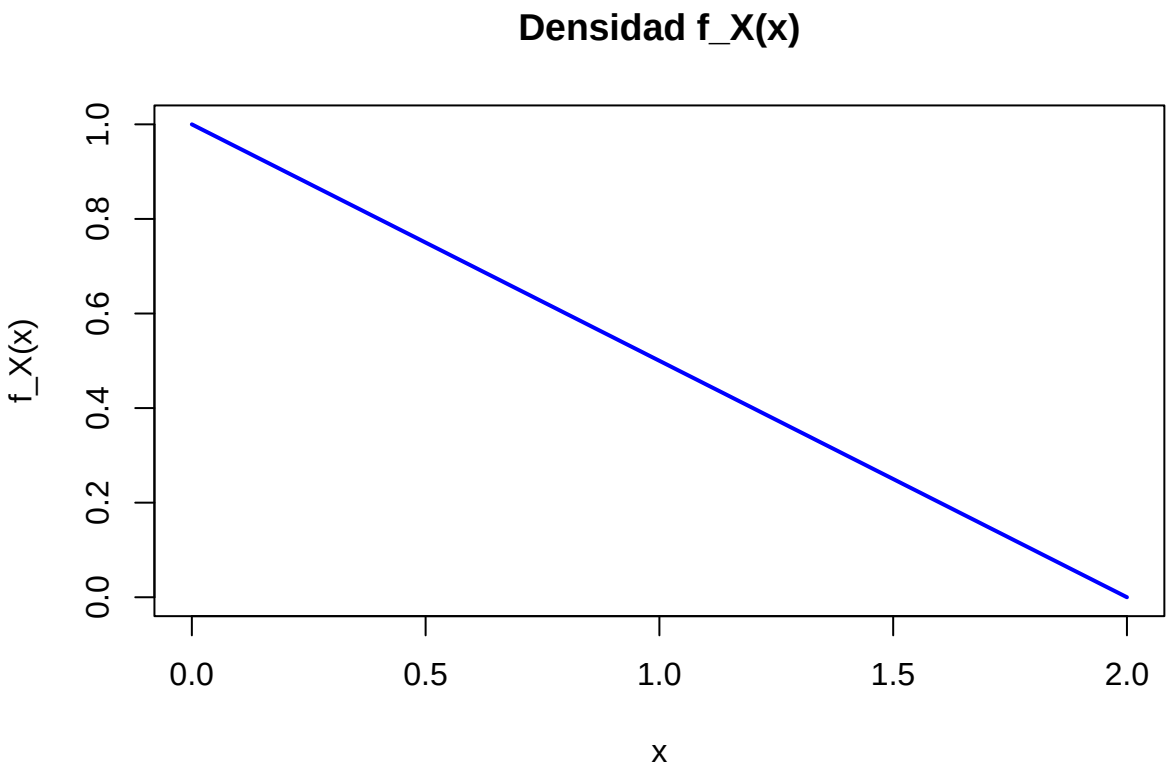
Ahora, calculamos la integral:

$$\int_0^2 (-0.5 \cdot x + 1)\,dx = \left[-0.25 \cdot x^2 + x\right]_0^2 = (-0.25 \cdot 4 + 2) - (0) = 1$$

Por lo tanto, $f_X(x)$ cumple con ambas condiciones y es una función de densidad.

2.3.2 Gráfica de $f_X(x)$

```r
# R code to plot the density function
f_x <- function(x) -0.5 * x + 1
curve(f_x, from = 0, to = 2, col = "blue", lwd = 2, ylab = "f_X(x)", xlab = "x",
      main = "Densidad f_X(x)")
```



2.3.3 Función de distribución

Calcular $F_X(x)$ y representarla gráficamente

La función de distribución acumulada (CDF) $F_X(x)$ se obtiene integrando la función de densidad:

$$F_X(x) = \int_0^x (-0.5 \cdot t + 1)\, dt$$

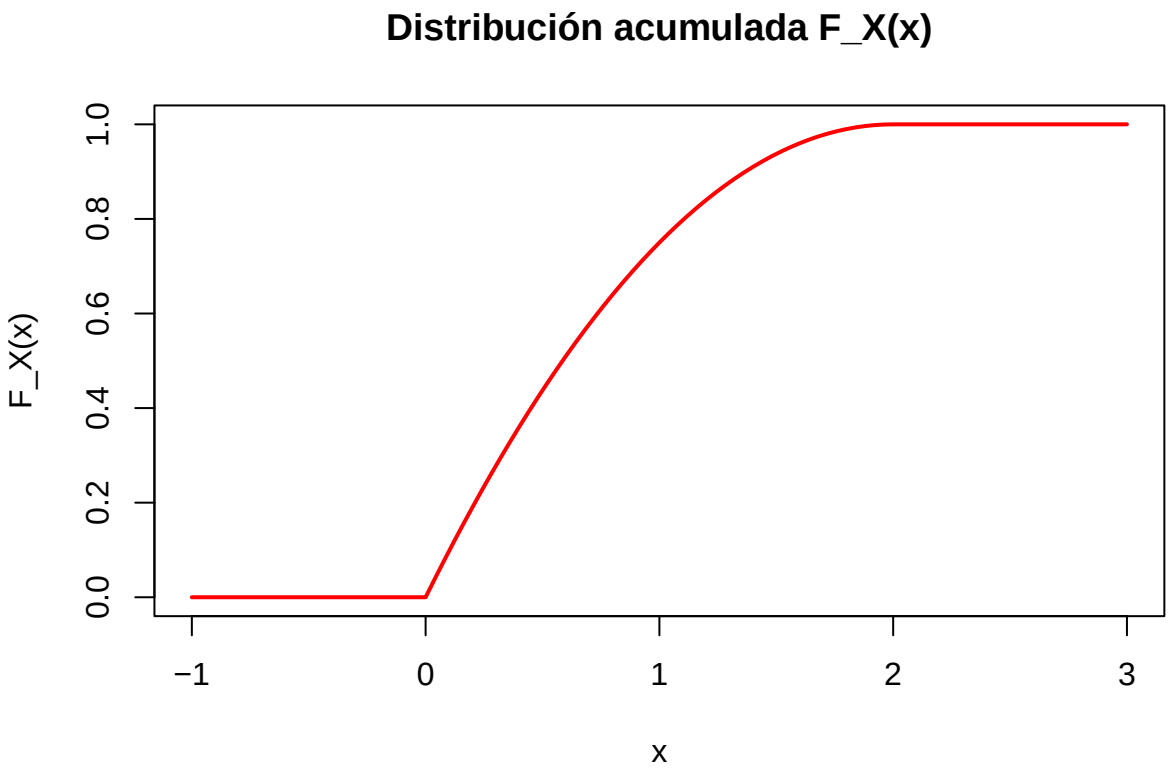
Para $x \in [0, 2]$, tenemos:

$$F_X(x) = \left[-0.25 \cdot t^2 + t\right]_0^x = -0.25 \cdot x^2 + x$$

Para $x < 0$, $F_X(x) = 0$, y para $x > 2$, $F_X(x) = 1$.

Gráfica de $F_X(x)$ç

```r
# R code to plot the CDF function
F_x <- function(x) ifelse(x < 0, 0, ifelse(x > 2, 1, -0.25 * x^2 + x))
curve(F_x, from = -1, to = 3, col = "red", lwd = 2, ylab = "F_X(x)", xlab = "x",
      main = "Distribución acumulada F_X(x)")
```



2.3.4 Probabilidades y $f_X(1)$

1. $P(X \geq 1) = 1 - F_X(1)$:

$$F_X(1) = -0.25 \cdot 1^2 + 1 = 0.75$$

Por lo tanto, $P(X \geq 1) = 1 - 0.75 = 0.25$.

2. $P(X > 1)$: Como X es una variable continua, $P(X > 1) = P(X \geq 1) = 0.25$.
3. $P(X = 1)$: Para una variable continua, la probabilidad puntual es 0, es decir, $P(X = 1) = 0$.
4. $f_X(1)$:

$$f_X(1) = -0.5 \cdot 1 + 1 = 0.5$$

2.3.5 Probabilidad de supervivencia

1. Menos de seis meses ($x = 0.5$):

$$P(X < 0.5) = F_X(0.5) = -0.25 \cdot 0.5^2 + 0.5 = 0.4375$$

2. Entre seis meses y un año ($x \in [0.5, 1]$):

$$P(0.5 \leq X \leq 1) = F_X(1) - F_X(0.5) = 0.75 - 0.375 = 0.375$$

3. Más de dos años ($x > 2$): Como el dominio de X es $[0, 2]$, $P(X > 2) = 0$.

2.3.6 $E(X)$ y $\mathrm{Var}(X)$

1. La esperanza de X es:

$$E(X) = \int_0^2 x \cdot f_X(x)\, dx = \int_0^2 x \cdot (-0.5 \cdot x + 1)\, dx$$

Desarrollamos:

$$E(X) = \int_0^2 (-0.5 \cdot x^2 + x)\, dx = \left[-\frac{0.5}{3} \cdot x^3 + 0.5 \cdot x^2 \right]_0^2$$

Calculamos:

$$E(X) = -\frac{0.5}{3} \cdot 8 + 0.5 \cdot 4 = -\frac{4}{3} + 2 = \frac{2}{3}$$

2. La varianza de X es:

$$\mathrm{Var}(X) = E(X^2) - E(X)^2$$

Primero calculamos $E(X^2)$:

$$E(X^2) = \int_0^2 x^2 \cdot f_X(x)\, dx = \int_0^2 x^2 \cdot (-0.5 \cdot x + 1)\, dx$$

Desarrollamos y calculamos:

$$E(X^2) = \int_0^2 (-0.5 \cdot x^3 + x^2)\, dx = \left[-\frac{0.5}{4} \cdot x^4 + \frac{1}{3} \cdot x^3 \right]_0^2$$

$$E(X^2) = -\frac{0.5}{4} \cdot 16 + \frac{1}{3} \cdot 8 = -2 + \frac{8}{3} = \frac{2}{3}$$

Finalmente:

$$\mathrm{Var}(X) = E(X^2) - E(X)^2 = \frac{2}{3} - \left(\frac{2}{3}\right)^2 = \frac{2}{3} - \frac{4}{9} = \frac{2}{9}$$

2.3.7 Modelo alternativo $g_X(x)$

Dado el modelo alternativo $g_X(x) = \exp(-k \cdot x)$ para $x \geq 0$, la constante k se determina imponiendo que la integral de la función de densidad debe ser 1:

$$\int_0^\infty \exp(-k \cdot x)\, dx = 1$$

Resolviendo:

$$\frac{1}{k} = 1 \implies k = 1$$

Por lo tanto, el nuevo modelo de densidad es $g_X(x) = \exp(-x)$.

2.4 Ejercicio 2.4

Para estudiar la regulación hormonal de una línea metabólica se inyectan ratas albinas con un fármaco que inhibe la síntesis de proteínas del organismo. En general, 4 de cada 20 ratas mueren a causa del fármaco antes de que el experimento haya concluido. Si se trata a 10 animales con el fármaco, ¿cuál es la probabilidad de que al menos 8 lleguen vivas al final del experimento?

SOLUCION

En este problema en el que tenemos grupos de 10 animales independientes, cada uno de los cuales puede sobrevivir o no resulta apropiada **la distribución binomial**.

- La probabilidad de que una rata sobreviva al fármaco es $p = \frac{16}{20} = 0.8$, dado que 4 de cada 20 ratas mueren.
- El experimento se realiza con 10 ratas, por lo que tenemos $n = 10$.
- Queremos calcular la probabilidad de que al menos 8 ratas sobrevivan. Matemáticamente, esto corresponde a:

$$P(X \geq 8)$$

donde X es el número de ratas que sobreviven y sigue una **distribución binomial**:

$$X \sim \text{Binomial}(n = 10, p = 0.8)$$

2.4.1 Cálculo de la probabilidad

La probabilidad de que exactamente k ratas sobrevivan está dada por la fórmula de la binomial:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Para responder la pregunta debemos calcular:

$$P(X \geq 8) = P(X = 8) + P(X = 9) + P(X = 10)$$

Esto puede calcularse:

- directamente usando la función de probabilidad acumulada implementada en R
- indirectamente calculando las probabilidades individuales y sumándolas.

En todo caso debemos recordar que al tratarse de una variable discreta si queremos usar $F_X(x)$ para calcular $P(X \geq k)$ deberemos tener en cuenta que:

$$P(X \geq k) = 1 - P(X \leq k-1)$$

En primer lugar calculamos esta suma utilizando la función de masa de probabilidad:

```
# Parámetros del problema
n <- 10
p <- 0.8

# Probabilidades P(X = 8), P(X = 9) y P(X = 10)
prob_8 <- dbinom(8, size = n, prob = p)
prob_9 <- dbinom(9, size = n, prob = p)
prob_10 <- dbinom(10, size = n, prob = p)
```

```r
# Probabilidad total P(X >= 8)
prob_total <- prob_8 + prob_9 + prob_10
prob_total
```

```
## [1] 0.6777995
```

Si usamos la funcion de distribución, `pbinom`

```r
1-pbinom (7, size = n, prob = p)
```

```
## [1] 0.6777995
```

Naturalmente ambos resultados coinciden. Obsérvese que al ser $p = 0.8$ valores altos resultan bastante probables, con lo que la

2.5 Ejercicio 2.5

En una cierta población se ha observado un número medio anual de 12 muertes por cáncer de pulmón. Si el número de muertes causadas por la enfermedad sigue una distribución de Poisson, ¿cuál es la probabilidad de que durante el año en curso: 1. haya exactamente 10 muertes por cáncer de pulmón? 2. 15 o más personas mueran a causa de la enfermedad? 3. 10 o menos personas mueran a causa de la enfermedad?

El número de muertes por cáncer de pulmón sigue una distribución de Poisson, que se usa para modelar la ocurrencia de eventos discretos dentro de un intervalo de tiempo, donde el valor esperado es proporcional al tamaño del intervalo. En este caso, el valor esperado es el número medio de muertes por año, que es 12. La función de masa de probabilidad (PMF) de una variable aleatoria X con distribución de Poisson y parámetro λ es:

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

donde k es el número de eventos, λ es el valor esperado (12 en nuestro caso) y $k!$ es el factorial de k. Usaremos este modelo para resolver los apartados.

2.5.1 Probabilidad de que haya exactamente 10 muertes

La probabilidad de observar exactamente $k = 10$ muertes se puede calcular usando la PMF de la distribución de Poisson con $\lambda = 12$:

$$P(X = 10) = \frac{12^{10} e^{-12}}{10!}$$

Podemos calcular este valor con R.

```r
lambda <- 12
k <- 10
prob_10_muertes <- dpois(k, lambda)
prob_10_muertes
```

```
## [1] 0.1048373
```

2.5.2 Probabilidad de que 15 o más personas mueran

Para obtener la probabilidad de que 15 o más personas mueran, necesitamos calcular la probabilidad acumulada de $X \geq 15$. Esto se puede obtener restando de 1 la probabilidad acumulada de $X < 15$, es decir:

$$P(X \geq 15) = 1 - P(X < 15) = 1 - P(X \leq 14)$$

Usamos la función de probabilidad acumulada (CDF) de la Poisson en R.

```
k_15 <- 14
prob_15_o_mas <- 1 - ppois(k_15, lambda)
prob_15_o_mas
```

```
## [1] 0.2279755
```

2.5.3 Probabilidad de que 10 o menos personas mueran

La probabilidad de que 10 o menos personas mueran es simplemente la probabilidad acumulada de $X \leq 10$, que se puede calcular directamente con la CDF de la distribución de Poisson.

$$P(X \leq 10)$$

Calculamos esto en R:

```
prob_10_o_menos <- ppois(k, lambda)
prob_10_o_menos
```

```
## [1] 0.3472294
```

2.5.4 Conclusión

1. La probabilidad de que haya exactamente 10 muertes es:

```
prob_10_muertes
```

```
## [1] 0.1048373
```

2. La probabilidad de que 15 o más personas mueran es:

```
prob_15_o_mas
```

```
## [1] 0.2279755
```

3. La probabilidad de que 10 o menos personas mueran es:

```
prob_10_o_menos
```

```
## [1] 0.3472294
```

2.6 Ejercicio 2.6

Los daños a los cromosomas del óvulo o del espermatozoide, pueden causar mutaciones que conducen a abortos, defectos de nacimiento, u otras deficiencias genéticas. Un estudio sobre los efectos teratogénicos de la radiación ha determinado que la probabilidad de que tal mutación se produzca por radiación es del 10%. El resto son atribuibles a otras causas. Una vez detectadas 150 mutaciones,

1. ¿cuántas se esperaría que se debiesen a radiaciones?
2. ¿Cuál es la probabilidad de que solamente 10 se debiesen a radiaciones?

Solución

Para analizar el número de mutaciones que se deben a radiaciones, podemos considera dos modelos diferentes: uno basado en la distribución binomial y otro en la distribución de Poisson.

2.6.1 Justificación del uso de distribución binomial

La distribución binomial es adecuada cuando tenemos un número fijo de ensayos independientes y cada ensayo tiene dos posibles resultados: éxito (la mutación es debida a radiación) o fracaso (la mutación no es debida a radiación). En cada ensayo, la probabilidad de éxito es constante.

Esto se ajusta perfectamente a las condiciones del problema: - Hay 150 ensayos independientes (cada mutación observada puede estar o no causada por radiación). - Cada ensayo tiene dos posibles resultados: mutación por radiación o mutación por otra causa. - La probabilidad de éxito es constante y pequeña ($p = 0.1$). Por tanto, el número de mutaciones debidas a radiación se puede modelizar bien mediante una distribución binomial $X \sim \text{Binomial}(n = 150, p = 0.1)$.

2.6.2 Justificación del uso de distribución de Poisson

La distribución de Poisson es adecuada para modelar el número de eventos raros que ocurren en un intervalo de tiempo, espacio, o cualquier otra unidad, cuando estos eventos ocurren de forma independiente y su probabilidad de ocurrencia es baja.

En este caso las "mutaciones debidas a radiación" pueden considerarse eventos raros dentro de un gran conjunto de mutaciones (150 mutaciones observadas, pero solo un 10% de ellas son debidas a radiación).

Puede considerarse además, que las mutaciones individuales pueden ocurrir de forma independiente entre sí, ya que la probabilidad de que una mutación se deba a radiación no afecta a la probabilidad de que otra mutación sea causada por radiación.

Estas condiciones son características de los *procesos de Poisson* y por tanto la distribución de Poisson es una elección natural para describir procesos en los que los eventos ocurren de manera aleatoria en un intervalo dado (por ejemplo, en un periodo de tiempo o un espacio), siempre que:

- Los eventos ocurran con una tasa promedio constante (en este caso, la tasa de mutaciones debidas a radiaciones es proporcional a la tasa global de mutaciones, multiplicada por la probabilidad $p = 0.1$).
- No haya límite teórico en el número de eventos que puedan ocurrir en un intervalo (aunque observamos un total de 150 mutaciones, teóricamente podríamos seguir detectando más mutaciones).

En el modelo de Poisson, el parámetro λ representa la tasa promedio de ocurrencia de los eventos (en este caso, mutaciones debidas a radiación). Si conocemos la tasa promedio de aparición de mutaciones por radiación ($\lambda = n \cdot p$ en el contexto binomial, pero también se puede calcular directamente si conocemos la tasa de aparición de eventos raros), entonces podemos usar directamente la distribución de Poisson para modelar el número de eventos.

En este caso, $\lambda = 150 \cdot 0.1 = 15$, que representa el número esperado de mutaciones debidas a radiación en el total observado de mutaciones.

2.6.3 Aproximación del modelo binomial por el de Poisson

La distribución de Poisson puede considerarse una aproximación de la binomial cuando el número de ensayos (n) es grande y la probabilidad de éxito (p) es pequeña. En este caso, el número esperado de éxitos, $n \cdot p$, se mantiene moderado (en este caso, $n \cdot p = 15$).

Este resultado que se conoce como *límite de Poisson* establece que si:

- n es grande (muchos ensayos),
- p es pequeño (baja probabilidad de éxito),
- el producto $n \cdot p = \lambda$ es moderado,

entonces la binomial $X \sim \text{Binomial}(n, p)$ se puede aproximar por una distribución de Poisson con parámetro $\lambda = n \cdot p$.

En este caso:

- $n = 150$ es suficientemente grande.

- $p = 0.1$ es pequeño.
- $n \cdot p = 15$, lo cual es un valor razonable para usar la aproximación de Poisson.

Por tanto, el número de mutaciones debidas a radiaciones puede aproximarse por una distribución de Poisson $X \sim \text{Poisson}(\lambda = 15)$.

2.6.4 Número esperado de mutaciones

En ambos modelos, la esperanza del número de mutaciones debidas a radiaciones es $E[X] = n \cdot p$. Esto representa el número promedio de mutaciones debidas a radiaciones. Lo calculamos:

$$E[X] = 150 \cdot 0.1 = 15$$

Por lo tanto, se espera que alrededor de 15 mutaciones se deban a radiaciones.

