Inferencia Estadística

Solución: Estimación y contraste paramétrico 2

DSLab

noviembre, 2024

Ejercicio 1: Propiedades de los estimadores

Pregunta: Sera $X_1, \dots, X_3$ y X_4 una muestra aletatoria de tamaño cuatro de una población cuya distribución es exponencial con parámetros θ desconocido. De las siguientes estadísticas, ¿cuáles son estimadores insesgados de θ?

$$T_1 = \frac{1}{3}(X_1 + X_2) + \frac{1}{6}(X_3 + X_4)$$

$$T_2 = \frac{(4X_1 + 3X_2 + 2X_3 + 1X_4)}{5}$$

$$T_1 = \frac{(X_1 + X_2 + X_3 + X_4)}{4}$$

Solución: Para determinar si las estadísticas T_1, T_2 y T_3 son estimadores insesgados del parámetro θ de una distribución exponencial, necesitamos calcular la esperanza matemática de cada una y verificar si es igual a θ.

Primero, recordemos que si X_i sigue una distribución exponencial con parámetro θ, entonces la esperanza matemática de X_i es:

$$E[X_i] = \theta$$

Ahora, procedemos a analizar cada uno de los estimadores.

Estimador T_1

$$T_1 = \frac{1}{3}(X_1 + X_2) + \frac{1}{6}(X_3 + X_4)$$

Calculamos la esperanza matemática de T_1:

$$E[T_1] = E\left[\frac{1}{3}(X_1 + X_2) + \frac{1}{6}(X_3 + X_4)\right]$$

Usando la linealidad de la esperanza, tenemos:

$$E[T_1] = \frac{1}{3}E[X_1] + \frac{1}{3}E[X_2] + \frac{1}{6}E[X_3] + \frac{1}{6}E[X_4]$$

Dado que $E[X_i] = \theta$ para todos i:

$$E[T_1] = \frac{1}{3}\theta + \frac{1}{3}\theta + \frac{1}{6}\theta + \frac{1}{6}\theta$$

$$E[T_1] = \frac{1}{3}\theta + \frac{1}{3}\theta + \frac{1}{6}\theta + \frac{1}{6}\theta = \theta$$

Por lo tanto, T_1 es un estimador insesgado de θ.

Estimador T_2

$$T_2 = \frac{4X_1 + 3X_2 + 2X_3 + X_4}{5}$$

Calculamos la esperanza matemática de T_2:

$$E[T_2] = E\left[\frac{4X_1 + 3X_2 + 2X_3 + X_4}{5}\right]$$

Usando la linealidad de la esperanza, tenemos:

$$E[T_2] = \frac{4}{5}E[X_1] + \frac{3}{5}E[X_2] + \frac{2}{5}E[X_3] + \frac{1}{5}E[X_4]$$

Dado que $E[X_i] = \theta$ para todos i:

$$E[T_2] = \frac{4}{5}\theta + \frac{3}{5}\theta + \frac{2}{5}\theta + \frac{1}{5}\theta$$

$$E[T_2] = \left(\frac{4+3+2+1}{5}\right)\theta = \left(\frac{10}{5}\right)\theta = 2\theta$$

Por lo tanto, T_2 no es un estimador insesgado de θ.

Estimador T_3

$$T_3 = \frac{X_1 + X_2 + X_3 + X_4}{4}$$

Calculamos la esperanza matemática de T_3:

$$E[T_3] = E\left[\frac{X_1 + X_2 + X_3 + X_4}{4}\right]$$

Usando la linealidad de la esperanza, tenemos:

$$E[T_3] = \frac{1}{4}E[X_1] + \frac{1}{4}E[X_2] + \frac{1}{4}E[X_3] + \frac{1}{4}E[X_4]$$

Dado que $E[X_i] = \theta$ para todos i:

$$E[T_3] = \frac{1}{4}\theta + \frac{1}{4}\theta + \frac{1}{4}\theta + \frac{1}{4}\theta$$

$$E[T_3] = \left(\frac{1+1+1+1}{4}\right)\theta = \left(\frac{4}{4}\right)\theta = \theta$$

Por lo tanto, T_3 es un estimador insesgado de θ.

Conclusión

De las estadísticas dadas, T_1 y T_3 son estimadores insesgados de θ. T_2 no es un estimador insesgado de θ.

Ejercicio 2: Propiedades de los estimadores

Pregunta: En un experimento binomial se observan x éxitos en n ensayos independientes. Se propone el siguiente estadístico como estimador del parámetro de proporción p:

$$T = \frac{x}{n}$$

Se pide demostrar que el estadístico T es un estimador consistente del parámetro binomial p.

Solución: Para demostrar que el estadístico $T = \frac{x}{n}$ es un estimador consistente del parámetro binomial p, necesitamos mostrar que T cumple con dos condiciones:

1. **No sesgado (o asintóticamente no sesgado)**: La esperanza matemática de T es igual a p o converge a p cuando n tiende a infinito.
2. **Varianza que tiende a cero**: La varianza de T tiende a cero cuando n tiende a infinito.

Vamos a verificar ambas condiciones para T.

Paso 1: Verificar que T es un estimador no sesgado de p

Sabemos que x sigue una distribución binomial con parámetros n y p, es decir, $x \sim \text{Binomial}(n, p)$. La esperanza matemática de x es:

$$E[x] = np$$

El estimador T se define como:

$$T = \frac{x}{n}$$

Calculamos la esperanza matemática de T:

$$E[T] = E\left[\frac{x}{n}\right] = \frac{1}{n}E[x] = \frac{1}{n} \cdot np = p$$

Esto muestra que T es un estimador no sesgado de p.

Paso 2: Verificar que la varianza de T tiende a cero cuando n tiende a infinito

La varianza de x para una distribución binomial es:

$$\text{Var}(x) = np(1-p)$$

Queremos encontrar la varianza de T:

$$\text{Var}(T) = \text{Var}\left(\frac{x}{n}\right) = \frac{1}{n^2}\text{Var}(x) = \frac{1}{n^2} \cdot np(1-p) = \frac{p(1-p)}{n}$$

Ahora, observamos el comportamiento de $\text{Var}(T)$ cuando n tiende a infinito:

$$\lim_{n\to\infty} \text{Var}(T) = \lim_{n\to\infty} \frac{p(1-p)}{n} = 0$$

Esto muestra que la varianza de T tiende a cero a medida que n aumenta.

Dado que T es un estimador no sesgado de p y su varianza tiende a cero cuando n tiende a infinito, podemos concluir que T es un estimador consistente de p. Por lo tanto, hemos demostrado que el estadístico $T = \frac{x}{n}$ es un estimador consistente del parámetro de proporción p.

Ejercicio 3: Método de Máxima Verosimilitud

Pregunta: Sea $X_1 \ldots, X_n$ una muestra aleatoria de una población cuya distribución es de Poisson con parámetro λ. Obtener el estimador de máxima verosimilitud de λ.

Solución:

Para obtener el estimador de máxima verosimilitud (MLE, por sus siglas en inglés) del parámetro λ de una distribución de Poisson basada en una muestra aleatoria $X_1, \ldots, X_n$, seguimos estos pasos:

Paso 1: Función de verosimilitud

La función de verosimilitud para una muestra $X_1, \ldots, X_n$ de una distribución de Poisson con parámetro λ es:

$$L(\lambda) = P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n \mid \lambda)$$

Como las X_i son independientes y cada X_i sigue una distribución de Poisson, la función de verosimilitud se puede escribir como:

$$L(\lambda) = \prod_{i=1}^{n} P(X_i = x_i \mid \lambda)$$

La función de probabilidad de una variable aleatoria X_i que sigue una distribución de Poisson es:

$$P(X_i = x_i \mid \lambda) = \frac{\lambda^{x_i} e^{-\lambda}}{x_i!}$$

Por lo tanto, la función de verosimilitud es:

$$L(\lambda) = \prod_{i=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!}$$

Paso 2: Log-verosimilitud

Para simplificar el proceso de maximización, trabajamos con el logaritmo de la función de verosimilitud, conocida como log-verosimilitud:

$$\ell(\lambda) = \log L(\lambda) = \log \left(\prod_{i=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} \right)$$

Usando las propiedades del logaritmo, esto se convierte en:

$$\ell(\lambda) = \sum_{i=1}^{n} \log \left(\frac{\lambda^{x_i} e^{-\lambda}}{x_i!} \right)$$