2.6.5 Probabilidad de que exactamente 10 mutaciones se deban a radiaciones

2.6.5.1 Usando la distribución Binomial La probabilidad de que exactamente 10 mutaciones se deban a radiaciones se puede calcular usando la PMF de la binomial:

$$P(X = 10) = \binom{150}{10} (0.1)^{10} (0.9)^{140}$$

Usando R tenemos:

```
n <- 150
p <- 0.1
k <- 10
prob_binom_10 <- dbinom(k, n, p)
prob_binom_10
```

```
## [1] 0.04591681
```

2.6.5.2 Usando la aproximación de Poisson La distribución de Poisson con $\lambda = n \cdot p = 15$ también se puede usar para aproximar esta probabilidad. La probabilidad de obtener exactamente 10 mutaciones se calcula como:

$$P(X = 10) = \frac{15^{10} e^{-15}}{10!}$$

Con R:

```
lambda <- 15
prob_pois_10 <- dpois(k, lambda)
prob_pois_10
```

```
## [1] 0.04861075
```

2.6.6 Conclusión

- Se espera que 15 de las 150 mutaciones se deban a radiaciones.
- La probabilidad de que exactamente 10 mutaciones se deban a radiaciones es:
  - Usando la distribución binomial:

```
prob_binom_10
```

```
## [1] 0.04591681
```

- Usando la aproximación de Poisson:

```
prob_pois_10
```

```
## [1] 0.04861075
```

Ambos métodos dan resultados similares, pero el modelo de Poisson es útil para simplificar los cálculos cuando el número total de mutaciones es grande y la probabilidad de cada evento es pequeña.

2.7 Ejercicio 2.7

Entre los diabéticos, el nivel de glucosa en sangre X, en ayunas, puede suponerse de distribución aproximadamente normal, con media 106mg/100ml y desviación típica 8mg/100ml, es decir : $X \sim N\left(\mu = 106, \sigma^2 = 64\right)$.

Hallar; 1. El porcentaje de diabéticos con niveles de glucosa inferiores a 120 ($P[X \leq 120]$ 2. ¿Qué porcentaje de diabéticos tienen niveles comprendidos entre 90 y 120? 3. Hallar el nivel de glucosa "p25", caracterizado por la propiedad de que el 25% de todos los diabéticos tiene un nivel de glucosa en ayunas inferior o igual a x.

SOLUCIÓN

Según el enunciado el nivel de glucosa X se distribuye según una distribución normal con media $\mu = 106$ y varianza $\sigma^2 = 64$, es decir, $X \sim N(106, 64)$, o equivalentemente $X \sim N(106, 8^2)$.

2.7.1 Porcentaje de diabéticos con niveles de glucosa inferiores a 120 ($P[X \leq 120]$)

Para calcular esta probabilidad, necesitamos estandarizar la variable X a una normal estándar $Z \sim N(0, 1)$. La fórmula de estandarización es:

$$Z = \frac{X - \mu}{\sigma}$$

Sustituyendo los valores de $\mu = 106$ y $\sigma = 8$:

$$Z = \frac{120 - 106}{8} = 1.75$$

Ahora calculamos $P(Z \leq 1.75)$, es decir, la probabilidad de que la variable estándar normal sea menor o igual que 1.75. Esta probabilidad la obtenemos a partir de la tabla de la normal estándar o usando R.

```
# Calculamos la probabilidad con la función pnorm
p1 <- pnorm(1.75)
p1
```

```
## [1] 0.9599408
```

2.7.2 Porcentaje de diabéticos con niveles de glucosa comprendidos entre 90 y 120

En este caso queremos calcular $P(90 \leq X \leq 120)$. Para hacerlo, calculamos las probabilidades individuales de $P(X \leq 120)$ y $P(X \leq 90)$, y restamos la segunda de la primera:

$$P(90 \leq X \leq 120) = P(X \leq 120) - P(X \leq 90)$$

Primero estandarizamos ambas variables:

$$Z_{120} = \frac{120 - 106}{8} = 1.75$$

$$Z_{90} = \frac{90 - 106}{8} = -2.00$$

Ahora calculamos $P(Z \leq 1.75)$ y $P(Z \leq -2.00)$ usando R.

```
# Calculamos ambas probabilidades
p2_120 <- pnorm(1.75)
p2_90 <- pnorm(-2.00)
p2 <- p2_120 - p2_90
p2
```

```
## [1] 0.9371907
```

2.7.3 Hallar el nivel de glucosa "p25"

Para encontrar el percentil 25 de la distribución, necesitamos resolver la ecuación:

$$P(X \leq p_{25}) = 0.25$$

Sabemos que $X \sim N(106, 64)$, así que estandarizamos el valor p_{25}:

$$Z_{p25} = \frac{p_{25} - 106}{8}$$

Luego, encontramos el valor de Z_{p25} que corresponde al percentil 25 de la distribución normal estándar, es decir, $P(Z \leq Z_{p25}) = 0.25$. Esto lo obtenemos con la función inversa de la distribución normal estándar.

```
# Calculamos el valor z correspondiente al percentil 25
z_p25 <- qnorm(0.25)
# Calculamos el p25 en la escala original
p25 <- 106 + z_p25 * 8
p25
```

```
## [1] 100.6041
```

2.7.4 Resumen de resultados:

1. La probabilidad de que el nivel de glucosa sea menor o igual a 120 es aproximadamente:

$$P[X \leq 120] = 0.9599$$

2. El porcentaje de diabéticos con niveles de glucosa comprendidos entre 90 y 120 es aproximadamente:

$$P[90 \leq X \leq 120] = 0.9104$$

3. El nivel de glucosa correspondiente al percentil 25, es decir, el valor p_{25}, es aproximadamente:

$$p_{25} \approx 100.61\,\text{mg/100ml}$$

2.8 Ejercicio 28

Se supone que la glucemia basal en individuos sanos, X_s sigue una distribución $X \sim N(\mu = 80, \sigma = 10)$, mientras que en los diabéticos X_d, sigue una distribución $X \sim N(\mu = 160, \sigma = 31.4)$. Si se conviene en clasificar como sanos al 2% de los diabéticos: a) ¿Por debajo de qué valor se considera sano a un individuo? ¿Cuántos sanos serán clasificados como diabéticos? b) Se sabe que en la población en general el 10% de los individuos son diabéticos ¿cuál es la probabilidad de que un individuo elegido al azar y diagnosticado como diabético, realmente lo sea?

2.9 Ejercicio 2.9

Supóngase que se van a utilizar 20 ratas en un estudio de agentes coagulantes de la sangre. Como primera experiencia, se dio un anticoagulante a 10 de ellos, pero por inadvertencia se pusieron todas sin marcas en el mismo recinto. Se necesitaron 12 ratas para la segunda fase del estudio y se le tomó al azar sin reemplazamiento. ¿Cuál es la probabilidad de que de las 12 elegidas 6 tengan la droga y 6 no la tengan?

3 Distribuciones de probabilidad multidimensionales

3.1 Ejercicio 1

Se tienen dos estudios clínicos importantes, cuyos análisis genéticos deben ser asignados aleatoriamente a uno o más de tres laboratorios, A, B y C . Denote con Y_1 el número de estudios asignados al laboratorio A y con Y_2 el número de estudios asignados al laboratorio B. Cada laboratorio puede recibir 0,1 o 2 estudios. a. Encuentre la función de probabilidad conjunta para Y_1 y Y_2. b. Encuentre $F(1, 0)$, es decir, la probabilidad de que el laboratorio A reciba como máximo un estudio y el laboratorio B no reciba ninguno.

3.1.1 Parte a: Función de probabilidad conjunta para Y_1 y Y_2

En este ejercicio, se nos indica que existen tres laboratorios (A, B y C) a los cuales se pueden asignar los estudios de forma aleatoria. Denotamos con Y_1 el número de estudios asignados al laboratorio A y con Y_2 el número de estudios asignados al laboratorio B. Cada laboratorio puede recibir entre 0 y 2 estudios.

Vamos a analizar el espacio muestral, S, que representa las posibles combinaciones de asignación de estudios a los laboratorios. Los resultados posibles son:

S	AA	AB	AC	BA	BB	BC	CA	CB	CC
(y_1, y_2)	(2,0)	(1,1)	(1,0)	(1,1)	(0,2)	(1,0)	(1,0)	(0,1)	(0,0)

Cada punto muestral es igualmente probable, con una probabilidad de $\frac{1}{9}$, ya que existen 9 combinaciones posibles.

La función de probabilidad conjunta para Y_1 y Y_2 queda entonces representada en la siguiente tabla:

	$y_1 = 0$	$y_1 = 1$	$y_1 = 2$
$y_2 = 0$	$\frac{1}{9}$	$\frac{2}{9}$	$\frac{1}{9}$
$y_2 = 1$	$\frac{2}{9}$	$\frac{2}{9}$	0
$y_2 = 2$	$\frac{1}{9}$	0	0

3.1.2 Parte b: Cálculo de $F(1, 0)$

Nos piden encontrar la probabilidad de que el laboratorio A reciba como máximo un estudio y el laboratorio B no reciba ninguno, es decir, $F(1, 0) = P(Y_1 \leq 1, Y_2 = 0)$.

Para resolverlo, sumamos las probabilidades de los eventos en los cuales $Y_1 \leq 1$ y $Y_2 = 0$, que son $(Y_1 = 0, Y_2 = 0)$ y $(Y_1 = 1, Y_2 = 0)$:

$$F(1,0) = P(Y_1 = 0, Y_2 = 0) + P(Y_1 = 1, Y_2 = 0)$$

Sustituyendo con las probabilidades correspondientes de la tabla obtenemos:

$$F(1,0) = \frac{1}{9} + \frac{2}{9} = \frac{3}{9} = \frac{1}{3}$$

3.1.3 Resumen

1. La función de probabilidad conjunta ha sido obtenida en función de todas las combinaciones posibles de asignación, considerando que cada una es igualmente probable.
2. La probabilidad solicitada, $F(1,0)$, es de $\frac{1}{3}$, que representa la probabilidad de que el laboratorio A reciba como máximo un estudio y el laboratorio B no reciba ninguno.

3.2 Ejercicio 2

Tres monedas balanceadas se lanzan en forma independiente al aire. Una de las variables de interés es Y_1, el número de caras. Denote con Y_2 la cantidad de dinero ganado en una apuesta colateral en la siguiente forma. Si la primera cara aparece en el primer tiro, usted gana 1€. Si la primera cara aparece en el tiro segundo o en el tercero gana 2€ o 3€, respectivamente. Si no aparece una cara, usted pierde 1€ (esto es, gana - 1€).

1. Encuentre la función de probabilidad conjunta para Y_1 y Y_2.
2. ¿Cuál es la probabilidad de que haya menos de tres caras y usted gane 1€ o menos? [Esto es, encuentre $F(2,1)$.

Solución

3.2.1 1. Función de probabilidad conjunta para Y_1 y Y_2

Dado que se lanzan tres monedas balanceadas de manera independiente, cada lanzamiento puede resultar en cara (C) o cruz (+) con probabilidad 0.5.

Listamos todas las posibles secuencias de resultados en los tres lanzamientos y calculamos los valores de Y_1 (número de caras obtenidas) y Y_2 (cantidad de dinero ganado) para cada caso.

Secuencia	Y_1 (Número de Caras)	Y_2 (Ganancia en €)
CCC	3	1
CC+	2	1
C+C	2	1
C++	1	1
+CC	2	2
+C+	1	2
++C	1	3
+++	0	-1

Para calcular la función de probabilidad conjunta $P(Y_1 = y_1, Y_2 = y_2)$, obtenemos las probabilidades de cada combinación de (Y_1, Y_2) a partir de la cantidad de secuencias que cumplen con esos valores específicos.

3.2.1.1 Probabilidad de cada combinación de (Y_1, Y_2):

- $P(Y_1 = 3, Y_2 = 1) = \frac{1}{8}$, secuencia: CCC
- $P(Y_1 = 2, Y_2 = 1) = \frac{2}{8}$, secuencias: CC+, C+C
- $P(Y_1 = 2, Y_2 = 2) = \frac{1}{8}$, secuencia: +CC
- $P(Y_1 = 1, Y_2 = 1) = \frac{1}{8}$, secuencia: C++
- $P(Y_1 = 1, Y_2 = 2) = \frac{1}{8}$, secuencia: +C+
- $P(Y_1 = 1, Y_2 = 3) = \frac{1}{8}$, secuencia: ++C
- $P(Y_1 = 0, Y_2 = -1) = \frac{1}{8}$, secuencia: +++

Con esto, la función de probabilidad conjunta $P(Y_1 = y_1, Y_2 = y_2)$ se define por:

$$P(Y_1 = y_1, Y_2 = y_2) = \begin{cases} \frac{1}{8}, & (y_1, y_2) = (3,1), (2,1), (2,2), (1,1), (1,2), (1,3), (0,-1) \\ 0, & \text{en cualquier otro caso} \end{cases}$$

3.2.2 2. Cálculo de $F(2,1)$

Buscamos $F(2,1) = P(Y_1 \leq 2, Y_2 \leq 1)$, es decir, la probabilidad de obtener menos de tres caras y ganar un máximo de 1€.

Para calcular esta probabilidad, sumamos $P(Y_1 = y_1, Y_2 = y_2)$ para todos los pares (y_1, y_2) que cumplen $Y_1 \leq 2$ y $Y_2 \leq 1$. De la tabla anterior, vemos que cumplen esta condición las siguientes combinaciones:

- $(Y_1 = 2, Y_2 = 1)$ con probabilidad $\frac{2}{8}$
- $(Y_1 = 1, Y_2 = 1)$ con probabilidad $\frac{1}{8}$
- $(Y_1 = 0, Y_2 = -1)$ con probabilidad $\frac{1}{8}$

Entonces:

$$F(2,1) = P(Y_1 \leq 2, Y_2 \leq 1) = \frac{2}{8} + \frac{1}{8} + \frac{1}{8} = \frac{4}{8} = \frac{1}{2}$$

3.2.3 Resumen de resultados

1. La función de probabilidad conjunta $P(Y_1 = y_1, Y_2 = y_2)$ se ha especificado para todos los valores posibles.
2. La probabilidad de que haya menos de tres caras y se gane 1€ o menos es $F(2,1) = \frac{1}{2}$.

3.3 Ejercicio 3

En el Ejercicio 1 determinamos que la distribución conjunta de Y_1, el número de análisis asignados al laboratorio A , y Y_2, el número de análisis asignados al laboratorio B , está dada por las entradas en la siguiente tabla.

y_1			
y_2	0	1	2
0	1/9	2/9	1/9
1	2/9	2/9	0
2	1/9	0	0

a. Encuentre la distribución de probabilidad marginal de Y_1.
b. De acuerdo con los resultados vistos anteriormente Y_1 tiene una distribución binomial con $n = 2$ y $p = 1/3$. ¿Hay algún conflicto entre este resultado y la respuesta dada en el punto a?

Solución

3.3.1 Parte a: Distribución de probabilidad marginal de Y_1

Para encontrar la distribución marginal de Y_1, debemos sumar las probabilidades conjuntas de Y_1 y Y_2 para cada valor posible de Y_1. A continuación, mostramos la tabla de probabilidades conjuntas de la solución anterior:

$y_1 \backslash y_2$	0	1	2
$y_1 = 0$	1/9	2/9	1/9
$y_1 = 1$	2/9	2/9	0
$y_1 = 2$	1/9	0	0

La distribución marginal de Y_1 se calcula sumando las probabilidades de cada fila (fijando y_1 y sumando sobre los valores de y_2):

1. Para $y_1 = 0$:
$$P(Y_1 = 0) = \frac{1}{9} + \frac{2}{9} + \frac{1}{9} = \frac{4}{9}$$
2. Para $y_1 = 1$:
$$P(Y_1 = 1) = \frac{2}{9} + \frac{2}{9} + 0 = \frac{4}{9}$$
3. Para $y_1 = 2$:
$$P(Y_1 = 2) = \frac{1}{9} + 0 + 0 = \frac{1}{9}$$

Por lo tanto, la distribución marginal de Y_1 es:

y_1	0	1	2
$p_{Y_1}(y_1)$	$\frac{4}{9}$	$\frac{4}{9}$	$\frac{1}{9}$

3.3.2 Parte b: Comparación con la distribución binomial

Se sugiere que Y_1 sigue una distribución binomial con parámetros $n = 2$ y $p = \frac{1}{3}$. Veamos si esta afirmación concuerda con los resultados obtenidos en la parte a.

Para una variable aleatoria binomial $Y_1 \sim \text{Binomial}(n = 2, p = 1/3)$, la función de probabilidad es:

$$P(Y_1 = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Calculamos las probabilidades para cada valor de k:

1. Para $k = 0$:
$$P(Y_1 = 0) = \binom{2}{0} \left(\frac{1}{3}\right)^0 \left(\frac{2}{3}\right)^2 = 1 \cdot 1 \cdot \frac{4}{9} = \frac{4}{9}$$
2. Para $k = 1$:
$$P(Y_1 = 1) = \binom{2}{1} \left(\frac{1}{3}\right)^1 \left(\frac{2}{3}\right)^1 = 2 \cdot \frac{1}{3} \cdot \frac{2}{3} = \frac{4}{9}$$
3. Para $k = 2$:
$$P(Y_1 = 2) = \binom{2}{2} \left(\frac{1}{3}\right)^2 \left(\frac{2}{3}\right)^0 = 1 \cdot \frac{1}{9} \cdot 1 = \frac{1}{9}$$

Las probabilidades obtenidas para la binomial son exactamente las mismas que encontramos en la distribución marginal de Y_1, lo que confirma que $Y_1 \sim \text{Binomial}(2, 1/3)$.

3.3.3 Conclusión

No hay conflicto entre la distribución marginal de Y_1 obtenida en la parte a y el hecho de que Y_1 tenga una distribución binomial con parámetros $n = 2$ y $p = \frac{1}{3}$.

3.4 Ejercicio 4

Un ingeniero ambiental mide la cantidad (en peso) de partículas contaminantes en muestras de aire de cierto volumen recolectado en dos chimeneas en una planta de energía alimentada con carbón. Una de las chimeneas está equipada con un aparato limpiador. Denote con Y_1 la cantidad de contaminante por muestra recolectada arriba de la chimenea que no tiene aparato limpiador y denote con Y_2 la cantidad de contaminante por muestra recolectada arriba de la chimenea que está equipada con el aparato limpiador.

Suponga que el comportamiento de frecuencia relativa de Y_1 y Y_2 puede ser modelado por

$$f(y_1, y_2) = \begin{cases} k, & 0 \leq y_1 \leq 2, \quad 0 \leq y_2 \leq 1, \quad 2y_2 \leq y_1 \\ 0, & \text{en cualquier otro punto.} \end{cases}$$

Esto es, Y_1 y Y_2 están uniformemente distribuidas sobre la región dentro del triángulo limitado por $y_1 = 2, y_2 = 0$ y $2y_2 = y_1$.

1. Encuentre el valor de k que haga de ésta una función de densidad de probabilidad.
2. Encuentre $P(Y_1 \geq 3Y_2)$. Esto es, encuentre la probabilidad de que el aparato limpiador reduzca la cantidad de contaminante en un tercio o más.

Solución

3.4.1 1. Encontrar el valor de k que haga de ésta una función de densidad de probabilidad

Para que $f(y_1, y_2)$ sea una función de densidad de probabilidad válida, la integral de $f(y_1, y_2)$ sobre toda la región de soporte debe ser igual a 1:

$$\iint_{\text{región de soporte}} f(y_1, y_2)\, dy_1\, dy_2 = 1$$

La función de densidad es constante y toma el valor k sobre la región triangular definida por $0 \leq y_1 \leq 2$, $0 \leq y_2 \leq 1$ y $2y_2 \leq y_1$. Esta región corresponde a un triángulo en el plano y_1-y_2 con los vértices en $(0, 0)$, $(2, 0)$, y $(2, 1)$.

3.4.1.1 Paso 1: Determinar el área de la región triangular La región triangular tiene una base de longitud 2 (a lo largo del eje y_1) y una altura de 1 (a lo largo del eje y_2). El área del triángulo es:

$$\text{Área} = \frac{1}{2} \times \text{base} \times \text{altura} = \frac{1}{2} \times 2 \times 1 = 1$$

3.4.1.2 Paso 2: Integrar $f(y_1, y_2)$ sobre la región triangular Dado que $f(y_1, y_2) = k$ en esta región y la función de densidad es uniforme, la integral sobre esta área es simplemente k multiplicado por el área:

$$\iint_{\text{región}} f(y_1, y_2)\, dy_1\, dy_2 = k \times \text{Área} = k \times 1 = k$$

Para que $f(y_1, y_2)$ sea una función de densidad, necesitamos que esta integral sea igual a 1, entonces:

$$k = 1$$

3.4.2 2. Encontrar $P(Y_1 \geq 3Y_2)$

Queremos encontrar la probabilidad de que $Y_1 \geq 3Y_2$ en la región triangular donde $f(y_1, y_2) = k = 1$.

3.4.2.1 Paso 1: Identificar la subregión definida por $Y_1 \geq 3Y_2$ La desigualdad $Y_1 \geq 3Y_2$ corresponde a la recta $y_1 = 3y_2$. Para encontrar la intersección de esta recta con la región triangular, notamos que: - En $y_1 = 2$, al sustituir en $y_1 = 3y_2$, tenemos $y_2 = \frac{2}{3}$.

Por lo tanto, la subregión de interés es el triángulo delimitado por los puntos $(0, 0)$, $(2, 0)$ y $(2, \frac{2}{3})$.

3.4.2.2 Paso 2: Calcular el área de la subregión La base de este subtriángulo es 2 (a lo largo de y_1), y la altura es $\frac{2}{3}$ (a lo largo de y_2). Su área es:

$$\text{Área del subtriángulo} = \frac{1}{2} \times \text{base} \times \text{altura} = \frac{1}{2} \times 2 \times \frac{2}{3} = \frac{2}{3}$$

3.4.2.3 Paso 3: Calcular la probabilidad La probabilidad buscada es la proporción del área del subtriángulo respecto al área total de la región de soporte:

$$P(Y_1 \geq 3Y_2) = \frac{\text{Área del subtriángulo}}{\text{Área de la región total}} = \frac{\frac{2}{3}}{1} = \frac{2}{3}$$

3.4.3 Respuesta final

1. El valor de k que hace de $f(y_1, y_2)$ una función de densidad de probabilidad es $k = 1$.
2. La probabilidad de que $Y_1 \geq 3Y_2$ es:

$$P(Y_1 \geq 3Y_2) = \frac{2}{3}$$

3.5 Ejercicio 5

En el Ejercicio 4 hemos establecido que

$$f\left(y_1, y_2\right) = \begin{cases} k, & 0 \leq y_1 \leq 2, \quad 0 \leq y_2 \leq 1, \quad 2y_2 \leq y_1 \\ 0, & \text{en cualquier otro punto.} \end{cases}$$

es una función de densidad de probabilidad conjunta válida para Y_1, la cantidad de contaminante por muestra recolectada arriba de la chimenea que no tenía el aparato limpiador, y para Y_2, la cantidad recolectada arriba de la chimenea con el aparato limpiador.

1. Si consideramos la chimenea con el limpiador instalado, encuentre la probabilidad de que la cantidad de contaminante en una muestra determinada sea mayor que 0.5.

b. Dado que se observa que la cantidad de contaminante en una muestra tomada arriba de la chimenea con el limpiador es 0.5 , encuentre la probabilidad de que la cantidad de contaminante exceda de 1.5 arriba de la otra chimenea (la que no tiene limpiador).

Solución

3.5.1 1. Probabilidad de que la cantidad de contaminante en la chimenea con limpiador sea mayor que 0.5

Queremos encontrar la probabilidad de que $Y_2 > 0.5$, donde Y_2 representa la cantidad de contaminante en la chimenea con el aparato limpiador. Recordando que $f(y_1, y_2)$ es la función de densidad conjunta y que

hemos hallado $k = 1$, calcularemos la probabilidad integrando sobre la región correspondiente a $y_2 > 0.5$ y y_1 dentro de su rango definido por la condición $2y_2 \leq y_1 \leq 2$.

La probabilidad se calcula como:

$$P(Y_2 > 0.5) = \iint_{\{y_2 > 0.5, 2y_2 \leq y_1 \leq 2\}} f(y_1, y_2)\, dy_1\, dy_2$$

3.5.1.1 Paso 1: Establecer los límites de integración

Para $y_2 > 0.5$, los límites de y_1 están restringidos por la región triangular dada: - $2y_2 \leq y_1 \leq 2$

Entonces, los límites de integración son: - y_2: desde 0.5 hasta 1 - y_1: desde $2y_2$ hasta 2

3.5.1.2 Paso 2: Integrar la función de densidad conjunta

Dado que $f(y_1, y_2) = 1$ en esta región, la probabilidad es la integral de 1 sobre el área triangular correspondiente:

$$P(Y_2 > 0.5) = \int_{0.5}^{1} \int_{2y_2}^{2} 1\, dy_1\, dy_2$$

Evaluamos esta integral en dos pasos.

```r
## Calculo de la probabilidad con R

## Definir la función de integración para y1
integrate_y1 <- function(y2) {
  integrate(function(y1) 1, lower = 2 * y2, upper = 2)$value
}

## Integrar respecto a y2
result <- integrate(function(y2) integrate_y1(y2), lower = 0.5, upper = 1)
result$value
```

3.5.1.3 Resultado

Al resolver esta integral, obtenemos la probabilidad:

$$P(Y_2 > 0.5) \approx 0.25$$

3.5.2 2. Probabilidad condicional dada una observación de $Y_2 = 0.5$

Queremos calcular $P(Y_1 > 1.5 \mid Y_2 = 0.5)$, la probabilidad de que la cantidad de contaminante en la chimenea sin limpiador (representada por Y_1) sea mayor que 1.5, dado que en la chimenea con limpiador se observó $Y_2 = 0.5$.

3.5.2.1 Paso 1: Identificar la función de densidad condicional

Para la función de densidad condicional $f_{Y_1|Y_2}(y_1|y_2)$, aplicamos:

$$f_{Y_1|Y_2}(y_1|Y_2 = 0.5) = \frac{f(y_1, 0.5)}{f_{Y_2}(0.5)}$$

Calcularemos $f_{Y_2}(0.5)$ integrando $f(y_1, y_2)$ sobre los valores de y_1 en la región donde $Y_2 = 0.5$:

$$f_{Y_2}(0.5) = \int_{y_1 = 2 \cdot 0.5}^{2} f(y_1, 0.5)\, dy_1 = \int_{1}^{2} 1\, dy_1$$

Evaluamos la integral:

$$f_{Y_2}(0.5) = \int_1^2 1\, dy_1 = (2-1) = 1$$

Por lo tanto, $f_{Y_2}(0.5) = 1$.

3.5.2.2 Paso 2: Calcular la probabilidad condicional

La probabilidad de interés es:

$$P(Y_1 > 1.5 | Y_2 = 0.5) = \int_{1.5}^2 f_{Y_1|Y_2}(y_1|Y_2 = 0.5)\, dy_1$$

Como $f_{Y_1|Y_2}(y_1|Y_2 = 0.5) = 1$ para $1 \leq y_1 \leq 2$, tenemos:

$$P(Y_1 > 1.5 | Y_2 = 0.5) = \int_{1.5}^2 1\, dy_1 = 2 - 1.5 = 0.5$$

3.5.3 Respuesta final

1. La probabilidad de que la cantidad de contaminante en la chimenea con limpiador sea mayor que 0.5 es aproximadamente 0.25.
2. La probabilidad de que la cantidad de contaminante en la chimenea sin limpiador exceda 1.5, dado que en la otra chimenea se observó 0.5, es 0.5.

3.6 Ejercicio 6

En el ejercicio 1 determinamos que la distribución conjunta de Y_1, el número de análisis asignados al laboratorio A, y Y_2, el número de análisis asignados al laboratorio B , está dada por las entradas en la siguiente tabla.

	y_1		
y_2	0	1	2
0	1/9	2/9	1/9
1	2/9	2/9	0
2	1/9	0	0

a. Encuentre $\mathrm{Cov}\,(Y_1, Y_2)$. $_i$
b. Le sorprende que $\mathrm{Cov}\,(Y_1, Y_2)$ sea negativa? ¿Por qué?

Solución

3.6.1 Parte a: Cálculo de la covarianza $\mathrm{Cov}(Y_1, Y_2)$

La covarianza entre dos variables aleatorias Y_1 y Y_2 se define como:

$$\mathrm{Cov}(Y_1, Y_2) = E(Y_1 Y_2) - E(Y_1)E(Y_2)$$

Para calcular la covarianza, necesitamos obtener los valores de $E(Y_1)$, $E(Y_2)$ y $E(Y_1 Y_2)$.

1. **Cálculo de $E(Y_1)$ y $E(Y_2)$**

   A partir de la distribución conjunta dada en la tabla, podemos calcular la esperanza de Y_1 y Y_2 sumando las posibles combinaciones de valores ponderadas por sus probabilidades.

$$E(Y_1) = \sum_{y_1} y_1 \cdot P(Y_1 = y_1)$$

$$E(Y_2) = \sum_{y_2} y_2 \cdot P(Y_2 = y_2)$$

   Usamos la tabla para estos cálculos:

```
## Probabilidades conjuntas
probs <- matrix(c(1/9, 2/9, 1/9, 2/9, 2/9, 0, 1/9, 0, 0), nrow = 3, byrow = TRUE)

## Valores de Y1 y Y2
y1_values <- 0:2
y2_values <- 0:2

## Esperanza de Y1
E_Y1 <- sum(y1_values * rowSums(probs))
## Esperanza de Y2
E_Y2 <- sum(y2_values * colSums(probs))

E_Y1
```

```
## [1] 0.6666667
```

```
E_Y2
```

```
## [1] 0.6666667
```

2. **Cálculo de $E(Y_1Y_2)$**

   Para calcular $E(Y_1Y_2)$, sumamos el producto $y_1 \cdot y_2$ ponderado por la probabilidad conjunta $p(y_1, y_2)$.

```
## Producto de Y1 * Y2 * probabilidad conjunta
E_Y1Y2 <- sum(outer(y1_values, y2_values, "*") * probs)

E_Y1Y2
```

```
## [1] 0.2222222
```

3. **Calcular la covarianza**

   Sustituyendo los valores obtenidos:

$$\text{Cov}(Y_1, Y_2) = E(Y_1Y_2) - E(Y_1)E(Y_2)$$

```
Cov_Y1Y2 <- E_Y1Y2 - E_Y1 * E_Y2
Cov_Y1Y2
```

```
## [1] -0.2222222
```

3.6.2 Parte b: Interpretación de la covarianza negativa

La covarianza calculada es negativa. Este resultado tiene sentido en el contexto del problema. Dado que los estudios se asignan a tres laboratorios y cada laboratorio recibe un número limitado de estudios, un

incremento en el número de estudios asignados a un laboratorio reduce la cantidad disponible para los otros. Así, si Y_1 aumenta, Y_2 tiende a disminuir, lo que explica una relación inversa entre ambas variables y resulta en una covarianza negativa.

Solución

3.6.3 Ejercicio

3.6.3.1 1. Probabilidad de que la cantidad de contaminante en la chimenea con limpiador sea mayor que 0.5 Queremos encontrar la probabilidad de que $Y_2 > 0.5$, donde Y_2 representa la cantidad de contaminante en la chimenea con el aparato limpiador. Recordando que $f(y_1, y_2)$ es la función de densidad conjunta y que hemos hallado $k = 1$, calcularemos la probabilidad integrando sobre la región correspondiente a $y_2 > 0.5$ y y_1 dentro de su rango definido por la condición $2y_2 \leq y_1 \leq 2$.

La probabilidad se calcula como:

$$P(Y_2 > 0.5) = \iint_{\{y_2 > 0.5, 2y_2 \leq y_1 \leq 2\}} f(y_1, y_2)\, dy_1\, dy_2$$

3.6.3.1.1 Paso 1: Establecer los límites de integración Para $y_2 > 0.5$, los límites de y_1 están restringidos por la región triangular dada: - $2y_2 \leq y_1 \leq 2$

Entonces, los límites de integración son: - y_2: desde 0.5 hasta 1 - y_1: desde $2y_2$ hasta 2

3.6.3.1.2 Paso 2: Integrar la función de densidad conjunta Dado que $f(y_1, y_2) = 1$ en esta región, la probabilidad es la integral de 1 sobre el área triangular correspondiente:

$$P(Y_2 > 0.5) = \int_{0.5}^{1} \int_{2y_2}^{2} 1\, dy_1\, dy_2$$

Evaluamos esta integral en dos pasos.

```r
## Calculo de la probabilidad con R

## Definir la función de integración para y1
integrate_y1 <- function(y2) {
  integrate(function(y1) 1, lower = 2 * y2, upper = 2)$value
}

## Integrar respecto a y2
result <- integrate(function(y2) integrate_y1(y2), lower = 0.5, upper = 1)
result$value
```

3.6.3.1.3 Resultado Al resolver esta integral, obtenemos la probabilidad:

$$P(Y_2 > 0.5) \approx 0.25$$

3.6.3.2 2. Probabilidad condicional dada una observación de $Y_2 = 0.5$ Queremos calcular $P(Y_1 > 1.5 \mid Y_2 = 0.5)$, la probabilidad de que la cantidad de contaminante en la chimenea sin limpiador (representada por Y_1) sea mayor que 1.5, dado que en la chimenea con limpiador se observó $Y_2 = 0.5$.

3.6.3.2.1 Paso 1: Identificar la función de densidad condicional Para la función de densidad condicional $f_{Y_1|Y_2}(y_1|y_2)$, aplicamos:

$$f_{Y_1|Y_2}(y_1|Y_2 = 0.5) = \frac{f(y_1, 0.5)}{f_{Y_2}(0.5)}$$

Calcularemos $f_{Y_2}(0.5)$ integrando $f(y_1, y_2)$ sobre los valores de y_1 en la región donde $Y_2 = 0.5$:

$$f_{Y_2}(0.5) = \int_{y_1=2\cdot 0.5}^{2} f(y_1, 0.5)\, dy_1 = \int_1^2 1\, dy_1$$

Evaluamos la integral:

$$f_{Y_2}(0.5) = \int_1^2 1\, dy_1 = (2 - 1) = 1$$

Por lo tanto, $f_{Y_2}(0.5) = 1$.

3.6.3.2.2 Paso 2: Calcular la probabilidad condicional La probabilidad de interés es:

$$P(Y_1 > 1.5|Y_2 = 0.5) = \int_{1.5}^{2} f_{Y_1|Y_2}(y_1|Y_2 = 0.5)\, dy_1$$

Como $f_{Y_1|Y_2}(y_1|Y_2 = 0.5) = 1$ para $1 \leq y_1 \leq 2$, tenemos:

$$P(Y_1 > 1.5|Y_2 = 0.5) = \int_{1.5}^{2} 1\, dy_1 = 2 - 1.5 = 0.5$$

3.6.3.3 Respuesta final

1. La probabilidad de que la cantidad de contaminante en la chimenea con limpiador sea mayor que 0.5 es aproximadamente 0.25.
2. La probabilidad de que la cantidad de contaminante en la chimenea sin limpiador exceda 1.5, dado que en la otra chimenea se observó 0.5, es 0.5.

3.7 Ejercicio 7

Las variables aleatorias Y_1 y Y_2 son tales que $E(Y_1) = 4, E(Y_2) = -1, V(Y_1) = 2$ y $V(Y_2) = 8$.

1. ¿Cuál es $\text{Cov}(Y_1, Y_1)$?
2. Suponiendo que las medias y las varianzas sean correctas, ¿es posible que $\text{Cov}(Y_1, Y_2) = 7$? [Sugerencia: si $\text{Cov}(Y_1, Y_2) = 7$, ¿cuál es el valor de ρ, el coeficiente de correlación?]
3. Suponiendo que las medias y las varianzas sean correctas, ¿cuál es el máximo valor posible para $\text{Cov}(Y_1, Y_2)$? Si $\text{Cov}(Y_1, Y_2)$ alcanza este valor máximo, ¿qué implica eso acerca de la relación entre Y_1 y Y_2 ?

Solución

3.7.1 Parte a: Cálculo de $\text{Cov}(Y_1, Y_1)$

Para cualquier variable aleatoria Y, la covarianza de Y consigo misma es igual a su varianza. Es decir:

$$\text{Cov}(Y_1, Y_1) = V(Y_1)$$

Dado que $V(Y_1) = 2$, tenemos que:

$$\text{Cov}(Y_1, Y_1) = 2$$

3.7.2 Parte b: Verificación de $\text{Cov}(Y_1, Y_2) = 7$

Si se supone que $\text{Cov}(Y_1, Y_2) = 7$, podemos calcular el coeficiente de correlación ρ, que se define como:

$$\rho = \frac{\text{Cov}(Y_1, Y_2)}{\sqrt{V(Y_1) \cdot V(Y_2)}}$$

Sustituyendo los valores dados:

$$\rho = \frac{7}{\sqrt{2 \cdot 8}} = \frac{7}{4} = 1.75$$

Dado que el coeficiente de correlación ρ debe estar en el rango de $-1 \leq \rho \leq 1$, obtener $\rho = 1.75$ es imposible. Esto implica que no es posible que $\text{Cov}(Y_1, Y_2) = 7$ con los valores de varianza y media proporcionados.

3.7.3 Parte c: Valor máximo posible de $\text{Cov}(Y_1, Y_2)$ y su interpretación

Para determinar el valor máximo posible de $\text{Cov}(Y_1, Y_2)$, consideramos que el valor absoluto del coeficiente de correlación ρ puede ser como máximo 1. Esto ocurre en los casos de correlación lineal perfecta (positiva o negativa). Entonces, el valor máximo de $\text{Cov}(Y_1, Y_2)$ es:

$$\text{Cov}(Y_1, Y_2) = \rho \cdot \sqrt{V(Y_1) \cdot V(Y_2)}$$

Si $\rho = 1$, lo cual indica una asociación lineal positiva perfecta, obtenemos:

$$\text{Cov}(Y_1, Y_2) = 1 \cdot \sqrt{2 \cdot 8} = 4$$

Esto significa que el valor máximo posible de $\text{Cov}(Y_1, Y_2)$ es 4. De manera similar, si $\rho = -1$, el valor mínimo posible de $\text{Cov}(Y_1, Y_2)$ sería -4, indicando una asociación lineal negativa perfecta.

Cuando $\text{Cov}(Y_1, Y_2) = 4$, significa que Y_1 y Y_2 están perfectamente correlacionadas en forma positiva, es decir, existe una relación lineal exacta en la que un aumento en Y_1 siempre corresponde a un aumento proporcional en Y_2.

3.8 Ejercicio 8

Un experimento de aprendizaje requiere que una rata corra por un laberinto (una red de pasillos) hasta que localice una de tres posibles salidas. La salida 1 presenta una recompensa de alimento, no así las salidas 2 y 3. (Si la rata finalmente selecciona la salida 1 casi siempre, puede tener lugar el aprendizaje.) Denote con Y_i el número de veces que la salida i es seleccionada en corridas sucesivas. Para lo siguiente, suponga que la rata escoge una salida aleatoriamente en cada corrida.

1. Encuentre la probabilidad de que $n = 6$ corridas resulte en $Y_1 = 3, Y_2 = 1$ y $Y_3 = 2$.
2. Para n general, encuentre $E(Y_1)$ y $V(Y_1)$.

3. Encuentre $\text{Cov}\,(Y_2, Y_3)$ para n general.
4. Para comprobar la preferencia de la rata entre las salidas 2 y 3 , podemos buscar en $Y_2 - Y_3$. Encuentre $E\,(Y_2 - Y_3)$ y $V\,(Y_2 - Y_3)$ para n general.

Solución

3.8.1 Parte a: Probabilidad de obtener $Y_1 = 3$, $Y_2 = 1$ y $Y_3 = 2$ en $n = 6$ corridas

Este problema puede resolverse utilizando la **distribución multinomial**, ya que describe el número de veces que ocurre cada posible resultado en un número fijo de ensayos independientes. Aquí, cada una de las tres salidas tiene la misma probabilidad de ser seleccionada en cada corrida, es decir, $p_1 = p_2 = p_3 = \frac{1}{3}$, y el número total de corridas es $n = 6$. Por lo tanto, estamos interesados en calcular:

$$P(Y_1 = 3, Y_2 = 1, Y_3 = 2) = \frac{6!}{3!\,1!\,2!} \left(\frac{1}{3}\right)^6$$

Calculando esta expresión:

```
n <- 6
p <- 1 / 3
prob <- factorial(n) / (factorial(3) * factorial(1) * factorial(2)) * p^n
prob
```

```
## [1] 0.08230453
```

El resultado de esta probabilidad es aproximadamente 0.0823.

3.8.2 Parte b: Esperanza y varianza de Y_1 para un n general

Para una variable aleatoria multinomial, la esperanza y la varianza de cada conteo se puede calcular usando las propiedades de esta distribución. En particular, para Y_1, tenemos que:

- La **esperanza** es $E(Y_1) = n \cdot p_1$.
- La **varianza** es $V(Y_1) = n \cdot p_1 \cdot (1 - p_1)$.