Descomponemos la log-verosimilitud:

$$\ell(\lambda) = \sum_{i=1}^{n} \left(\log(\lambda^{x_i}) + \log(e^{-\lambda}) - \log(x_i!) \right)$$

$$\ell(\lambda) = \sum_{i=1}^{n} \left(x_i \log \lambda - \lambda - \log(x_i!) \right)$$

Podemos separar los términos que dependen de λ de los que no:

$$\ell(\lambda) = \sum_{i=1}^{n} x_i \log \lambda - \sum_{i=1}^{n} \lambda - \sum_{i=1}^{n} \log(x_i!)$$

Paso 3: Derivada de la log-verosimilitud

Para encontrar el MLE, derivamos $\ell(\lambda)$ con respecto a λ y establecemos la derivada igual a cero:

$$\frac{\partial \ell(\lambda)}{\partial \lambda} = \frac{\partial}{\partial \lambda} \left(\sum_{i=1}^{n} x_i \log \lambda - n\lambda - \sum_{i=1}^{n} \log(x_i!) \right)$$

Derivamos término a término:

$$\frac{\partial \ell(\lambda)}{\partial \lambda} = \sum_{i=1}^{n} x_i \cdot \frac{1}{\lambda} - n$$

Simplificando, obtenemos:

$$\frac{\partial \ell(\lambda)}{\partial \lambda} = \frac{1}{\lambda} \sum_{i=1}^{n} x_i - n$$

Establecemos la derivada igual a cero para encontrar el valor de λ que maximiza la log-verosimilitud:

$$\frac{1}{\lambda} \sum_{i=1}^{n} x_i - n = 0$$

$$\sum_{i=1}^{n} x_i = n\lambda$$

Paso 4: Resolver para λ

Despejamos λ:

$$\lambda = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Resultado

El estimador de máxima verosimilitud (MLE) del parámetro λ es:

$$\hat{\lambda} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Este resultado muestra que el MLE de λ es simplemente la media muestral de los datos $X_1, \ldots, X_n$.

Ejercicio 4: Tamaño muestral

Pregunta: Una empresa quiere realizar una encuesta para estimar el nivel de satisfacción de sus clientes con un nuevo producto. La empresa desea que la estimación tenga un nivel de confianza del 95% y un margen de error del 5%. Basado en estudios previos, se espera que la proporción de clientes satisfechos sea aproximadamente del 70%. ¿Cuál es el tamaño muestral mínimo necesario para la encuesta?

Solución Para calcular el tamaño muestral mínimo necesario para estimar el nivel de satisfacción de los clientes con un nuevo producto con un nivel de confianza del 95% y un margen de error del 5%, podemos utilizar la fórmula para el tamaño de muestra en estimaciones de proporciones:

$$n = \frac{Z^2 \cdot p \cdot (1-p)}{E^2}$$

Donde:

- Z es el valor crítico del nivel de confianza (para el 95%, $Z \approx 1.96$)
- p es la proporción estimada (en este caso, 0.70)
- E es el margen de error deseado (en este caso, 0.05)

Sustituyendo los valores en la fórmula:

$$n = \frac{(1.96)^2 \cdot 0.70 \cdot (1-0.70)}{(0.05)^2} = 322.6944$$

Dado que el tamaño muestral siempre se redondea al siguiente número entero, el tamaño muestral mínimo necesario es:

$$n \approx 323$$

Por lo tanto, la empresa necesita encuestar al menos a 323 clientes para estimar el nivel de satisfacción con un nivel de confianza del 95% y un margen de error del 5%.

Ejercicio 5: Intervalos de confianza

Pregunta: Una empresa quiere comparar la efectividad de dos programas de entrenamiento distintos para sus empleados. Para ello, selecciona dos grupos independientes de empleados y aplica un programa de entrenamiento diferente a cada grupo. Después del entrenamiento, se mide el rendimiento de cada empleado. Los resultados son los siguientes:

- Grupo 1 (Programa A): 15 empleados, media de rendimiento = 80, desviación estándar = 10
- Grupo 2 (Programa B): 12 empleados, media de rendimiento = 75, desviación estándar = 12

Se desea calcular un intervalo de confianza del 95% para la diferencia de las medias de rendimiento entre los dos programas de entrenamiento.