Como $p_1 = \frac{1}{3}$, obtenemos:

$$E(Y_1) = \frac{n}{3}, \quad V(Y_1) = n \cdot \frac{1}{3} \cdot \frac{2}{3} = \frac{2n}{9}$$

3.8.3 Parte c: Covarianza entre Y_2 y Y_3 para un n general

La **covarianza** entre dos variables en una distribución multinomial, Y_2 y Y_3, se puede calcular como:

$$\text{Cov}(Y_2, Y_3) = -n \cdot p_2 \cdot p_3$$

Dado que $p_2 = p_3 = \frac{1}{3}$, tenemos:

$$\text{Cov}(Y_2, Y_3) = -n \cdot \frac{1}{3} \cdot \frac{1}{3} = -\frac{n}{9}$$

3.8.4 Parte d: Esperanza y varianza de $Y_2 - Y_3$ para un n general

Para evaluar la preferencia de la rata entre las salidas 2 y 3, podemos observar la diferencia $Y_2 - Y_3$. Calcularemos su esperanza y varianza.

1. **Esperanza de $Y_2 - Y_3$**:

   Usamos la linealidad de la esperanza:

$$E(Y_2 - Y_3) = E(Y_2) - E(Y_3) = \frac{n}{3} - \frac{n}{3} = 0$$

2. **Varianza de $Y_2 - Y_3$**:

   Para calcular la varianza de $Y_2 - Y_3$, usamos la fórmula de la varianza de una diferencia:

$$V(Y_2 - Y_3) = V(Y_2) + V(Y_3) - 2\,\mathrm{Cov}(Y_2, Y_3)$$

   Sabemos que $V(Y_2) = V(Y_3) = \frac{2n}{9}$ y que $\mathrm{Cov}(Y_2, Y_3) = -\frac{n}{9}$, por lo que:

$$V(Y_2 - Y_3) = \frac{2n}{9} + \frac{2n}{9} - 2 \cdot \left(-\frac{n}{9}\right) = \frac{6n}{9} = \frac{2n}{3}$$

3.8.5 Resumen de Resultados

1. La probabilidad de observar $Y_1 = 3$, $Y_2 = 1$ y $Y_3 = 2$ en 6 corridas es 0.0823.
2. La esperanza y varianza de Y_1 para n corridas son $E(Y_1) = \frac{n}{3}$ y $V(Y_1) = \frac{2n}{9}$.
3. La covarianza entre Y_2 y Y_3 es $\mathrm{Cov}(Y_2, Y_3) = -\frac{n}{9}$.
4. La esperanza y varianza de $Y_2 - Y_3$ son $E(Y_2 - Y_3) = 0$ y $V(Y_2 - Y_3) = \frac{2n}{3}$.

4 Muestreo y Distribuciones en el Muestreo

4.1 Ejercicio 1

Un guardabosque, que estudia los efectos de la fertilización en ciertos bosques de pinos en el sureste, está interesado en estimar el promedio de área de la base de los pinos. Al estudiar áreas basales de pinos similares durante muchos años, descubrió que estas mediciones (en pulgadas cuadradas) están distribuidas normalmente con desviación estándar aproximaa de 4 pulgadas cuadradas.

Encuentre la probabilidad de que la media muestral se encuentre a no más de 2 pulgadas cuadradas de la media poblacional si se muestrean $n = 9$ árboles

4.1.1 Solución

El problema indica que las áreas basales de los pinos están distribuidas normalmente con desviación estándar conocida ($\sigma = 4$ pulgadas cuadradas). Deseamos calcular la probabilidad de que la media muestral de $n = 9$ árboles difiera en no más de 2 pulgadas cuadradas de la media poblacional μ. Esto implica que queremos encontrar la probabilidad:

$$P(|\bar{X} - \mu| \leq 2)$$

o, de manera equivalente:

$$P(\mu - 2 \leq \bar{X} \leq \mu + 2)$$

Dado que la distribución de las áreas basales de los pinos es normal, la distribución de la media muestral $\bar{X}$ también es normal, con media igual a μ y desviación estándar igual a $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$. En este caso:

$$\sigma_{\bar{X}} = \frac{4}{\sqrt{9}} = \frac{4}{3} \approx 1.333$$

La probabilidad se puede reescribir en términos de la variable tipificada:

$$P(\mu - 2 \leq \bar{X} \leq \mu + 2) = P\left(\frac{\mu - 2 - \mu}{\sigma_{\bar{X}}} \leq Z \leq \frac{\mu + 2 - \mu}{\sigma_{\bar{X}}}\right)$$

Esto simplifica a:

$$P\left(-\frac{2}{\sigma_{\bar{X}}} \leq Z \leq \frac{2}{\sigma_{\bar{X}}}\right)$$

Sustituyendo $\sigma_{\bar{X}} = 1.333$, tenemos:

$$P\left(-\frac{2}{1.333} \leq Z \leq \frac{2}{1.333}\right) = P(-1.5 \leq Z \leq 1.5)$$

La probabilidad de un intervalo para Z en una distribución normal estándar se obtiene mediante la función de distribución acumulada $F_Z(z)$. El cálculo es:

$$P(-1.5 \leq Z \leq 1.5) = F_Z(1.5) - F_Z(-1.5)$$

Dado que $F_Z(-1.5) = 1 - F_Z(1.5)$ debido a la simetría de la distribución normal estándar:

$$P(-1.5 \leq Z \leq 1.5) = 2 \cdot F_Z(1.5) - 1$$

Procedemos a calcular $F_Z(1.5)$ en R.

```r
# Cálculo de la probabilidad
p_upper <- pnorm(1.5) # CDF para Z = 1.5
p_interval <- 2 * p_upper - 1
p_interval
```

```
## [1] 0.8663856
```

El resultado obtenido indica que, si se selecciona una muestra aleatoria de 9 árboles, la probabilidad de que la media muestral $\bar{X}$ se encuentre dentro de 2 pulgadas cuadradas de la media poblacional μ. La probabilidad calculada es aproximadamente:

$$P(-1.5 \leq Z \leq 1.5) \approx 0.8664$$

4.2 Ejercicio 2

Suponga que al guardabosque del 1 le gustaría que la media muestral estuviera a no más de 1 pulgada cuadrada de la media poblacional, con probabilidad 90%. ¿Cuántos árboles debe medir para asegurar este grado de precisión?

4.2.1 Solución

El guardabosque desea encontrar el tamaño de la muestra n necesario para que la probabilidad de que la media muestral $\bar{X}$ esté a no más de 1 pulgada cuadrada de la media poblacional μ sea al menos del 90%. Esto significa que queremos garantizar que:

$$P(|\bar{X} - \mu| \leq 1) \geq 0.90$$

Reescribiendo la probabilidad:

$$P(\mu - 1 \leq \bar{X} \leq \mu + 1) \geq 0.90$$

La media muestral $\bar{X}$ sigue una distribución normal con media μ y desviación estándar $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$. Esto nos permite transformar la probabilidad a una escala estándar:

$$P(\mu - 1 \leq \bar{X} \leq \mu + 1) = P\left(-\frac{1}{\sigma_{\bar{X}}} \leq Z \leq \frac{1}{\sigma_{\bar{X}}}\right)$$

Donde Z es la variable normal "estándar", $N(0, 1)$.. Sustituyendo $\sigma_{\bar{X}} = \frac{4}{\sqrt{n}}$, la probabilidad se convierte en:

$$P\left(-\frac{1 \cdot \sqrt{n}}{4} \leq Z \leq \frac{1 \cdot \sqrt{n}}{4}\right) \geq 0.90$$

Sea z^* el valor crítico de la distribución normal estándar tal que $P(-z^* \leq Z \leq z^*) = 0.90$. Esto implica que $z^* = F_Z^{-1}(0.95)$, ya que 90% de la probabilidad está centrada simétricamente, dejando 5% en cada cola.

El intervalo estándar nos lleva a:

$$\frac{\sqrt{n}}{4} = z^*$$

Resolviendo para n:

$$n = (4z^*)^2$$

Usaremos R para calcular z^* y el tamaño de la muestra.

```
# Cálculo de z* y tamaño de muestra
z_star <- qnorm(0.95) # Valor crítico para 90% de probabilidad centrada
n <- (4 * z_star)^2
z_star
```

```
## [1] 1.644854
```

```
n
```

```
## [1] 43.2887
```

4.2.2 Resultado

El valor crítico z^* es aproximadamente:

$$z^* \approx 1.645$$

Sustituyendo en la fórmula para n:

$$n = (4 \cdot 1.645)^2 = 43.29$$

Como el tamaño de muestra debe ser un número entero, redondeamos hacia arriba:

$$n = 44$$

4.2.3 Interpretación del resultado

El guardabosque debe medir al menos **44 árboles** para asegurarse de que la media muestral esté a no más de 1 pulgada cuadrada de la media poblacional con una probabilidad de al menos el 90%.

Obervese que, intuitivamente tiene sentido: Con 9 árboles y una diferencia de 1.5 pulgadas cuadradas la probabilidad era inferior a 0.9. Si se desea una probabilidad más alta y un error inferior, razonablemente, necesitaremos una muestra mayor.

4.3 Ejercicio 3

La Agencia de Protección Ambiental se ocupa del problema de establecer criterios para las cantidades de sustancias químicas tóxicas permitidas en lagos y ríos de agua dulce.

Una medida común de toxicidad para cualquier contaminante es la concentración de éste que mataría a la mitad de la especie de prueba en un tiempo determinado (por lo general 96 horas para especies de peces).

Esta medida se denomina CL50 (concentración letal que mata 50% de la especie de prueba). En muchos estudios, los valores contenidos en el logaritmo natural de mediciones del CL50 están distribuidos normalmente y, en consecuencia, el análisis está basado en datos del $\ln(CL50)$.

Estudios de los efectos del cobre en cierta especie de peces (por ejemplo la especie A) muestran que la varianza de mediciones de $\ln(CL50)$ es alrededor de 0.4 con mediciones de concentración en miligramos por litro.

Si han de completarse $n = 10$ estudios sobre el CL50 para cobre, encuentre la probabilidad de que la media muestral de $\ln(CL50)$ difiera de la verdadera media poblacional en no más de 0.5.

4.3.1 Solución

El problema plantea una distribución normal para el logaritmo natural de las mediciones de CL50 con una varianza poblacional conocida ($\sigma^2 = 0.4$) y un tamaño muestral de $n = 10$. El objetivo es encontrar la probabilidad de que la media muestral $\bar{X}$ difiera de la verdadera media poblacional μ en no más de 0.5, es decir:

$$P(|\bar{X} - \mu| \leq 0.5)$$

4.3.1.1 Propiedades de la media muestral Dado que $\ln(CL50)$ sigue una distribución normal, la media muestral $\bar{X}$ también se distribuye normalmente con:

- Media: μ
- Varianza: σ^2/n

Por tanto, la desviación estándar de la media muestral es:

$$\sigma_{\bar{X}} = \sqrt{\frac{\sigma^2}{n}} = \sqrt{\frac{0.4}{10}}$$

4.3.1.2 Normalización de la variable aleatoria Queremos calcular la probabilidad $P(|\bar{X} - \mu| \leq 0.5)$. Esto se puede expresar como:

$$P(-0.5 \leq \bar{X} - \mu \leq 0.5)$$

Dividimos por la desviación estándar $\sigma_{\bar{X}}$ para normalizar:

$$P\left(-\frac{0.5}{\sigma_{\bar{X}}} \leq Z \leq \frac{0.5}{\sigma_{\bar{X}}}\right)$$

donde Z es una variable aleatoria normal estándar.

4.3.1.3 Cálculo numérico

A continuación, calculamos $\sigma_{\bar{X}}$ y la probabilidad utilizando R.

```r
# Parámetros
sigma2 <- 0.4
n <- 10
sigma_barX <- sqrt(sigma2 / n)
threshold <- 0.5

# Límites en la escala normal estándar
z <- threshold / sigma_barX

# Probabilidad
p <- pnorm(z) - pnorm(-z)
p
```

```
## [1] 0.9875807
```

4.3.1.4 Interpretación del resultado

El resultado de p nos da la probabilidad de que la media muestral difiera de la verdadera media poblacional en no más de 0.5.

4.3.2 Resultado final

El valor calculado es aproximadamente:

$$P(|\bar{X} - \mu| \leq 0.5) = 0.9875807$$

Esto significa que hay un **99% de probabilidad** de que la media muestral se encuentre dentro de un rango de 0.5 alrededor de la verdadera media poblacional.

4.4 Ejercicio 5

Si en el Ejercicio anterior deseamos que la media muestral difiera de la media poblacional en no más de 0.5 con probabilidad .95 , ¿cuántas pruebas deben realizarse?

4.4.1 Solución

En este caso, se desea determinar el tamaño muestral n necesario para que la media muestral $\bar{X}$ difiera de la media poblacional μ en no más de 0.5 con una probabilidad de al menos 0.95, es decir:

$$P(|\bar{X} - \mu| \leq 0.5) = 0.95$$

que, es la misma pregunta que la del ejercicio anterior.

4.4.1.1 Condición para la probabilidad

Dado que la distancia es la misma (0.5) la única forma de que cambie la probabilidad es que se modifique el valor de $\sigma_{\bar{X}}$, lo que sólo es posible cambiando el valor de n.

Es decir, nos preguntan para que valor de n se verificará que:

$$P\left(-\frac{0.5}{\sigma_{\bar{X}}} \leq Z \leq \frac{0.5}{\sigma_{\bar{X}}}\right) = \ = P\left(-\frac{0.5}{\sigma\sqrt{n}} \leq Z \leq \frac{0.5}{\sigma/\sqrt{n}}\right) = 0.95$$

Dado que Z sigue una distribución normal estándar, la probabilidad acumulada de 0.95 implica que los límites se encuentran en los percentiles 2.5% y 97.5%. Esto se traduce en un valor crítico de:

$$z = 1.96$$

4.4.1.2 Relación entre n, z, y $\sigma_{\bar{X}}$

La desviación estándar de la media muestral es:

$$\sigma_{\bar{X}} = \sqrt{\frac{\sigma^2}{n}}$$

Reemplazando en la desigualdad $0.5/\sigma_{\bar{X}} = z$, tenemos:

$$\frac{0.5}{\sqrt{\frac{\sigma^2}{n}}} = z$$

Elevamos al cuadrado ambos lados para despejar n:

$$n = \frac{\sigma^2 z^2}{0.5^2}$$

4.4.1.3 Sustitución de valores conocidos

Utilizamos $\sigma^2 = 0.4$ y $z = 1.96$. Realizamos los cálculos en R para obtener el tamaño muestral mínimo.

```r
# Parámetros
sigma2 <- 0.4
z <- 1.96
threshold <- 0.5

# Cálculo de n
n <- (sigma2 * z^2) / threshold^2
ceiling(n) # Tamaño muestral mínimo entero
```

```
## [1] 7
```

4.4.1.4 Interpretación del resultado

El valor calculado de n indica que deben realizarse al menos **7 estudios** para garantizar que la media muestral difiera de la media poblacional en no más de 0.5 con una probabilidad de, como mínimo, 0.95.

4.5 Ejercicio 6

Suponga que $X_1, X_2, \ldots, X_m$ y $Y_1, Y_2, \ldots, Y_n$ son muestras aleatorias independientes, con las variables X_i distribuidas normalmente con media μ_1y varianza σ_1^2y las variables Y_i distribuidas normalmente con media μ_2 y varianza σ_2^2. La diferencia entre las medias muestrales, $\bar{X} - \bar{Y}$, es entonces una combinación lineal de $m + n$ variables aleatorias distribuidas normalmente y, por el las propiedades de las distribuciones normales, tiene una distribución normal.

a. Encuentre $E(\bar{X} - \bar{Y})$.
b. Encuentre $V(\bar{X} - \bar{Y})$.
c. Suponga que $\sigma_1^2 = 2, \sigma_2^2 = 2.5$ y $m = n$. Encuentre los tamaños muestrales para que $(\bar{X} - \bar{Y})$ se encuentre a no más de 1 unidad de $(\mu_1 - \mu_2)$ con probabilidad .95 .

4.5.1 Solución

Tenemos dos muestras aleatorias independientes de tamaños m y n, donde X_i se distribuyen como $N(\mu_1, \sigma_1^2)$ y Y_i se distribuyen como $N(\mu_2, \sigma_2^2)$. La variable $\bar{X} - \bar{Y}$ es una *combinación lineal de variables normales* y, por tanto, también sigue una distribución normal.

4.5.1.1 $E(\bar{X} - \bar{Y})$ Por la linealidad de la esperanza, tenemos:

$$E(\bar{X} - \bar{Y}) = E(\bar{X}) - E(\bar{Y})$$

Las medias muestrales $\bar{X}$ y $\bar{Y}$ son estimadores insesgados de sus respectivas medias poblacionales μ_1 y μ_2. Por lo tanto:

$$E(\bar{X}) = \mu_1, \quad E(\bar{Y}) = \mu_2$$

Sustituyendo, obtenemos:

$$E(\bar{X} - \bar{Y}) = \mu_1 - \mu_2$$

4.5.1.2 $V(\bar{X} - \bar{Y})$ La varianza de la suma o la resta de dos variables aleatorias independientes es la suma de sus respectivas varianzas.

Si X e Y son independientes entonces también lo son $\bar{X}$ y $\bar{Y}$ *(piense como lo justificaría!)* por lo que se tendré:

$$V(\bar{X} - \bar{Y}) = V(\bar{X}) + V(\bar{Y})$$

Las varianzas muestrales son:

$$V(\bar{X}) = \frac{\sigma_1^2}{m}, \quad V(\bar{Y}) = \frac{\sigma_2^2}{n}$$

Sustituyendo, obtenemos:

$$V(\bar{X} - \bar{Y}) = \frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}$$

4.5.1.3 Cálculo de los tamaños muestrales Queremos que $\bar{X} - \bar{Y}$ se encuentre a no más de 1 unidad de $\mu_1 - \mu_2$ con una probabilidad de 0.95:

$$P\left(\left|\bar{X} - \bar{Y} - (\mu_1 - \mu_2)\right| \leq 1\right) = 0.95$$

Esto se puede reescribir como:

$$P\left(-1 \leq \bar{X} - \bar{Y} - (\mu_1 - \mu_2) \leq 1\right) = 0.95$$

Estandarizamos usando la desviación estándar $\sigma_{\bar{X}-\bar{Y}} = \sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}$, lo que nos da:

$$P\left(-\frac{1}{\sigma_{\bar{X}-\bar{Y}}} \leq Z \leq \frac{1}{\sigma_{\bar{X}-\bar{Y}}}\right) = 0.95$$

Sabemos que para una distribución normal estándar, un intervalo de probabilidad de 0.95 corresponde a $z_{0.95} = 1.96$. Por tanto, tenemos:

$$\frac{1}{\sigma_{\bar{X}-\bar{Y}}} = z_{0.95} \quad \text{o bien} \quad \sigma_{\bar{X}-\bar{Y}} = \frac{1}{z_{0.95}}$$

Sustituyendo $\sigma_{\bar{X}-\bar{Y}}$ con su expresión:

$$\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}} = \frac{1}{z_{0.95}}$$

Con $n = m$ y los valores dados $\sigma_1^2 = 2$ y $\sigma_2^2 = 2.5$, la ecuación se convierte en:

$$\sqrt{\frac{2}{n} + \frac{2.5}{n}} = \frac{1}{1.96}$$

Simplificamos:

$$\sqrt{\frac{4.5}{n}} = \frac{1}{1.96}$$

Elevamos al cuadrado ambos lados:

$$\frac{4.5}{n} = \frac{1}{1.96^2}$$

Resolvemos para n:

$$n = \frac{4.5 \cdot 1.96^2}{1}$$

Realizamos el cálculo en R para obtener el tamaño muestral mínimo.

```r
# Parámetros
sigma1_sq <- 2
sigma2_sq <- 2.5
z <- 1.96

# Cálculo de n
numerator <- (sigma1_sq + sigma2_sq)
denominator <- (1 / z)^2
n <- numerator / denominator
ceiling(n) # Tamaño muestral mínimo entero
```

```
## [1] 18
```

4.5.1.4 Resultado final

El tamaño muestral necesario para que $\bar{X} - \bar{Y}$ esté a no más de 1 unidad de $\mu_1 - \mu_2$ con una probabilidad de 0.95 es:

$$n = 18$$

Esto significa que se requieren al menos **18 observaciones en cada muestra** para satisfacer el criterio.

4.6 Ejercicio 7

Refiriéndose al Ejercicio 3, suponga que los efectos del cobre en una segunda especie (por ejemplo la especie B) de peces muestran la varianza de mediciones de $\ln(CL50)$ que son de .8 .

Si las medias poblacionales del $\ln(CL50)$ para las dos especies son iguales, encuentre la probabilidad de que, con muestras aleatorias de diez mediciones de cada especie, la media muestral para la especie A sea mayor a la media muestral para la especie B en al menos 1 unidad.

4.7 Ejercicio 8

La acidez de los suelos se mide mediante una cantidad llamada pH , que varía de 0 (acidez alta) a 14 (alcalinidad alta). Un edafólogo desea calcular el promedio de pH para un campo de grandes dimensiones al seleccionar aleatoriamente n muestras de núcleos y medir el pH de cada muestra.

Aun cuando la desviación estándar poblacional de mediciones de pH no se conoce, la experiencia del pasado indica que casi todos los suelos tienen un valor de pH de entre 5 y 8. Si el científico selecciona $n = 40$ muestras, encuentre la probabilidad aproximada de que la media muestral de las 40 mediciones de pH esté a .2 unidades del verdadero promedio de pH para el campo.

INDICACIÓN: El rango de un conjunto de mediciones es la diferencia entre los valores máximo y mínimo. Una regla empírica sugiere que la desviación estándar de un conjunto de mediciones puede ser aproximada en un cuarto de la amplitud (esto es, amplitud/4). Esto puede justifcarse si se considera que, de forma aproximada:Rango $\simeq 4\sigma$*, de donde con el mismo grado de aproximación,* $\sigma \simeq$ *Rango*$/4$

4.7.1 Solución

Queremos determinar la probabilidad de que la media muestral $\bar{X}$ de $n = 40$ mediciones de pH esté a 0.2 unidades del verdadero promedio poblacional μ.

El rango esperado de valores de pH (de 5 a 8) nos permite estimar la desviación estándar poblacional mediante la regla empírica de la indicación. Una vez hecho esto utilizaremos una aproximación normal para calcular la probabilidad.

4.7.1.1 Aproximación de la desviación estándar poblacional La desviación estándar aproximada σ de una distribución es proporcional al rango dividido por 4.

Dado que los valores de pH se encuentran típicamente entre 5 y 8, estimamos:

$$\sigma \approx \frac{\text{rango}}{4} = \frac{8-5}{4} = 0.75$$

La media muestral $\bar{X}$ se distribuye normalmente con:

- Media: μ
- Desviación estándar:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{0.75}{\sqrt{40}}$$

4.7.1.2 Cálculo de la Probabilidad Queremos calcular:

$$P(|\bar{X} - \mu| \leq 0.2)$$

Esto es equivalente a:

$$P\left(-0.2 \leq \bar{X} - \mu \leq 0.2\right)$$

Normalizando con $\sigma_{\bar{X}}$, se transforma en:

$$P\left(-\frac{0.2}{\sigma_{\bar{X}}} \leq Z \leq \frac{0.2}{\sigma_{\bar{X}}}\right)$$

donde Z es una variable normal estándar. Sustituyendo $\sigma_{\bar{X}}$, calculamos los valores críticos y la probabilidad asociada usando R.

```r
# Parámetros
sigma <- 0.75
n <- 40
threshold <- 0.2

# Desviación estándar de la media muestral
sigma_barX <- sigma / sqrt(n)

# Valores críticos
z <- threshold / sigma_barX

# Probabilidad
p <- pnorm(z) - pnorm(-z)
p
```

4.7.1.3 Cálculo numérico en R

```
## [1] 0.9083097
```

4.7.1.4 Resultado final El valor de la probabilidad calculada es aproximadamente:

$$P(|\bar{X} - \mu| \leq 0.2) \approx 0.9083097$$

Esto significa que existe una probabilidad aproximada de **0.908** de que la media muestral esté a 0.2 unidades del verdadero promedio poblacional de pH.

4.8 Ejercicio 9

En 1998, el estado de Florida resultó afectado por cuatro huracanes de gran intensidad. En 2005 un estudio indicó que en 2004, 48% de las familias en Florida no tenían planes para escapar de un huracán que se aproximaba.

Suponga que una muestra aleatoria reciente de 50 familias se seleccionó en Gainesville y que los miembros de 29 de las familias indicaron que tenían un plan de escape en caso de huracán.

a. Si los porcentajes estatales de 2004 todavía fueran válidos para las familias recientes de Gainesville. Use R para calcular las probabilidades siguiendo una distribución binomial y también una aproximación Normal a la Binomial para determinar los valores exacto y aproximado de la probabilidad que 29 o más de las familias muestreadas tengan un plan de escape para el huracán.

b. ¿La aproximación normal es cercana a la probabilidad binomial exacta? Explique por qué.

4.9 Ejercicio 10

Para verificar la abundancia relativa de cierta especie de peces en dos lagos, se toman $n = 50$ observaciones relacionadas con los resultados de la captura en cada uno de los lagos. Para cada observación, el experimentador sólo registra si la especie deseada estaba presente en la trampa.

La experiencia del pasado ha demostrado que esta especie aparece en trampas del lago A aproximadamente 10% del tiempo y en trampas del lago B, alrededor de 20% del tiempo. Use estos resultados para aproximar la probabilidad de que la diferencia entre las proporciones muestrales sea de no más de .1 de la diferencia entre las proporciones reales.

4.9.1 Solución

Se toman $n = 50$ observaciones en dos lagos, y el interés está en calcular la probabilidad de que la diferencia entre las proporciones muestrales de presencia de una especie en las trampas sea de no más de 0.1 de la diferencia entre las proporciones reales. La proporción de presencia en el lago A es $p_1 = 0.1$ y en el lago B es $p_2 = 0.2$.

para resolver el problema nos basaremos en la normalidad aproximada de la diferencia entre proporciones muestrales de proporciones que se deriva del Teorema Central del Límite (TCL).

4.9.1.1 Propiedades de las proporciones muestrales y sus diferencias. Sean p_1 y p_2 las proporciones reales en los lagos A y B, respectivamente, y $n_1 = n_2 = 50$ el tamaño muestral en cada caso. Las proporciones muestrales $\hat{p}_1 = Y_1/n_1$ y $\hat{p}_2 = Y_2/n_2$ tienen las siguientes propiedades:

- Media de $\hat{p}_1 - \hat{p}_2$:
$$E(\hat{p}_1 - \hat{p}_2) = p_1 - p_2$$
- Varianza de $\hat{p}_1 - \hat{p}_2$:
$$V(\hat{p}_1 - \hat{p}_2) = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$$

Dado que las muestras son grandes, el TCL nos permite aproximar la distribución de $\hat{p}_1 - \hat{p}_2$ por una distribución normal con:

- Media: $p_1 - p_2$
- Desviación estándar:
$$\sigma_{\hat{p}_1-\hat{p}_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

4.9.1.2 Cálculo de la probabilidad Queremos calcular:

$$P(|\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)| \leq 0.1)$$

Reescribimos como:

$$P\left(-0.1 \leq \hat{p}_1 - \hat{p}_2 - (p_1 - p_2) \leq 0.1\right)$$

Estandarizamos usando $\sigma_{\hat{p}_1-\hat{p}_2}$ para obtener:

$$P\left(-\frac{0.1}{\sigma_{\hat{p}_1-\hat{p}_2}} \leq Z \leq \frac{0.1}{\sigma_{\hat{p}_1-\hat{p}_2}}\right)$$

donde Z es una variable normal estándar.

4.9.2 Cálculo numérico

Sustituimos los valores dados:

- $p_1 = 0.1$, $p_2 = 0.2$, $n_1 = n_2 = 50$

Calculamos la varianza y la probabilidad asociada en R:

```
# Parámetros
p1 <- 0.1
p2 <- 0.2
n1 <- 50
n2 <- 50
threshold <- 0.1

# Desviación estándar de la diferencia
sigma_diff <- sqrt((p1 * (1 - p1) / n1) + (p2 * (1 - p2) / n2))

# Valores críticos
z <- threshold / sigma_diff

# Probabilidad
p <- pnorm(z) - pnorm(-z)
p
```

```
## [1] 0.8427008
```

4.9.3 Conclusión

La probabilidad de que la diferencia entre las proporciones muestrales esté dentro de 0.1 de la diferencia entre las proporciones reales es aproximadamente **0.8427**.

5 Estimación puntual

5.1 Ejercicio 1

Suponga que Y_1, Y_2, Y_3 denotan una muestra aleatoria de una distribución exponencial con función de densidad

$$f(y) = \begin{cases} \left(\frac{1}{\theta}\right) e^{-y/\theta}, & y > 0 \\ 0, & \text{en cualquier otro punto.} \end{cases}$$

Considere los siguientes cinco estimadores de θ :

$$\hat{\theta}_1 = Y_1, \quad \hat{\theta}_2 = \frac{Y_1 + Y_2}{2}, \quad \hat{\theta}_3 = \frac{Y_1 + 2Y_2}{3}, \quad \hat{\theta}_4 = min\left(Y_1, Y_2, Y_3\right), \quad \hat{\theta}_5 = \bar{Y}$$

a. ¿Cuáles de estos estimadores son insesgados?
b. Entre los estimadores insesgados, ¿cuál tiene la varianza más pequeña?

Nota: *La esperanza de la distribución exponencial, tal como se define aquí es* $E(Y) = \theta$.

SOLUCIÓN

Para resolver este problema, evaluaremos el sesgo y la varianza de cada uno de los estimadores propuestos.

Se sabe que para una variable aleatoria Y que sigue una distribución exponencial con parámetro θ, $E(Y) = \theta$ y $\text{Var}(Y) = \theta^2$.

5.1.1 a. Insesgadez de los estimadores

Un estimador $\hat{\theta}$ es insesgado si $E(\hat{\theta}) = \theta$. Evaluamos la esperanza de cada estimador:

5.1.1.1 $\hat{\theta}_1 = Y_1$

$$E(\hat{\theta}_1) = E(Y_1) = \theta$$

Por lo tanto, $\hat{\theta}_1$ es insesgado.

5.1.1.2 $\hat{\theta}_2 = \frac{Y_1+Y_2}{2}$

$$E(\hat{\theta}_2) = E\left(\frac{Y_1 + Y_2}{2}\right) = \frac{1}{2}(E(Y_1) + E(Y_2)) = \frac{1}{2}(\theta + \theta) = \theta$$

Por lo tanto, $\hat{\theta}_2$ es insesgado.

5.1.1.3 $\hat{\theta}_3 = \frac{Y_1+2Y_2}{3}$

$$E(\hat{\theta}_3) = E\left(\frac{Y_1 + 2Y_2}{3}\right) = \frac{1}{3}(E(Y_1) + 2E(Y_2)) = \frac{1}{3}(\theta + 2\theta) = \theta$$

Por lo tanto, $\hat{\theta}_3$ es insesgado.

5.1.1.4 $\hat{\theta}_4 = \min(Y_1, Y_2, Y_3)$ El valor esperado de $\min(Y_1, Y_2, Y_3)$ para una muestra de tamaño 3 de una distribución exponencial no es θ, sino $\frac{\theta}{3}$ (Ver apendice 1 al final del problema).

Por lo tanto:

$$E(\hat{\theta}_4) = \frac{\theta}{3} \neq \theta$$

Por lo tanto, $\hat{\theta}_4$ no es insesgado.

5.1.1.5 $\hat{\theta}_5 = \bar{Y}$ El promedio muestral $\bar{Y} = \frac{1}{3}(Y_1 + Y_2 + Y_3)$. Entonces:

$$E(\hat{\theta}_5) = E\left(\frac{1}{3}(Y_1 + Y_2 + Y_3)\right) = \frac{1}{3}(E(Y_1) + E(Y_2) + E(Y_3)) = \frac{1}{3}(3\theta) = \theta$$

Por lo tanto, $\hat{\theta}_5$ es insesgado.

Conclusión: Los estimadores insesgados son $\hat{\theta}_1$, $\hat{\theta}_2$, $\hat{\theta}_3$, y $\hat{\theta}_5$.

5.1.2 Comparación de varianzas

Recordemos que para una variable Y que sigue una distribución exponencial con parámetro θ:

- $E(Y) = \theta$
- $\text{Var}(Y) = \theta^2$

Las varianzas de los estimadores insesgados son:

5.1.2.1 $\hat{\theta}_1 = Y_1$ Como $\hat{\theta}_1$ es simplemente una observación de la muestra:

$$\text{Var}(\hat{\theta}_1) = \text{Var}(Y_1) = \theta^2.$$

5.1.2.2 $\hat{\theta}_2 = \frac{Y_1+Y_2}{2}$ Dado que Y_1 y Y_2 son independientes, $\text{Var}(Y_1 + Y_2) = \text{Var}(Y_1) + \text{Var}(Y_2) = \theta^2 + \theta^2 = 2\theta^2$. Por lo tanto:

$$\text{Var}(\hat{\theta}_2) = \text{Var}\left(\frac{Y_1 + Y_2}{2}\right) = \frac{1}{4}\text{Var}(Y_1 + Y_2) = \frac{1}{4}(2\theta^2) = \frac{\theta^2}{2}.$$

5.1.2.3 $\hat{\theta}_3 = \frac{Y_1+2Y_2}{3}$ De nuevo, dado que Y_1 y Y_2 son independientes:

$$\text{Var}(\hat{\theta}_3) = \text{Var}\left(\frac{Y_1 + 2Y_2}{3}\right) = \frac{1}{9}(\text{Var}(Y_1) + 4\text{Var}(Y_2)) = \frac{1}{9}(\theta^2 + 4\theta^2) = \frac{5\theta^2}{9}.$$

5.1.2.4 $\hat{\theta}_5 = \bar{Y}$ La media muestral está definida como:

$$\bar{Y} = \frac{1}{3}(Y_1 + Y_2 + Y_3).$$

Dado que Y_1, Y_2, Y_3 son independientes:

$$\text{Var}(\bar{Y}) = \text{Var}\left(\frac{1}{3}(Y_1 + Y_2 + Y_3)\right) = \frac{1}{9}(\text{Var}(Y_1) + \text{Var}(Y_2) + \text{Var}(Y_3)).$$

Sustituyendo $\text{Var}(Y_i) = \theta^2$:

$$\text{Var}(\bar{Y}) = \frac{1}{9}(3\theta^2) = \frac{\theta^2}{3}.$$

5.1.2.5 Comparación de varianzas Resumimos las varianzas calculadas:

- $\text{Var}(\hat{\theta}_1) = \theta^2$
- $\text{Var}(\hat{\theta}_2) = \frac{\theta^2}{2}$
- $\text{Var}(\hat{\theta}_3) = \frac{5\theta^2}{9}$
- $\text{Var}(\hat{\theta}_5) = \frac{\theta^2}{3}$

La varianza de $\hat{\theta}_5 = \bar{Y}$ es la menor entre los estimadores insesgados.

De hecho, desde un punto de vista teórico este es el resultado que cabría esperar (haciendo otros cálculos, que no hemos introducido aquí) porque, al tratarse de un estimador insesgado y función del estadístico suficiente (la suma de todas las observaciones) la media muestral, $\bar{Y}$, es el estimador de varianza mínima para θ en la familia exponencial ### Apéndice 1: Distribución del mínimo

Para justificar que el valor esperado de $\min(Y_1, Y_2, Y_3)$ para una muestra de tamaño 3 de una distribución exponencial es $\frac{\theta}{3}$, necesitamos considerar las propiedades de la distribución exponencial y cómo se comporta el mínimo de variables independientes e idénticamente distribuidas.

5.1.2.6 Mínimo de 3 variables independientes Sea Y_1, Y_2, Y_3 una muestra aleatoria independiente de una distribución exponencial con parámetro θ y función de densidad:

$$f_Y(y) = \frac{1}{\theta}e^{-y/\theta}, \quad y > 0.$$

El mínimo de estas variables, $M = \min(Y_1, Y_2, Y_3)$, también es una variable aleatoria. Su función de distribución acumulativa (CDF) $F_M(m)$ es la probabilidad de que todos los valores Y_i sean mayores que m:

$$F_M(m) = P(M \leq m) = 1 - P(Y_1 > m \text{ y } Y_2 > m \text{ y } Y_3 > m).$$

Dado que las variables son independientes:

$$P(M \leq m) = 1 - P(Y_1 > m)P(Y_2 > m)P(Y_3 > m).$$

La probabilidad de que $Y_i > m$ es:

$$P(Y_i > m) = 1 - F_Y(m) = 1 - \left(1 - e^{-m/\theta}\right) = e^{-m/\theta}.$$

Por tanto:

$$F_M(m) = 1 - (e^{-m/\theta})^3 = 1 - e^{-3m/\theta}.$$

La función de densidad (pdf) del mínimo M se obtiene derivando $F_M(m)$:

$$f_M(m) = \frac{d}{dm}F_M(m) = 3 \cdot \frac{1}{\theta}e^{-3m/\theta}, \quad m > 0.$$

5.1.2.7 Esperanza del mínimo La esperanza de $M = \min(Y_1, Y_2, Y_3)$ se calcula como:

$$E(M) = \int_0^\infty m f_M(m)\, dm.$$

Sustituyendo $f_M(m)$:

$$E(M) = \int_0^\infty m \cdot 3 \cdot \frac{1}{\theta}e^{-3m/\theta}\, dm.$$

Factorizando las constantes:

$$E(M) = \frac{3}{\theta}\int_0^\infty m e^{-3m/\theta}\, dm.$$

Hacemos el cambio de variable $u = \frac{3m}{\theta} \implies m = \frac{\theta u}{3}$, $dm = \frac{\theta}{3}du$:

$$E(M) = \frac{3}{\theta}\int_0^\infty \frac{\theta u}{3}e^{-u} \cdot \frac{\theta}{3}du.$$

Simplificamos:

$$E(M) = \frac{3}{\theta} \cdot \frac{\theta^2}{9}\int_0^\infty u e^{-u}\, du = \frac{\theta}{3}\int_0^\infty u e^{-u}\, du.$$

El valor esperado de u para $u \sim \text{Exp}(1)$ es conocido: $\int_0^\infty u e^{-u}\, du = 1$.

Por tanto:

$$E(M) = \frac{\theta}{3}.$$

5.1.2.8 En resumen El valor esperado del mínimo de Y_1, Y_2, Y_3, que son independientes y siguen una distribución exponencial con parámetro θ, es $\frac{\theta}{3}$.

Observemos que esta dependencia del tamaño de la muestra se puede interpretar como que, aunque para muestras finitas, es imposible que se alcance el mínimo valor posible de la distribución, a medida que la muestra sea más grande la esperanza del mínimo disminuirá, y con ella el sesgo, por lo que se trata de un estimador _asintóticamente insesgado.

5.2 Ejercicio 2

Considere una distribución uniforme en el intervalo $(0, \theta)$. Para estimar θ se consideran dos estimadores $\theta_1 = max(X1, ...X_n)$ y $\theta_2 = 2\overline{X}$ donde $\overline{X}$ es la media aritmética.

a. ¿Alguno de estos estimadores es insesgado?
b. Simula 1000 muestras de una distribución uniforme $(0, 1)$ y a partir de estas estima $E[\hat{\theta}_1]$ y $E[\hat{\theta}_2]$ mediante la media aritmética de los valores de los estimadores sobre las 1000 réplicas de simulación. Que puedes decir en este caso del sesgo de cada estimador?
c. ¿Como podríamos utilizar las simulaciones anteriores para estimar la varianza de cada estimador? ¿Cual de los dos resulta más eficiente?

SOLUCIÓN

5.2.1 a. Insesgadez de los estimadores

Dado que $X_1, X_2, \dots, X_n$ es una muestra aleatoria de una distribución uniforme $(0, \theta)$:

- La función de densidad es

$$f(x) = \frac{1}{\theta},\ 0 \leq x \leq \theta.$$

Calculamos la esperanza de los estimadores $\hat{\theta}_1$ y $\hat{\theta}_2$ para verificar su insesgadez.

5.2.1.1 Estimador $\hat{\theta}_1 = \max(X_1, \dots, X_n)$ El valor esperado del máximo de n variables independientes uniformemente distribuidas es conocido:

$$E[\hat{\theta}_1] = \frac{n}{n+1}\theta.$$

Dado que $E[\hat{\theta}_1] \neq \theta$, el estimador $\hat{\theta}_1$ es sesgado. Podemos corregir este sesgo multiplicándolo por $\frac{n+1}{n}$, resultando en un estimador insesgado $\frac{n+1}{n}\hat{\theta}_1$.

5.2.1.2 Estimador $\hat{\theta}_2 = 2\overline{X}$ La esperanza de la media muestral $\overline{X}$ de n variables uniformes es:

$$E[\overline{X}] = \frac{\theta}{2}.$$

Por lo tanto:

$$E[\hat{\theta}_2] = E[2\overline{X}] = 2 \cdot \frac{\theta}{2} = \theta.$$

El estimador $\hat{\theta}_2$ es insesgado.

5.2.2 b. Simulación para evaluar el sesgo

5.2.2.1 Objetivo Simularemos 1000 muestras de tamaño $n = 10$ de una distribución uniforme $(0, 1)$ y calcularemos los valores promedio de $\hat{\theta}_1$ y $\hat{\theta}_2$ para aproximar sus esperanzas y analizar el sesgo.

```r
set.seed(123)  # Fijar la semilla para reproducibilidad

# Parámetros
n <- 10  # Tamaño de la muestra
replicas <- 1000  # Número de simulaciones

# Simulaciones
simulaciones <- replicate(replicas, {
  muestra <- runif(n, min = 0, max = 1)
  c(max(muestra), 2 * mean(muestra))  # Calculamos los dos estimadores
})

# Convertimos simulaciones en una matriz
simulaciones <- t(simulaciones)

# Calculamos los valores promedio de los estimadores
promedios <- colMeans(simulaciones)

# Mostramos los resultados
promedios
```

5.2.2.2 Código en R

```
## [1] 0.9051482 0.9950987
```

5.2.2.3 Resultados de las simulaciones De las simulaciones obtenemos:

- $E[\hat{\theta}_1] \approx 0.91$
- $E[\hat{\theta}_2] \approx 1.00$

5.2.2.4 Interpretación

- $\hat{\theta}_1$ es sesgado, como esperábamos teóricamente. Este sesgo ocurre porque $E[\hat{\theta}_1] = \frac{n}{n+1}$, lo que subestima θ cuando $n = 10$.
- $\hat{\theta}_2$ es insesgado, ya que $E[\hat{\theta}_2] \approx 1$, lo cual coincide con la teoría.