Solución:

Para calcular el intervalo de confianza para la diferencia de medias de dos muestras independientes, usamos la siguiente fórmula:

$$(\bar{X}_1 - \bar{X}_2) \pm Z \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Donde: - $\bar{X}_1$ y $\bar{X}_2$ son las medias muestrales de los dos grupos. - s_1 y s_2 son las desviaciones estándar de los dos grupos. - n_1 y n_2 son los tamaños de muestra de los dos grupos. - Z es el valor crítico para el nivel de confianza deseado (para el 95%, $Z \approx 1.96$).

Primero, identificamos los valores:

- $\bar{X}_1 = 80, s_1 = 10, n_1 = 15$
- $\bar{X}_2 = 75, s_2 = 12, n_2 = 12$
- $Z = 1.96$ para un intervalo de confianza del 95%

Luego, calculamos la desviación estándar combinada de la diferencia de medias:

$$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{10^2}{15} + \frac{12^2}{12}} \approx 4.32$$

Finalmente, calculamos el intervalo de confianza:

$$(\bar{X}_1 - \bar{X}_2) \pm Z \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$(80 - 75) \pm 1.96 \cdot 4.32 \approx 5 \pm 8.4672$$

Entonces, el intervalo de confianza del 95% para la diferencia de medias es:

$$(-3.4672, 13.4672)$$

Ejercicio 6: Contraste de Hipótesis

Pregunta: Realiza un contraste de hipótesis para el problema planteado en el ejercicio anterior y proporciona un $p - valor$.

Solución: Para realizar un contraste de hipótesis sobre la diferencia de medias de dos muestras independientes y calcular el valor p, seguimos estos pasos:

Paso 1: Plantear las hipótesis

Queremos comparar los rendimientos medios de los dos programas de entrenamiento. Las hipótesis nulas y alternativas son:

- **Hipótesis nula (H_0)**: No hay diferencia en las medias de rendimiento entre los dos programas, es decir, $\mu_1 - \mu_2 = 0$.
- **Hipótesis alternativa (H_a)**: Hay una diferencia en las medias de rendimiento entre los dos programas, es decir, $\mu_1 - \mu_2 \neq 0$.

Paso 2: Calcular el estadístico de prueba

Utilizamos la fórmula del estadístico t para la diferencia de medias:

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Dado que bajo H_0, $\mu_1 - \mu_2 = 0$, el estadístico t se simplifica a:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Sustituyendo los valores:

- $\bar{X}_1 = 80$
- $\bar{X}_2 = 75$
- $s_1 = 10$
- $s_2 = 12$
- $n_1 = 15$
- $n_2 = 12$

$$t = \frac{80 - 75}{\sqrt{\frac{10^2}{15} + \frac{12^2}{12}}} \approx 1.157$$

Paso 3: Determinar los grados de libertad

Para el contraste de hipótesis, utilizamos la aproximación de Welch-Satterthwaite para los grados de libertad df:

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}} \approx 21.44$$

Redondeamos a 21 grados de libertad.

Paso 4: Calcular el valor p

Con $t = 1.157$ y $df = 21$, utilizamos una tabla t o una calculadora de distribuciones t para encontrar el valor p. Dado que es un contraste bilateral (dos colas), multiplicamos por 2 la probabilidad acumulada de la cola superior.

Usamos una calculadora de distribución t:

$$p - valor = 2 \times P(T > 1.157 \mid df = 21)$$

El valor p asociado con $t = 1.157$ y $df = 21$ se encuentra mediante tablas t o software estadístico. Aquí usaremos una aproximación:

$$p \approx 2 \times 0.128 \approx 0.256$$

Conclusión

El valor p es aproximadamente 0.256. Dado que este valor es mayor que el nivel de significancia comúnmente utilizado ($\alpha = 0.05$), no rechazamos la hipótesis nula.

No hay suficiente evidencia para afirmar que hay una diferencia significativa en las medias de rendimiento entre los dos programas de entrenamiento.