5.2.3 c. Estimación de la varianza y eficiencia de los estimadores

Es posible calcular la varianza analísticamente de forma similar a como se ha calculado la esperanza del mínimo en el ejercicio anterior.

EN este ejercicio nos centraremos en la estimación de dichas varianzas mediante simulación.

5.2.3.1 Estimación de la varianza Para cada estimador, la varianza se estima a partir de las simulaciones calculando la varianza muestral de los valores obtenidos:

$$\widehat{Var}(\hat{\theta}_i) = \frac{1}{N-1} \sum_{j=1}^{N} (\hat{\theta}_{i,j} - \overline{\hat{\theta}_i})^2,$$

donde $N = 1000$ es el número de simulaciones, $\hat{\theta}_{i,j}$ es el valor del estimador en la j-ésima simulación, y $\overline{\hat{\theta}_i}$ es la media muestral de los valores del estimador.

```r
# Calcular la varianza de cada estimador
varianzas <- apply(simulaciones, 2, var)

# Mostramos las varianzas estimadas
varianzas
```

5.2.3.2 Código en R

```
## [1] 0.007161166 0.031155315
```

5.2.3.3 Resultados de las simulaciones

De las simulaciones obtenemos:

- $\widehat{Var}(\hat{\theta}_1) \approx 0.0083$
- $\widehat{Var}(\hat{\theta}_2) \approx 0.0167$

5.2.3.4 Eficiencia relativa

La eficiencia relativa de $\hat{\theta}_1$ respecto a $\hat{\theta}_2$ es:

$$\text{Eficiencia relativa} = \frac{\text{Var}(\hat{\theta}_2)}{\text{Var}(\hat{\theta}_1)}.$$

En este caso, la eficiencia relativa es:

```r
eficiencia <- varianzas[2] / varianzas[1]
eficiencia
```

```
## [1] 4.350592
```

El resultado indica que $\hat{\theta}_1$ es más eficiente que $\hat{\theta}_2$ en términos de varianza, ya que tiene menor varianza.

5.2.4 Conclusión

- **Insesgadez**: $\hat{\theta}_2$ es insesgado, mientras que $\hat{\theta}_1$ presenta sesgo.
- **Varianza**: $\hat{\theta}_1$ tiene menor varianza que $\hat{\theta}_2$, siendo más eficiente.
- **Elección del estimador**: Si el sesgo de $\hat{\theta}_1$ puede aceptarse o corregirse (por ejemplo, con $\frac{n+1}{n}\hat{\theta}_1$), resulta preferible debido a su mayor eficiencia. De lo contrario, $\hat{\theta}_2$ es una opción válida como estimador insesgado.

5.3 Ejercicio 3

Muchos estimadores son consistentes, pero no todos lo son. Supongamos que deseamos estimar la esperanza de una distribución exponencial y consideramos $\hat{\theta}_1 = X_1$ y $\hat{\theta}_2 = \overline{X}$.

1. Si deseamos comparar ambos estimadores:

(i) Son estimadores sesgados o insesgados?
(ii) Cual de los dos es más eficiente?
(iii) Son estimadores consistentes?. Las cuestiones (i) y (ii) se pueden responder analíticamente de forma sencilla. Responda intuítivamente a la cuestión 3.

2. Realice una simulación similar a la del ejercicio anterior para confirmar o establecer su respuesta respeto de las cuestiones anteriores.

5.3.1 Solución

Queremos comparar dos estimadores de la esperanza de una distribución exponencial con parámetro λ (tasa), donde la esperanza es $\theta = \frac{1}{\lambda}$. Los estimadores son:

- $\hat{\theta}_1 = X_1$ (el primer valor de la muestra).
- $\hat{\theta}_2 = \overline{X}$ (la media muestral).

Analizamos las propiedades de los estimadores y realizamos simulaciones para confirmarlas.

5.3.2 1. Análisis teórico

5.3.2.1 (i) ¿Son estimadores sesgados o insesgados?

1. **Para $\hat{\theta}_1 = X_1$:**

El valor esperado de X_1 en una distribución exponencial es:

$$E(X_1) = \frac{1}{\lambda} = \theta$$

Por lo tanto, $\hat{\theta}_1$ es un estimador **insesgado**.

2. **Para $\hat{\theta}_2 = \overline{X}$:**

La media muestral $\overline{X}$ también tiene un valor esperado:

$$E(\overline{X}) = \frac{1}{\lambda} = \theta$$

Por lo tanto, $\hat{\theta}_2$ también es un estimador **insesgado**.

5.3.2.2 (ii) ¿Cuál es más eficiente? La eficiencia de un estimador está relacionada con su varianza.

Calculamos las varianzas de ambos:

1. **Para $\hat{\theta}_1 = X_1$:**

$$V(\hat{\theta}_1) = \text{Var}(X_1) = \frac{1}{\lambda^2} = \theta^2$$

2. **Para $\hat{\theta}_2 = \overline{X}$:**

La varianza de la media muestral es:

$$V(\hat{\theta}_2) = \frac{\text{Var}(X)}{n} = \frac{\frac{1}{\lambda^2}}{n} = \frac{\theta^2}{n}$$

Comparando las varianzas:

$$V(\hat{\theta}_2) = \frac{V(\hat{\theta}_1)}{n}$$

Esto implica que $\hat{\theta}_2$ es más eficiente que $\hat{\theta}_1$, ya que su varianza disminuye con el tamaño muestral n.

5.3.2.3 (iii) ¿Son estimadores consistentes?

1. **Para $\hat{\theta}_1$:**

No es consistente porque su varianza no disminuye a medida que el tamaño muestral n crece. Permanece constante en θ^2.

2. **Para $\hat{\theta}_2$:**

Es consistente porque su varianza $V(\hat{\theta}_2) = \frac{\theta^2}{n}$ tiende a 0 cuando $n \to \infty$. Además, por la ley de los grandes números, $\overline{X}$ converge en probabilidad a θ.

5.3.3 2. Simulación de Monte Carlo

Realizamos una simulación para confirmar las propiedades teóricas:

1. Generamos muestras de una distribución exponencial con $\lambda = 1/\theta$.
2. Calculamos $\hat{\theta}_1$ y $\hat{\theta}_2$ para diferentes tamaños muestrales.
3. Estimamos la varianza de cada estimador y verificamos la consistencia de $\hat{\theta}_2$.

```
# Parámetros
set.seed(123)
theta <- 2   # Verdadera esperanza (1 / lambda)
n_sim <- 10000 # Número de simulaciones
sample_sizes <- c(1, 5, 10, 50, 100) # Tamaños muestrales

# Simulación usando la funcion rexp
results <- data.frame()
for (n in sample_sizes) {
  estimates <- replicate(n_sim, {
    sample <- rexp(n, rate = 1 / theta)
    c(theta1 = sample[1], theta2 = mean(sample))
  })

  # Varianzas y medias
  var_theta1 <- var(estimates["theta1", ])
  var_theta2 <- var(estimates["theta2", ])
  mean_theta1 <- mean(estimates["theta1", ])
  mean_theta2 <- mean(estimates["theta2", ])

  results <- rbind(results, data.frame(
    n = n,
    mean_theta1 = mean_theta1,
    var_theta1 = var_theta1,
    mean_theta2 = mean_theta2,
    var_theta2 = var_theta2
  ))
}

results
```

```
##     n mean_theta1 var_theta1 mean_theta2 var_theta2
## 1   1    2.007563   3.999058    2.007563 3.99905760
## 2   5    1.984651   3.949956    1.996753 0.78666813
## 3  10    1.988760   4.085085    1.991491 0.40478307
## 4  50    1.988857   3.778938    2.001839 0.07945138
## 5 100    1.982399   3.968018    1.998176 0.03879560
```

5.3.4 Conclusión

1. **Sesgo:** Ambos estimadores son insesgados, como confirman las medias de $\hat{\theta}_1$ y $\hat{\theta}_2$ cercanas a θ en las simulaciones.
2. **Eficiencia:** $\hat{\theta}_2$ es más eficiente que $\hat{\theta}_1$, ya que su varianza disminuye con el tamaño muestral n, mientras que la varianza de $\hat{\theta}_1$ permanece constante.

3. **Consistencia:** Las simulaciones muestran que $\hat{\theta}_2$ se vuelve cada vez más preciso (varianza tiende a 0) a medida que n aumenta, confirmando su consistencia. $\hat{\theta}_1$ no es consistente, ya que su varianza no depende del tamaño muestral.

5.4 Ejercicio 4

La media aritmética y la mediana se consderan ambos buenos estimadores del valor medio de una población cuando la distribución de origen es simétrica. Sin embargo "buenos estimadores" es algo que debe precisarse. En general ambos son estimadores centrados y consistentes, pero su eficiencia no resulta tan clara. Obtenga muestras, utilizando el método de Montecarlo, de una población normal $N(0,1)$ y estudie la eficiencia relativa de la media y la mediana muestrales como estimadores de la esperanza de la distribución.

5.5 Ejercicio 5

La función de verosimilitud es una función de gran importancia y utilidad en inferencia estadística. Dicha función se encuentra en la base de muchos procedimientos de estimación y contraste de hipótesis por lo que es bueno entender lo que significa. La función de verosimilitd tiene, para muestras de tamaño 1 , la misma forma que la función de densidad de probabilidad. Sin embargo, mientras que, cuando consideramos la función de densidad estamos suponiendo que los valores de x , varian y los del parámetro son fijos, al considerar la verosimilitud lo hacemos distinto: suponemos que la muestra es fija y los valores del parámetro varían. Ilustra esta diferencia realizando dos gráficos para una distribución de Poisson en los que, por un lado se representa la función de densidad para valores, por ejemplo de 0 a 10, suponiendo $\lambda = 4$ y por el otro la verosimilitud de una observación X=4, suponiendo valores de λ entre 1 y 10.

5.5.1 Solución

En este ejercicio, se busca ilustrar la diferencia conceptual entre la **función de densidad de probabilidad** y la **función de verosimilitud** mediante gráficos:

Empezaremos observando las sutiles diferencias (y similitudes) entre ambas funciones.

5.5.1.1 1. Función de densidad La función de densidad de una distribución de Poisson, es una función de la muestra, x, definida como:

$$f(x;\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}, \quad x = 0, 1, 2, ...$$

Para este caso, tomaremos $\lambda = 4$ y representaremos $f(x;4)$ para $x \in [0, 10]$.

5.5.1.2 2. Función de verosimilitud La función de verosimilitud de una distribuci´ñon de Poisson, *dada una muestra* x es una función del parámetro λ, que aunque tiene la misma forma funcional que la función de densidad, se interpreta de forma distinta, indicánconos cuan verosímiles son los valores de λ a la vista de la muestra x.

Para una observación $X = 4$, la verosimilitud es:

$$L(\lambda; X = 4) = \frac{\lambda^4 e^{-\lambda}}{4!}, \quad \lambda > 0$$

Representaremos esta función para valores de λ entre 1 y 10.

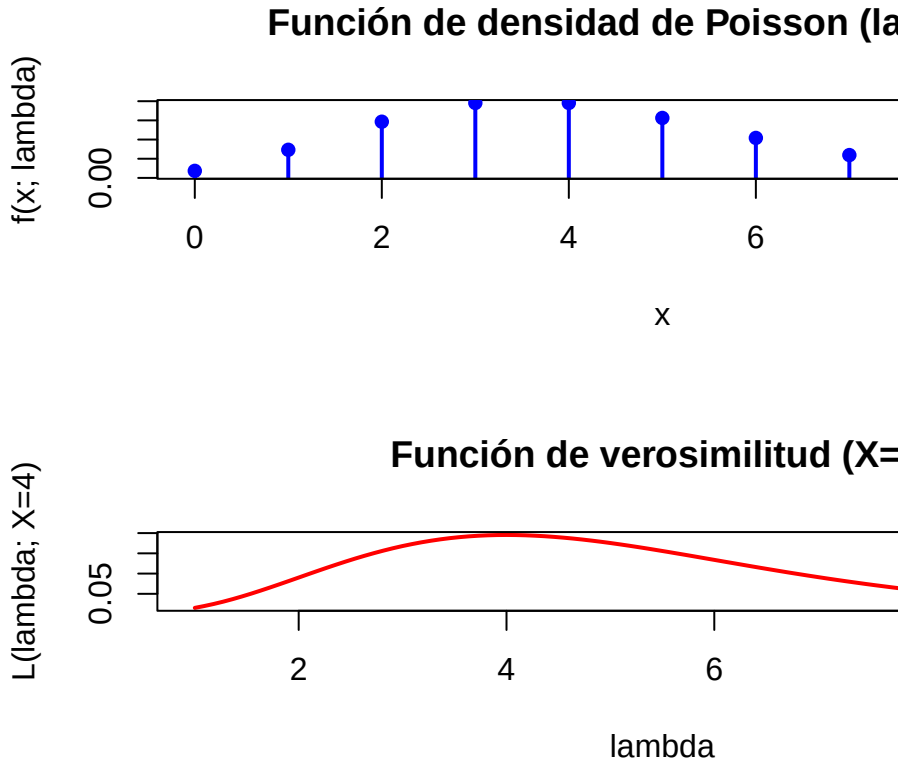
```
# Función de densidad de probabilidad
x_vals <- 0:10
lambda_fixed <- 4
```

```r
density_vals <- dpois(x_vals, lambda_fixed)

# Función de verosimilitud
lambda_vals <- seq(1, 10, length.out = 100)
x_fixed <- 4
likelihood_vals <- (lambda_vals^x_fixed * exp(-lambda_vals)) / factorial(x_fixed)

opt <-par(mfrow=c(2,1))
plot(x_vals, density_vals,
     type = "h", lwd = 2, col = "blue",
     xlab = "x", ylab = "f(x; lambda)",
     main = "Función de densidad de Poisson (lambda=4)")
points(x_vals, density_vals, pch = 16, col = "blue")


# Gráfico de la función de verosimilitud
plot(lambda_vals, likelihood_vals, type = "l", lwd = 2, col = "red",
     xlab = "lambda", ylab = "L(lambda; X=4)", main = "Función de verosimilitud (X=4)")
```



5.5.1.3 3. Representación gráfica en R

```r
par(opt)
```

5.5.2 Explicación de los gráficos

1. **Función de densidad**:
   - Representa cómo varía la probabilidad de observar diferentes valores de x bajo una distribución Poisson con $\lambda = 4$.

- Los valores de x son la variable, mientras que λ es constante.
- Al ser una variable discreta, la función toma valores únicamente en 0, 1, 2 etc.

2. **Función de verosimilitud**:
   - Muestra cómo cambia cuan "verosimil" es observar el valor fijo $X = 4$ según sean los valores de λ.
   - Aquí, X es fijo y λ variable.
   - Es una función que toma valor para cualquier posible valor de λ y por tanto es continua, aunque no es una densidad.

5.6 Ejercicios 6

Sean $X_1, X_2, ..., X_n$ variables aleatorias de Bernoulli independientes tales que $P(X_i = 1) = p$ y $P(X_i = 0) = 1 - p$ para cada $i = 1, 2, 3, ...$ Con la variable aleatoria Y denote el número de intentos necesario para obtener el primer éxito, es decir, el valor de i para el cual $X_i = 1$ ocurre primero. Entonces Y tiene una distribución geométrica con $P(Y = y) = (1 - p)^{y-1}p$, para $y = 1, 2, 3, ...$

Encuentre el estimador del método de momentos para p basado en esta única observación de Y.

5.7 Ejercicio 7

Sean $Y_1, Y_2, ..., Y_n$ variables aleatorias uniformes independientes y distribuidas idénticamente en el intervalo $(0, 3\theta)$. Deduzca el estimador del método de momentos para θ.

5.7.1 Solución

Queremos encontrar el estimador del método de momentos para el parámetro θ basado en una muestra $Y_1, Y_2, ..., Y_n$ de una distribución uniforme en el intervalo $(0, 3\theta)$.

5.7.2 1. Media de la distribución uniforme $(0, 3\theta)$

La media (es decir la esperanza matemática o momento de orden uno) de una variable aleatoria uniforme $U(a, b)$ está dada por:

$$\mu = \frac{a + b}{2}$$

En este caso, los límites de la distribución son $a = 0$ y $b = 3\theta$, por lo que:

$$E(Y) = \frac{0 + 3\theta}{2} = \frac{3\theta}{2}$$

5.7.3 2. Estimador del método de momentos

El método de momentos sustituye el momento muestral correspondiente al momento poblacional en la función que relaciona el parámetro con dicho momento poblacional.

Para el primer momento, que, insistimos, es la esperanza matemática ($E(Y^1)$ tenemos, del apartado anterior:

$$\theta = \frac{2 \cdot E(Y)}{3}$$

Sustituyendo el primer momento poblacional, $\mu_1 = E(Y)$ por el primer momento muestral $\hat{\mu}_1 = \overline{Y}$, obtenemos el estimador del método de momentos:

$$\hat{\theta} = \frac{2\overline{Y}}{3}$$

Esto significa que para una muestra $Y_1, Y_2, \ldots, Y_n$, podemos calcular $\hat{\theta}$ a partir de la media muestral $\overline{Y}$.

5.8 Ejercicio 8

Sean $Y_1, Y_2, \ldots, Y_n$ variables aleatorias independientes y distribuidas idénticamente de una familia de distribución de potencias con parámetros α y $\theta = 3$. Entonces, si $\alpha > 0$,

$$f(y \mid \alpha) = \begin{cases} \alpha y^{\alpha-1}/3^{\alpha}, & 0 \leq y \leq 3 \\ 0, & \text{en cualquier otro punto..} \end{cases}$$

Asumiendo que hemos calculado $E(Y_1) = 3\alpha/(\alpha+1)$ deduzca el estimador del método de momentos para α.

5.9 Ejercicio 9

Suponga que $Y_1, Y_2, \ldots, Y_n$ denotan una muestra aleatoria de la distribución de Poisson con media λ.

1. Encuentre el estimador máximo verosímil (EMV) $\hat{\lambda}$ para λ.
2. Encuentre el valor esperado y la varianza de $\hat{\lambda}$.
3. Demuestre que el estimador del inciso a es consistente para λ.
4. ¿Cuál es el EMV para $P(Y = 0) = e^{-\lambda}$?

5.9.1 Solución

5.9.1.1 Estimador máximo verosímil (EMV) La función de verosimilitud para una muestra aleatoria de tamaño n es:

$$L(\lambda; Y_1, \ldots, Y_n) = \prod_{i=1}^{n} \frac{\lambda^{Y_i} e^{-\lambda}}{Y_i!}$$

Tomando el logaritmo de la función de verosimilitud:

$$\ell(\lambda) = \sum_{i=1}^{n} \left(Y_i \ln(\lambda) - \lambda - \ln(Y_i!) \right)$$

Derivamos con respecto a λ e igualamos a 0 para encontrar el EMV:

$$\frac{\partial \ell(\lambda)}{\partial \lambda} = \sum_{i=1}^{n} \frac{Y_i}{\lambda} - n = 0$$

Resolviendo para λ, obtenemos:

$$\hat{\lambda} = \frac{\sum_{i=1}^{n} Y_i}{n} = \overline{Y}$$

Por lo tanto, el estimador máximo verosímil de λ es:

$$\hat{\lambda} = \overline{Y}$$

Estrictamente hablando, para comprobar que el valor obtenido es un máximo debemos verificar que la segunda derivada log-verosimilitud con respecto a λ es negativa en $\hat{\lambda}$.

5.9.1.2 Segunda derivada de la log-verosimilitud

La función log-verosimilitud es:

$$\ell(\lambda) = \sum_{i=1}^{n} (Y_i \ln(\lambda) - \lambda - \ln(Y_i!))$$

Hemos visto que la primera derivada con respecto a λ es:

$$\frac{\partial \ell(\lambda)}{\partial \lambda} = \sum_{i=1}^{n} \frac{Y_i}{\lambda} - n$$

Por lo que, volviendo a derivar respecto a λ se obtiene:

$$\frac{\partial^2 \ell(\lambda)}{\partial \lambda^2} = \sum_{i=1}^{n} \frac{-Y_i}{\lambda^2}$$

Evaluandola en $\hat{\lambda}$, es decir, sustituyendo $\lambda = \hat{\lambda} = \overline{Y}$ en la segunda derivada:

$$\frac{\partial^2 \ell(\hat{\lambda})}{\partial \lambda^2} = \sum_{i=1}^{n} \frac{-Y_i}{\overline{Y}^2}$$

Dado que $\overline{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}$, se tiene que $Y_i/\overline{Y}$ es positivo para todos los i. Por lo tanto:

$$\frac{\partial^2 \ell(\hat{\lambda})}{\partial \lambda^2} = \frac{-1}{\overline{Y}^2} \sum_{i=1}^{n} Y_i$$

Como $\overline{Y}$ y $\sum_{i=1}^{n} Y_i$ son positivos, la segunda derivada es negativa:

$$\frac{\partial^2 \ell(\hat{\lambda})}{\partial \lambda^2} < 0$$

y, por lo tanto se confirma que $\hat{\lambda}$ es un máximo local para la función log-verosimilitud. Esto valida que el estimador encontrado es el estimador máximo verosímil (EMV).

5.9.1.3 Valor esperado y varianza de $\hat{\lambda}$

5.9.1.3.1 Valor esperado: Dado que $\hat{\lambda} = \overline{Y}$ y $\overline{Y}$ es la media muestral de variables Poisson con media λ, tenemos:

$$E(\hat{\lambda}) = E(\overline{Y}) = \lambda$$

Por lo tanto, $\hat{\lambda}$ es un estimador **insesgado** de λ.

5.9.1.3.2 Varianza: La varianza de la media muestral $\overline{Y}$ es:

$$\text{Var}(\hat{\lambda}) = \text{Var}(\overline{Y}) = \frac{\text{Var}(Y)}{n}$$

Dado que $Y \sim \text{Poisson}(\lambda)$, la varianza de Y es λ, por lo que:

$$\text{Var}(\hat{\lambda}) = \frac{\lambda}{n}$$

5.9.1.4 Consistencia del estimador

Un estimador es consistente si:

1. Es insesgado.
2. Su varianza tiende a 0 cuando $n \to \infty$.

De los resultados anteriores:

- $E(\hat{\lambda}) = \lambda$, por lo que es insesgado.
- $\text{Var}(\hat{\lambda}) = \frac{\lambda}{n}$, que tiende a 0 cuando $n \to \infty$.

Por lo tanto, $\hat{\lambda}$ es un estimador **consistente** de λ.

5.9.1.5 EMV para $P(Y = 0) = e^{-\lambda}$

La probabilidad de $Y = 0$ en una distribución Poisson es:

$$P(Y = 0) = e^{-\lambda}$$

El EMV de una función momótona de un parámetro es la misma función del EMV del parámetro.

Aplicando esta propiedad, que se conoce como _invariancia funcional del EMV a la función $h(\lambda) = P(Y = 0) = e^{-\lambda}$ se obtiene, reemplazando λ por su estimador EMV $\hat{\lambda} = \overline{Y}$ el estimador ma´aximo verosimil de $h(\lambda)$, es decir:

$$\widehat{h(\lambda)} = P(\widehat{Y = 0}) = e^{-\hat{\lambda}} = e^{-\overline{Y}} = h(\hat{\lambda})$$

5.9.1.6 Resumiendo:

1. El estimador máximo verosímil de λ es $\hat{\lambda} = \overline{Y}$.
2. El valor esperado de $\hat{\lambda}$ es $E(\hat{\lambda}) = \lambda$, y su varianza es $\text{Var}(\hat{\lambda}) = \frac{\lambda}{n}$.
3. $\hat{\lambda}$ es un estimador consistente para λ.
4. El EMV de $P(Y = 0)$ es $P(\widehat{Y = 0}) = e^{-\overline{Y}}$.

5.10 Ejercicio 10

Suponga que $Y_1, Y_2, \ldots, Y_n$ denotan una muestra aleatoria de una población distribuida exponencialmente con media θ. Encuentre el MLE de la varianza poblacional θ^2.

6 Intervalos de confianza

6.1 EJERCICIO 1

La distribución del número de huevos puestos por una determinada especie de gallina durante su período de reproducción tiene una media de 35 huevos con una desviación estándar de 18.2. Supongamos que un grupo de investigadores recoge una muestra aleatoria de 45 gallinas de esta especie, cuenta el número de huevos establecidos durante el período de reproducción y registra la media de la muestra. Repiten estas muestras 1.000 veces, y construyen una distribución de las medias de la muestra.

a) ¿Cómo se llama esta distribución?

b) ¿Esperaríamos que la forma de esta distribución fuera simétrica, sesgada o no sesgada? Razona la respuesta.

c) Calcula la variabilidad de esta distribución y di cómo se llama el parámetro que la mide.

d) Supongamos que el presupuesto de los investigadores se reduce y solo pueden recoger muestras aleatorias de 10 gallinas. Se registra la media de las muestras del número de huevos y se repite 1.000 veces, construyendo una nueva distribución de las medias de la muestra. ¿Cómo será la variabilidad de esta nueva distribución comparada con la variabilidad de la distribución original?

6.2 EJERCICIO 2

Un administrador de hospital con la esperanza de mejorar los tiempos de espera decide estimar el tiempo de espera medio de la sala de urgencias (ER) de su hospital. Recopila una muestra aleatoria simple de 64 pacientes y determina el tiempo (en minutos) entre cuando ingresaron al ER hasta que fueron visitados por un médico. Un intervalo de confianza del 95% basado en esta muestra es (128 minutos, 147 minutos), basado en una distribución normal para la media. Determina y razona si las siguientes afirmaciones son verdaderas o falsas.

a) Este intervalo de confianza no es válido, ya que no sabemos si la distribución en la población de los tiempos de espera de ER es normal.

b) Tenemos una confianza del 95% de que el tiempo de espera medio de estos 64 pacientes en una sala de emergencias está entre 128 y 147 minutos.

c) Tenemos una confianza del 95% de que el tiempo de espera medio de todos los pacientes en la sala de emergencias de este hospital está entre 128 y 147 minutos.

d) El 95% de las muestras aleatorias tienen una media muestral entre 128 y 147 minutos.

e) Un intervalo de confianza del 99% sería más estrecho que el intervalo de confianza del 95%, ya que debemos estar más seguros de nuestra estimación.

f) El margen de error es de 9,5 y la media de la muestra es de 137,5.

g) Para reducir el margen de error de un intervalo de confianza del 95% a la mitad de lo que es ahora, tendremos que duplicar el tamaño de la muestra.

6.2.1 SOLUCIÓN

Aquí tienes la traducción de las respuestas al castellano:

a) Falso. El tamaño muestral permite aceptar la normalidad por el TLC.

b) Falso. La inferencia se realiza sobre el parámetro poblacional, no sobre el parámetro muestral, que siempre estará dentro del intervalo.

c) Cierto.

d) Falso. No hacemos inferencia sobre las muestras.

e) Falso. Sería más amplio.

f) Cierto. La media es el punto medio del intervalo.

g) Falso. Deberíamos cuadruplicar el tamaño de la muestra.

6.3 EJERCICIO 3

Las autoridades sanitarias fijan la cantidad de 14 UFP/100ml (UFP=unidades formadoras de placas) como la concentración máxima de un determinado virus entérico en aguas residuales de cualquier punto del estado. Se realiza un control en aguas depuradas de 10 granjas que generan purines. La concentración del virus entérico corresponde a un número muy grande, de forma que podemos asumir que sigue una distribución

Normal. Por otro lado, las granjas están suficientemente alejadas como para asumir que los resultados individuales son mutuamente independientes.

Los valores obtenidos han sido:

14.3	15.3	13.8	15.4	15.5	14.6	13.9	15	14

1. Calcula el intervalo de confianza al 95% de la concentración media del virus en las aguas que vierten las granjas.

2. Interpreta el resultado en función del valor fijado por la administración.

6.3.1 SOLUCIÓN

6.3.1.1 Cálculo del intervalo de confianza al 95% para la concentración media del virus El intervalo de confianza (IC) para la media de una población normal con varianza desconocida se calcula como:

$$\bar{x} \pm t_{n-1,1-\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

Donde: - $\bar{x}$ es la media muestral. - $t_{n-1,1-\alpha/2}$ es el valor crítico de la distribución t de Student con $n-1$ grados de libertad y nivel de confianza $1-\alpha$. - s es la desviación estándar muestral. - n es el tamaño de la muestra.

Los datos proporcionados son:

$$\text{concentración} = \{14.3, 15.3, 13.8, 15.4, 15.5, 14.6, 13.9, 15.0, 14.6, 13.8\}$$

El tamaño muestral es $n = 10$. Calculamos la media ($\bar{x}$) y la desviación estándar (s) con R y usamos la función `t.test` para obtener el intervalo de confianza.

```
# Datos
conc_virus <- c(14.3, 15.3, 13.8, 15.4, 15.5, 14.6, 13.9, 15.0, 14.6, 13.8)

# Intervalo de confianza al 95%
res <- t.test(conc_virus)
res$conf.int
```

```
## [1] 14.14573 15.09427
## attr(,"conf.level")
## [1] 0.95
```

Resultados obtenidos: - Límite inferior del IC: 14.1457 - Límite superior del IC: 15.0943

Por lo tanto, el intervalo de confianza al 95% es:

$$(14.146, 15.094)$$

6.3.1.2 Interpretación del resultado El valor fijado por la administración como límite máximo aceptable para la concentración de virus es 14 UFP/100ml. Observamos que **todo el intervalo de confianza calculado está por encima de este valor (14.146, 15.094)**. Esto implica que:

- Es razonable concluir que la media poblacional de la concentración del virus está por encima de 14 UFP/100ml.

- Solo el 5% de los intervalos que construimos a partir de muestras no contiene la verdadera media poblacional, y el intervalo que hemos construido no incluye el valor fijado por la administración.

Conclusión: Existe evidencia estadística para sugerir, con una alta confianza, que las granjas exceden, en promedio, el límite permitido por las autoridades sanitarias.

6.4 EJERCICIO 4

En un estudio sobre los efectos fisiológicos del alcohol se midió el tiempo que se tarda en reaccionar a un estímulo en un conjunto de seis individuos antes y después de consumir una fuerte dosis de alcohol. El tiempo de latencia medido en milisegundos fue el siguiente:

Individuo	1	2	3	4	5	6
Antes	3.85	3.81	3.60	3.68	3.78	3.83
Después	3.82	3.95	3.80	3.87	3.88	3.94

a) Calcula un intervalo de confianza del 95% para la diferencia de medias: *Después - Antes*.
¿Podríamos afirmar que la media después es superior a la media antes?

b) ¿Cómo cambiará el intervalo si reducimos el nivel de confianza al 90%?
¿Será más amplio? ¿Será más estrecho? ¿O no cambiará?

6.4.1 SOLUCIÓN

6.4.1.1 Cálculo del intervalo de confianza al 95% para la diferencia de medias: *Después - Antes* Estamos trabajando con datos apareados, ya que se trata de mediciones realizadas en los mismos individuos antes y después de consumir alcohol. Por lo tanto, calculamos las diferencias entre los valores *Después - Antes* para cada individuo, y procedemos como si fuera un problema de una sola muestra con las diferencias.

El intervalo de confianza para la media de las diferencias se calcula como:

$$\bar{d} \pm t_{n-1,1-\alpha/2} \cdot \frac{s_d}{\sqrt{n}}$$

Donde: - $\bar{d}$ es la media de las diferencias. - $t_{n-1,1-\alpha/2}$ es el valor crítico de la distribución t de Student con $n-1$ grados de libertad y nivel de confianza $1-\alpha$. - s_d es la desviación estándar de las diferencias. - n es el número de pares.

Los datos son: - Antes: $[3.85, 3.81, 3.60, 3.68, 3.78, 3.83]$ - Después: $[3.82, 3.95, 3.80, 3.87, 3.88, 3.94]$

Calculemos el intervalo de confianza al 95% con R.

```
# Datos
antes <- c(3.85, 3.81, 3.60, 3.68, 3.78, 3.83)
despues <- c(3.82, 3.95, 3.80, 3.87, 3.88, 3.94)

# Intervalo de confianza al 95% para las diferencias pareadas
res_95 <- t.test(despues, antes, paired = TRUE)
res_95$conf.int
```

```
## [1] 0.03092935 0.20573732
## attr(,"conf.level")
## [1] 0.95
```

Resultados: - Límite inferior del IC: 0.03093 - Límite superior del IC: 0.20574

El intervalo de confianza al 95% para la diferencia de medias es:

$$(0.031, 0.206)$$

Interpretación: Como todo el intervalo está en la parte positiva de la recta real, podemos afirmar con un 95% de confianza que la media después de consumir alcohol es superior a la media antes de consumir alcohol.

6.4.1.2 Cambio del intervalo al reducir el nivel de confianza al 90%

Al reducir el nivel de confianza al 90%, el valor crítico $t_{n-1,1-\alpha/2}$ disminuye, lo que hace que el intervalo de confianza sea más estrecho. Calculemos el nuevo intervalo con R.

```
# Intervalo de confianza al 90% para las diferencias pareadas
res_90 <- t.test(despues, antes, paired = TRUE, conf.level = 0.9)
res_90$conf.int
```

```
## [1] 0.0498184 0.1868483
## attr(,"conf.level")
## [1] 0.9
```

Resultados: - Límite inferior del IC: 0.04982 - Límite superior del IC: 0.18685

El intervalo de confianza al 90% es:

$$(0.050, 0.187)$$

Conclusión: Como esperábamos, el intervalo al 90% es más estrecho que el intervalo al 95%. Esto ocurre porque reducimos el nivel de confianza, lo que implica un menor margen de error.

6.5 EJERCICIO 5

El estudio sanguíneo de un individuo presenta 125 neutrófilos de un recuento total de 200 glóbulos blancos. Se pide:

1. Encuentra una estimación puntual para la proporción de neutrófilos.

2. Encuentra un intervalo de confianza al 90% para la anterior proporción.

3. En un individuo sano, el porcentaje de neutrófilos se encuentra entre el 60% y el 70% del total de glóbulos blancos. Según el intervalo del apartado anterior, ¿hay alguna evidencia de desequilibrio de neutrófilos en la muestra de sangre analizada?

6.5.1 SOLUCIÓN

6.5.1.1 Estimación puntual para la proporción de neutrófilos

El estimador (por momentos y por máxima verosimilitud) de la proporción poblacional es la frecuencia relativa ($\hat{p}$) que se calcula como:

$$\hat{p} = \frac{x}{n}$$

Donde: - $x = 125$ es el número de neutrófilos. - $n = 200$ es el total de glóbulos blancos.

Calculamos $\hat{p}$:

```
# Datos
x <- 125
n <- 200

# Proporción muestral
```

```r
p_hat <- x / n
p_hat
```

```
## [1] 0.625
```

Resultado:

$$\hat{p} = 0.625$$

La estimación puntual para la proporción de neutrófilos es **0.625** o **62.5%**.

6.5.1.2 Intervalo de confianza al 90% para la proporción

El intervalo de confianza asintótico para una proporción se calcula como:

$$\hat{p} \pm z_{1-\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Donde:

- $z_{1-\alpha/2}$ es el valor crítico de la distribución normal estándar para un nivel de confianza del 90%.
- $\hat{p}$ es la proporción muestral.
- n es el tamaño de la muestra.

Usamos R para calcular este intervalo:

```r
# Nivel de confianza
alpha <- 0.1
z <- qnorm(1 - alpha / 2)

# Error estándar
se <- sqrt(p_hat * (1 - p_hat) / n)

# Intervalo de confianza
lower <- p_hat - z * se
upper <- p_hat + z * se
c(lower, upper)
```

```
## [1] 0.5686923 0.6813077
```

Resultado: - Límite inferior: **0.569** - Límite superior: **0.681**

El intervalo de confianza al 90% para la proporción de neutrófilos es:

$$(0.569, 0.681)$$

6.5.1.3 Evidencia de desequilibrio de neutrófilos

Un individuo sano tiene una proporción de neutrófilos entre 60% y 70% ($0.6 \leq p \leq 0.7$). Observamos que:

- El intervalo de confianza calculado es $(0.569, 0.681)$.
- Dado que hay una considerable superposición entre ambos intervalos no podemos concluir que haya evidencia de desequilibrio de neutrófilos.

6.6 EJERCICIO 6

Un proceso químico se lleva a cabo usando un catalizador del que se quiere estimar el rendimiento medio. Una muestra piloto de tamaño 8 estima la desviación típica con un valor de 2.

Decide el tamaño de muestra necesario para obtener intervalos de confianza para la media con un 90% y un 95% de confianza de anchura igual a 3 (precisión 1.5), suponiendo que dicha variable sigue un modelo normal.

6.6.1 SOLUCIÓN

Queremos determinar el tamaño de muestra necesario para obtener intervalos de confianza para la media con anchura igual a 3 (precisión de 1.5), considerando niveles de confianza del 90% y 95%. Se sabe que la desviación estándar estimada del proceso químico es $\sigma = 2$ y que la variable sigue una distribución normal.

Dado que la desviación estándar es estimada, nos basaremos en la distribución t de Student en lugar de la normal estándar.

La amplitud del intervalo de confianza para la media basado en la distribución t se calcula como:

$$\text{Amplitud} = 2 \cdot t_{n-1,1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

Resolviendo para n, a partir de la precisión, es decir la mitad de la amplitud se tiene:

$$n = \left(\frac{t_{n-1,1-\alpha/2} \cdot \sigma}{\text{Precisión}} \right)^2$$

Donde: - Precisión $= \frac{\text{Anchura}}{2} = 1.5$. - $t_{n-1,1-\alpha/2}$ es el valor crítico de la distribución t con $n-1$ grados de libertad y nivel de confianza $1-\alpha$. - $\sigma = 2$ es la desviación estándar estimada.

Como $t_{n-1,1-\alpha/2}$ depende de n, este cálculo requiere iteración.

Realizaremos manualmente la iteración aunque también es posible usar algun paquete como `DescTools` para confirmar los resultados.

6.6.1.1 Cálculo manual con iteración

Dado que para un valor fijo de α el valor de t depende de n podemos iterar para calcular n, ajustando el valor de $t_{n-1,1-\alpha/2}$ en cada paso según el tamaño de muestra estimado.

Podemos empezar con un valor bien bajo de n y, a partir de éste ir aumentando el tamaño.

```
s <- 2          # Desviación estándar
precision <- 1.5  # Precisión deseada

find_n <- function(conf_level) {
  n <- 4  # Tamaño inicial
  repeat {
    t_value <- qt(1 - (1 - conf_level) / 2, df = n - 1)
    n_new <- ceiling((t_value * s / precision)^2)
    if (n_new == n) break
    n <- n_new
  }
  return(n)
}

# Tamaños de muestra para 90% y 95%
n_90_iter <- find_n(0.90)
n_95_iter <- find_n(0.95)
c(n_90_iter, n_95_iter)
```

```
## [1]  7 10
```

Resultados del cálculo iterativo: 1. Para un nivel de confianza del **90%**, $n = 7$. 2. Para un nivel de confianza del **95%**, $n = 10$.

Al aumentar el nivel de confianza del 90% al 95%, el tamaño de muestra necesario aumenta debido al incremento del valor crítico t, lo que refleja una mayor necesidad de datos para reducir la incertidumbre del intervalo.

6.7 EJERCICIO 7

En un estudio sobre las alteraciones hormonales que se presentan durante la práctica deportiva se ha medido el aumento de cortisol al realizar una prueba específica de resistencia de 30 minutos. El trabajo se ha realizado con voluntarios de edad y peso similares, pero diferenciados según sus hábitos, separando en dos grupos a los participantes: sedentarios y practicantes habituales de algún deporte. Se supone que la variable medida sigue el modelo normal con varianza común. Se han medido 8 personas de cada grupo.

Se han publicado los siguientes intervalos de confianza (con nivel de confianza del 90%):

- Sedentarios: $(2.85, 4.40)$
- Practicantes de deporte: $(3.52, 5.23)$

a) Calcula las medias muestrales de cada grupo.

b) Si la concentración está expresada en $\mu\mu$g/dl (microgramos por decilitro) y suponemos que se prefiere finalmente presentar los resultados en ng/ml (nanogramos por mililitro), ¿cómo quedarían afectados los intervalos de confianza iniciales?

6.8 EJERCICIO 8

Se reporta el siguiente listado del análisis del nivel de colesterol en una muestra de 30 individuos obesos. Desafortunadamente, algunas partes del listado se han vuelto ilegibles y su valor se ha sustituido por 9999.

```
One Sample t-test
data: x
t $=301.49$, df $=9999, \mathrm{p}$-value $<2.2 \mathrm{e}-16$
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
246.8329999 .000
sample estimates:
mean of x
248.5179
```

Reconstruye los valores incorrectos del listado.