Ejercicio 7: Estimación en el muestreo

Pregunta: Mediante el empleo de `R`, genera 100 muestras, cada una de tamaño 15, de una distribución normal de media 120 y desviación estándar 8. Para cada muestra, construye un IC del 95% para la σ^2. ¿Cuántos de estos intervalos contienen el verdadero valor de 64 para σ^2? ¿Este resultado está de acuerdo con lo que esperabas? Debes comenzar por derivar el IC para la varianza poblacional σ^2. Para ello, recuerda que la cuasivarianza muestral

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

es un estimador insesgado para la varianza poblacional σ^2 y que, bajo condiciones de normalidad,

$$s^2 \sim \frac{\sigma^2 \chi^2_{n-1}}{n-1}.$$

Solución: Para calcular los intervalos de confianza para σ^2, necesitamos ajustar un poco la fórmula del intervalo de confianza para σ. La fórmula para un intervalo de confianza del 95% para σ^2 es:

$$\left(\frac{(n-1) \cdot s^2}{\chi^2_{\alpha/2, n-1}}, \frac{(n-1) \cdot s^2}{\chi^2_{1-\alpha/2, n-1}} \right)$$

donde s^2 es la varianza muestral, n es el tamaño de la muestra y $\chi^2_{\alpha/2,n-1}$ y $\chi^2_{1-\alpha/2,n-1}$ son los percentiles correspondientes de la distribución chi-cuadrado con $n-1$ grados de libertad.

```r
# Establecer la semilla para la reproducibilidad
set.seed(42)
```

```r
# Parámetros
n_samples <- 100
sample_size <- 15
true_var <- 64
confidence_level <- 0.95

# Función para calcular el intervalo de confianza para la varianza
calculate_ci_var <- function(sample_var, n) {
  lower <- (n - 1) * sample_var / qchisq((1 - confidence_level) / 2, df = n - 1)
  upper <- (n - 1) * sample_var / qchisq((1 + confidence_level) / 2, df = n - 1)
  return(c(lower, upper))
}

# Generar las muestras y calcular las varianzas muestrales
samples <- matrix(rnorm(n_samples * sample_size, mean = 120, sd = 8)
                  , ncol = sample_size)
sample_vars <- apply(samples, 1, var)

# Calcular los intervalos de confianza para la varianza
cis_var <- apply(samples, 1, function(x) calculate_ci_var(var(x), length(x)))

# Contar cuántos intervalos contienen el verdadero valor de la varianza
contained <- apply(cis_var, 2, function(x) true_var >= x[2] & true_var <= x[1])
contained_count <- sum(contained)

# Imprimir el resultado
print(paste("Número de intervalos que contienen el verdadero valor de la varianza:"
            , contained_count))
```

```
## [1] "Número de intervalos que contienen el verdadero valor de la varianza: 95"
```

Ejercicio 8: Muestreo

Problema: Una compañía de auditoría está llevando a cabo una revisión de los balances contables reportados por un banco. Quieren estimar la proporción de cuentas para las cuales existe una discrepancia entre lo reportado por el cliente y el banco. Se decide realizar un muestreo para obtener una estimación precisa de esta proporción. Se sabe que la proporción real de cuentas con discrepancias es desconocida, pero la compañía de auditoría quiere obtener una estimación con un nivel de confianza del 99%. Además,

desean que el margen de error máximo en la estimación sea de 0.01 unidades. ¿Cuántas cuentas deberán seleccionarse para satisfacer estos requisitos?

Solución:

Datos:

- Nivel de confianza ($1 - \alpha$): 99% (es decir, $\alpha = 0.01$)
- Margen de error (E): 0.01

Para calcular el tamaño muestral necesario, utilizaremos la fórmula para el tamaño de muestra en estimaciones de proporciones:

$$n = \frac{Z^2 \cdot p \cdot (1 - p)}{E^2}$$

Donde:

- n es el tamaño muestral necesario.
- Z es el valor crítico de la distribución normal estándar correspondiente al nivel de confianza ($\alpha/2$).
- p es la proporción estimada (desconocida en este caso).
- E es el margen de error deseado.

Dado que no tenemos una estimación inicial de p, usaremos $p = 0.5$ como valor más conservador, ya que maximiza el tamaño de la muestra para un margen de error dado.