7 Contrastes de Hipótesis

7.1 Ejercicio 1.

Un investigador ha preparado un nivel de dosis de droga que según él, inducirá el sueño en 80% de las personas que sufren de insomnio. Después de examinar la dosis, pensamos que lo dicho por él respecto a la efectividad de la dosis es exagerado. En un intento por refutar su dicho, administramos la dosis prescrita a 20 personas que padecen de insomnio y observamos Y, el número de individuos a quienes la dosis induce el sueño. Deseamos probar la hipótesis $H_0 : p = .8$ contra la alternativa, $H_a : p < .8$. Suponga que se usa la región de rechazo $\{y \leq 12\}$.

a. De acuerdo con la información de este problema, ¿qué es un error tipo I?
b. Encuentre α.

c. Con base en la información de este problema, ¿qué es un error tipo II?
d. Encuentre β cuando $p = .6$.
e. Encuentre β cuando $p = .4$.

7.2 Ejercicio 2

Continuando con el ejercicio 1

a. Defina la región de rechazo de la forma $\{y \leq c\}$ de modo que $\alpha \approx .01$.
b. Para la región de rechazo del inciso a, encuentre β cuando $p = .6$.
c. Para la región de rechazo del inciso a, encuentre β cuando $p = .4$.

7.3 Ejercicio 3.

Nos interesa probar si una moneda está o no balanceada, con base en el número de caras Y en 36 tiros de la moneda. ($H_0 : p = .5$ contra $H_a : p \neq .5$). Si usamos la región de rechazo $|y - 18| \geq 4$, ¿ cuál es

a. el valor de α ?
b. el valor de β si $p = .7$?

7.4 Ejercicio 4.

Verdadero o falso Consulte el Ejercicio 3

a. El nivel de la prueba calculado en el Ejercicio 3(a) es la probabilidad de que H_0 sea verdadera.
b. El valor de β calculado en el Ejercicio 3(b) es la probabilidad de que H_a sea verdadera.
c. En el Ejercicio 3(b), β se calculó suponiendo que la hipótesis nula era falsa.
d. Si β se calculó cuando $p = 0.55$, el valor sería más grande que el valor de β obtenido en el Ejercicio 3(b).
e. La probabilidad de que la prueba equivocadamente rechace H_0 es β.
f. Suponga que la región de rechazo (RR) se cambió a $|y - 18| \geq 2$.
    1. Esta RR llevaría a rechazar la hipótesis nula con más frecuencia que la RR empleada en el Ejercicio 3
    2. Si α se calculó usando esta nueva RR, el valor sería más grande que el valor obtenido en el Ejercicio 3(a)
    3. Si β se calculó cuando $p = .7$ y usando esta nueva RR , el valor sería más grande que el valor obtenido en el Ejercicio 3(b).

7.5 Ejercicio 5.

Una prueba clínica en dos etapas está planeada para probar $H_0 : p = .10$ contra $H_a : p > .10$, donde p es la proporción de pacientes que responden a un tratamiento y que fueron tratados según el protocolo. En la primera etapa, 15 pacientes se acumularon y trataron. Si 4 o más de los que responden se observan entre los (primeros) 15 pacientes, H_0 es rechazada, el estudio se termina y no se acumulan más pacientes. De otro modo, otros 15 pacientes se acumularán y tratarán en la segunda etapa. Si un total de 6 o más de los que responden se observan entre los 30 pacientes acumulados en las dos etapas (15 en la primera etapa y 15 más en la segunda etapa), entonces H_0 es rechazada. Por ejemplo, si 5 de los que responden se encuentran entre los pacientes de la primera etapa, H_0 es rechazada y el estudio se termina. No obstante, si 2 de los que responden se encuentran entre los pacientes de la primera etapa, se acumulan 15 pacientes de la segunda etapa y se identifican otros 4 o más de los que responden (para un total de 6 o más entre los 30), H_0 es rechazada y el estudio termina. [1]

a. Utilice la tabla binomial para hallar el valor numérico de α para este procedimiento de prueba.

b. Utilice la tabla binomial para determinar la probabilidad de rechazar la hipótesis nula cuando use esta región de rechazo si $p = .30$.
c. Para la región de rechazo definida líneas antes, encuentre β si $p = .30$.

8 Aplicaciones de los contrastes de hipótesis

8.1 Elección del tipo de test

Estos ejercicios representan situaciones típicas en los que se puede utilizar un contraste de hipótesis para decidir entre dos hipótesis relativas a un problema.

En una situación real debemos (solemos) empezar con una visualización de los datos y, eventualmente, cuando procede (en variables cuantitativas) un test de normalidad que nos sirve para decidir que tipo de contraste utilizar.

Estos ejercicios se han adaptado del excelente libro de Baron y otros (Bioestadística) de la Universidad de Málaga, pero por su naturaleza más académica, no vienen acompañados de conjuntos de datos mínimamente grandes, como para poder hacer dichas visualizaciones y pruebas de normalidad.

En consecuencia, la "decisión" sobre si utilizar un test paramétrico o no paramétrico puede ser más difícil de tomar.

A título de pista podemos decir que los ejercicios 1 al 9 están tomados del capítulo de pruebas paramétricas, del 10 al 13 del capítulo de pruebas de la ji cuadrado y del 14 al 18 del capítulo de pruebas no paramétricas de dicho documento.

Sin embargo, y con el fin de motivar un trabajo reflexivo, os animo a justificar en cada caso porque decidís aplicar el tests que aplicáis. Y, de hecho, cuando sea posible, os animo también a aplicar tanto la versión paramétrica como la no paramétrica de un test y valorar, aparte de cual es el adecuado, si os llevan o no ala misma conclusión.

8.2 Procedimiento del test (Neymann-Pearson)

Recordad que la aplicación de un contraste no consiste en escribir un código de R, mirar un p-valor y decidir si rechazamos o no la hipótesis.

Aunque no es bueno adherirse a un esquema rígido si que resulta útil tener una mínima pauta que podemos considerar común a la mayoría de los problemas en los que aplicamos el contraste de hipótesis.

Esta pauta puede describirse de la forma siguiente:

I: ***ANTES DE OBTENER LOS DATOS*** *deberíamos:*

0. Definir el modelo probabilístico sobre el que realizaremos el contraste
1. Definir el tamaño del efecto a detectar en la población (en estos ejercicios suele ser cero)
2. Reformular la pregunta de investigación en términos de las hipótesis nula y alternativa (H_0 y H_1)
3. Elegir el nivel de significación α (e, idealmente la potencia $1-\beta$) con los que controlaremos el error de tipo I y de tipo II (β) respectivamente. Habitualmente tan sólo podemos prefijar α
4. Escoger el estadístico de test adecuado al problema, el modelo y las hipótesis. Idealmente desearemos que sea un test *óptimo*
5. Calcular el tamaño de la muestra necesario para alcanzar la potencia deseada.
6. Obtener el valor crítico t_α asociado al estadístico de test seleccionado. Esto equivale a definir la región crítica del test.

II: Proceder a recoger los datos

III: ***UNA VEZ SE HAN OBTENIDO LOS DATOS***

7. Calcular el valor del estadístico de test en la muestra, t^{obs}
8. Decidir si, a la vista de la comparación entre t_α y t^{obs} nos inclinamos por la hipótesis nula o la alternativa.
9. Interpretar la decisión en términos de la hipótesis científica que ha generado el problema. Diferenciar cuando sea posible si hay significación estadística, clínica/biológica o ambas

En la práctica, y en los ejercicios que siguen, podemos encontrarnos con los datos recogidos y desear llevar a cabo el contraste.

En este caso, omitiremos el cálculo del tamaño muestral y la potencia.

Esta lista no es una plantilla para resolver cualquier problema de contraste de hipótesis, pero sí que debe ayudarnos a ver que un test no es tan solo un cálculo y una decisión basada en un p-valor sino un proceso de toma de decisión que, para que tenga sentido debe de realizarse teniendo en cuenta los elementos que se han descrito.

8.3 ¿Y que hay del p-valor? (Fisher)

Obsérvese que el procedimiento descrito no ha hecho referencia al p-valor, concepto particularmente discutido en la estadística actual.

El p-valor, de hecho, no forma parte del esquema presentado, basado en la teoría de contraste de hipótesis de Neymann-Pearson sino de la teoría de las pruebas de significación estadística introducida por R.A. Fisher.

Dichas pruebas de significación iban principalmente encaminadas a decidir si se rechazaba o no una hipótesis (no había hipótesis alternativa) e, informalmente los pasos a realizar eran:

1. Seleccionar un estadístico de test que mida adecuadamente la discrepancia entre la hipótesis y las observaciones.
2. Construir la hipótesis nula
3. Calcular la probabilidad teórica del resultado observado t^{obs}, bajo la hipótesis nula, es decir el p-valor.
4. Decidir sobre la significación de los resultados, si la probabilidad de los mismos *bajo* H_0 era "pequeña" (Fisher no imponía puntos de corte). Es decir rechazar la hipótesis nula si el p-valor es pequeño.
5. Interpretar los resultados.

8.4 Combinando ambas aproximaciones

Fisher y Neymann-Pearson debatieron duramente, hace casi un siglo, sobre cual de las dos aproximaciones era la adecuada.

Sin entrar en la polémica entre ambas, el hecho es que con el tiempo, muchos autores tendieron a fundir (y a confundir!) ambas aproximaciones, lo que llevó a combinar las hipótesis nula y alternativa de Neymann-pearson con el p-valor de Fisher.

En la actualidad, en la que la mayoría de los cálculos se llevan a cabo con ordenador, es inmediato, tras calcular el valor del estadístico de test, obtener el p-valor de la prueba por lo que la decisión sobre si se acepta la hipótesis nula o la alternativa se realiza a menudo comparando el p-valor con la probabilidad de error de tipo I (α) en vez de comparar el valor observado del estadístico de test (t^{obs}) con el valor crítico (t_α). En la práctica, *esta sustitución puede considerarse adecuada si no va más allá de cambiar un criterio de decisión por otro.*

8.5 Referencias

- Fisher, Neyman-Pearson or NHST? A tutorial for teaching data testing

9 Ejercicios

9.1 Ejercicio 1

El calcio se presenta normalmente en la sangre de los mamíferos en concentraciones de alrededor de 6 mg por cada 100 ml del total de sangre. La desviación típica normal de ésta variable es 1 mg de calcio por cada 100 ml del volumen total de sangre. Una variabilidad mayor a ésta puede ocasionar graves trastornos en la coagulación de la sangre. Una serie de nueve pruebas sobre un paciente revelaron una media muestral de $6,2$mg de calcio por 100 ml del volumen total de sangre, y una desviación típica muestral de 2 mg de calcio por cada 100 ml de sangre. ¿Hay alguna evidencia, para un nivel $\alpha = 0,05$, de que el nivel medio de calcio para este paciente sea más alto del normal?

9.2 Ejercicio 2

Las guías médicas recomiendan realizar una campaña de educación e higiene dental si el porcentaje de niños con dientes cariados es superior al 15%. En una población con 12.637 niños, ¿debe hacerse la campaña si, de 387 de ellos, 70 tenían algún diente cariado?

9.3 Ejercicio 3

Muchos autores afirman que los pacientes con depresión tienen una función cortical por debajo de lo normal debido a un riego sanguíneo cerebral por debajo de lo normal. A dos muestras de individuos, unos con depresión y otros normales, se les midió un índice que indica el flujo sanguíneo en la materia gris (dado en mg/(100 g/min)) obteniéndose:

Depresivos	$n_1 = 19$	$\bar{x}_1 = 47$	$\hat{\mathcal{S}}_1 = 7'8$
Normales	$n_2 = 22$	$\bar{x}_2 = 53'8$	$\hat{\mathcal{S}}_2 = 6'1$

¿Hay evidencia significativa a favor de la afirmación de los autores?

9.4 Ejercicio 4

La prueba de la d-xilosa permite la diferenciación entre una esteatorrea originada por una mala absorción intestinal y la debida a una insuficiencia pancreática, de modo que cifras inferiores a 4 grs. de d-xilosa, indican una mala absorción intestinal. Se realiza dicha prueba a 10 individuos, obteniéndose una media de 3,5 grs. y una desviación típica de 0,5 grs. ¿Se puede decir que esos pacientes padecen una mala absorción intestinal?

9.5 Ejercicio 5

La tabla siguiente muestra los efectos de un placebo y de la hidroclorotiacida sobre la presión sanguínea sistólica de 11 pacientes.

Placebo	211	210	210	203	196	190	191	177	173	170	163
H-cloro	181	172	196	191	167	161	178	160	149	119	156

Según estos datos experimentales, ¿podemos afirmar que existe diferencia en la presión sistólica media durante la utilización de estos dos fármacos?

9.6 Ejercicio 6

De un estudio sobre la incidencia de la hipertensión en la provincia de Málaga, se sabe que en la zona rural el porcentaje de hipertensos es del $27,7\%$. Tras una encuesta a 400 personas de una zona urbana, se obtuvo un 24% de hipertensos.

1. ¿Se puede decir que el porcentaje de hipertensos en la zona urbana es distinto que en la zona rural?
2. ¿Es menor el porcentaje de hipertensos en la zona urbana que en la zona rural?

9.7 Ejercicio 7

Se desea comparar la actividad motora espontánea de un grupo de 25 ratas control y otro de 36 ratas desnutridas. Se midió el número de veces que pasaban delante de una célula fotoeléctrica durante 24 horas. Los datos obtenidos fueron los siguientes:

Ratas de control	$n_1 = 25$	$\bar{x}_1 = 869,8$	$\mathcal{S}_1 = 106,7$
Ratas desnutridas	$n_2 = 36$	$\bar{x}_2 = 465$	$\mathcal{S}_2 = 153,7$

¿Se observan diferencias significativas entre el grupo control y el grupo desnutrido?

9.8 Ejercicio 8

Se pretende comprobar la hipótesis expuesta en algunos trabajos de investigación acerca de que la presencia del antígeno AG-4 está relacionada con un desenlace fatal. Con éste fin, se hizo una revisión sobre las historias clínicas de 21 mujeres muertas por carcinoma de cuello uterino, observando que 6 de ellas presentaban el citado antígeno. Por otro lado y con fines de comparación se tomó otra muestra de 42 personas, con edades similares a las del grupo anterior y que reaccionaron bien al tratamiento del carcinoma de cuello uterino, en 28 de las cuales se observó la presencia del citado antígeno. ¿Puede pues afirmarse que la presencia del antígeno se relaciona con una efectividad del tratamiento?

9.9 Ejercicio 9

Para comprobar si la tolerancia a la glucosa en sujetos sanos tiende a decrecer con la edad se realizó un test oral de glucosa a dos muestras de pacientes sanos, unos jóvenes y otros adultos. El test consistió en medir el nivel de glucosa en sangre en el momento de la ingestión (nivel basal) de 100 grs. de glucosa y a los 60 minutos de la toma. Los resultados fueron los siguientes:

Jóvenes:	Basal	81	89	80	75	74	97	76	89	83
60 minutos	136	150	149	141	138	154	141	155	145	147

Adultos:	Basal	98	94	93	88	79	90	86	89	81
60 minutos	196	190	191	189	159	185	182	190	170	197

1. ¿Se detecta una variación significativa del nivel de glucosa en sangre en cada grupo?
2. ¿Es mayor la concentración de glucosa en sangre a los 60 minutos, en adultos que en jóvenes?
3. El contenido basal de glucosa en sangre, ¿es menor en jóvenes que en adultos?
4. ¿Se detecta a los 60 minutos una variación del nivel de glucosa en sangre diferente de los adultos, en los jóvenes?

9.10 Ejercicio 10

Ante la sospecha de que el hábito de fumar de una embarazada puede influir en el peso de su hijo al nacer, se tomaron dos muestras, una de fumadoras y otra de no fumadoras, y se clasificó a sus hijos en tres categorías en función de su peso en relación con los percentiles $\mathcal{P}_{10}$ y $\mathcal{P}_{90}$ de la población. El resultado se expresa en la tabla siguiente:

Peso del niño			
¿Madre fumadora?	Menor de $\mathcal{P}_{10}$	Entre $\mathcal{P}_{10}$ y $\mathcal{P}_{90}$	Mayor de $\mathcal{P}_{90}$
	117	529	19
	124	1147	117

¿Hay una evidencia significativa a favor de la sospecha a la vista de los resultados de la muestra?

9.11 Ejercicio 11

Varios libros de Medicina Interna recomiendan al médico la palpación de la arteria radial con el fin de evaluar el estado de la pared arterial. Se tomaron 215 pacientes y se les clasificó según la palpabilidad de dicha arteria (grados 0,1 y 2 para no palpable, palpable y muy palpable o dura, respectivamente) y según una puntuación de 0 a 4 en orden creciente de degeneración arterial (evaluada tras la muerte del paciente y su análisis anatomo-patológico). Los datos son los de la tabla siguiente:

	Palpabilidad		
Degeneración	**0**	**1**	**2**
0	20	5	5
1	60	20	10
2	45	15	15
3	10	5	5

¿Existe relación entre el grado de palpabilidad y el análisis anatomopatológico?

9.12 Ejercicio 12

Con el fin de conocer si un cierto tipo de bacterias se distribuyen al azar en un determinado cultivo o si, por el contrario, lo hacen con algún tipo de preferencia (el centro, los extremos, etc…), se divide un cultivo en 576 áreas iguales y se cuenta el número de bacterias en cada área. Los resultados son los siguientes:

n^o de bacterias	0	1	2	3	4	≥ 5
	n^o de áreas	229	211	93	35	7

¿Obedecen los datos a una distribución de Poisson?

9.13 Ejercicio 13

Deseamos conocer, si las distribuciones atendiendo al grupo sanguíneo, en tres muestras referidas atendiendo al tipo de tensión arterial, se distribuyen de igual manera. Para lo cual, se reunió una muestra de 1500 sujetos a los que se les determinó su grupo sanguíneo y se les tomó la tensión arterial, clasificándose ésta en baja, normal, y alta. Obteniéndose los siguientes resultados:

	Grupo sanguíneo				
Tensión arterial	A	B	AB	O	Total
Baja	28	9	7	31	75
Normal	543	211	90	476	1.320
Alta	44	22	8	31	105
Total	615	242	105	538	1.500

9.14 Ejercicio 14

Se realiza un estudio para determinar los efectos de poner fin a un bloqueo renal en pacientes cuya función renal está deteriorada a causa de una metástasis maligna avanzada de causa no urológica. Se mide la tensión arterial de cada paciente antes y después de la operación. Se obtienen los siguientes resultados:

Tensión arterial

Antes	150	132	130	116	107	100	101	96	90	78
Después	90	102	80	82	90	94	84	93	89	8?????

¿Se puede concluir que la intervención quirúrgica tiende a disminuir la tensión arterial?

9.15 Ejercicio 15

Se ensayaron dos tratamientos antirreumáticos administrados al azar, sobre dos grupos de 10 pacientes, con referencia a una escala convencional (a mayor puntuación, mayor eficacia), valorada después del tratamiento. Los resultados fueron:

Nivel de eficacia del tratamiento

Tratamiento primero	12	15	21	17	38	42	10	23	35	28
Tratamiento segundo	21	18	25	14	52	65	40	43	35	42

Decidir si existe diferencia entre los tratamientos.

9.16 Ejercicio 16

Los siguientes datos nos dan el peso de comida (en Kg.) consumidos por adulto y día en diferentes momentos en un año. Usar un contraste no paramétrico para comprobar si el consumo de comida es el mismo en los 4 meses considerados.

Febrero	Mayo	Agosto	Noviembre
4′7	4′7	4′8	4′9
4′9	4′4	4′7	5′2
5′0	4′3	4′6	5′4
4′8	4′4	4′4	5′1
4′7	4′1	4′7	5′6

9.17 Ejercicio 17

Se hizo un estudio neurofisiológico sobre la conducción motora tibial posterior en dos grupos de pacientes embarazadas con las siguientes determinaciones:

Conducción motora tibial posterior

Primer grupo	51	40	41	53	48	50	45	58	45	44
Segundo grupo	58	43	40	45	41	42	44	52	56	48

Comprobar la igualdad o no de ambas muestras.

9.18 Ejercicio 18

En un experimento diseñado para estimar los efectos de la inhalación prolongada de óxido de cadmio, 15 animales de laboratorio sirvieron de sujetos para el experimento, mientras que 10 animales similares sirvieron de controles. La variable de interés fue el nivel de hemoglobina después del experimento. Se desea saber si puede concluirse que la inhalación prolongada de óxido de cadmio disminuye el nivel de hemoglobina según los siguientes datos que presentamos:

Nivel de hemoglobina

Expuestos	14′4	14′2	13′8	16′5	14′1	16′6	15′9	15′6	14′1	15′3
	15′7	16′7	13′7	15′3	14′0					
No expuestos	17′4	16′2	17′1	17′5	15′0	16′0	16′9	15′0	16′3	16′8

9.19 Ejercicio 19

A 11 ratas tratadas crónicamente con alcohol se les midió la presión sanguínea sistólica antes y después de 30 minutos de administrarles a todas ellas una cantidad fija de etanol, obteniéndose los datos siguientes:

Presión sanguínea sistólica

Antes	126	120	124	122	130	129	114	116	119	112	118
Después	119	116	117	122	127	122	110	120	112	110	111

¿Hay un descenso significativo de la presión sanguínea sistólica tras la ingestión de etanol?