Primero, encontramos el valor crítico Z para un nivel de confianza del 99%:

1. $\alpha/2 = 0.01/2 = 0.005$
2. Buscamos el valor correspondiente en la tabla de la distribución normal estándar, que es aproximadamente 2.576.

Luego, utilizamos esta información para calcular el tamaño muestral:

$$n = \frac{2.576^2 \cdot 0.5 \cdot (1 - 0.5)}{0.01^2} \approx 16589.44$$

Redondeando al próximo entero superior, obtenemos que el tamaño muestral necesario es $n = 16590$.

Por lo tanto, se necesitan seleccionar al menos 16590 cuentas para que, con un nivel de confianza del 99%, la proporción muestral esté a no más de 0.01 unidad de la proporción real.

Ejercicio 9: Contraste de Hipótesis

Problema: Supongamos que un jugador de fútbol afirma que su tasa de éxito en los tiros de penali es superior al 70%. Sin embargo, un analista deportivo cree que la tasa de éxito es menor que eso. Para probar su hipótesis, el analista revisa los datos de los últimos 50 tiros de penaltis del jugador y encuentra que lograron marcar en 30 de ellos. ¿Quién tiene razón?

Solución: El analista decide realizar un contraste de hipótesis al nivel de significancia del 5%.

Hipótesis:

- Hipótesis nula (H_0): La tasa de éxito en los tiros de penal es igual o superior al 70%.
- Hipótesis alternativa (H_1): La tasa de éxito en los tiros de penal es menor que 70%.

Nivel de Significancia: $\alpha = 0.05$

Datos: - Número de tiros de penal (n): 50 - Número de tiros de penal exitosos (x): 30

Paso 1: Formulación de Hipótesis

$$H_0 : p \geq 0.70$$
$$H_1 : p < 0.70$$

Donde p es la proporción de tiros de penal exitosos.

Paso 2: Estadístico de Contraste

Dado que estamos trabajando con una proporción muestral y el tamaño de la muestra es grande (n > 30), podemos utilizar un estadístico z para el contraste de hipótesis:

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 \cdot (1-p_0)}{n}}}$$

Donde $\hat{p}$ es la proporción muestral, p_0 es la proporción bajo la hipótesis nula y n es el tamaño de la muestra.

Paso 3: Regla de Decisión

Rechazamos la hipótesis nula si $z < -z_\alpha$, donde $-z_\alpha$ es el valor crítico correspondiente al nivel de significancia α en el lado izquierdo de la distribución normal estándar.

Paso 4: Cálculo del Estadístico de Contraste y Valor Crítico

Primero, calculamos la proporción muestral:

$$\hat{p} = \frac{x}{n} = \frac{30}{50} = 0.60$$

Luego, calculamos el estadístico de contraste:

$$z = \frac{0.60 - 0.70}{\sqrt{\frac{0.70 \cdot (1-0.70)}{50}}}$$

Utilizando una tabla de la distribución normal estándar o un software estadístico, encontramos que $-z_{0.05} \approx -1.645$.

Paso 5: Toma de Decisiones

Como $z = -2.53$ es menor que $-z_{\alpha}$, rechazamos la hipótesis nula.

Con un nivel de significancia del 5%, hay evidencia suficiente para concluir que la tasa de éxito en los tiros de penal del equipo de fútbol es menor que el 70%.

Ejercicio 10: Datos

Problema: Se desea investigar si la edad media de los clientes de un banco excede los 40 años. Se realiza un análisis estadístico utilizando datos de una campaña de marketing bancario. Se calcula un intervalo de confianza para la edad promedio de los clientes basado en una muestra aleatoria de 1000 clientes. Posteriormente, se lleva a cabo un contraste de hipótesis para determinar si la edad media es mayor que 40 años, utilizando un nivel de significancia del 5%. Emplea para resolver este ejercicio los datos de la base de datos `bank` del repositorio UCI.

Solución:

```
url <- "https://archive.ics.uci.edu/ml/machine-learning-databases/00222/bank.zip"
download.file(url, "bank.zip")
unzip("bank.zip", "bank-full.csv")
bank <- read.csv("bank-full.csv",sep=";")

# Verificar la estructura del conjunto de datos
str(bank)
```

```
## 'data.frame':    45211 obs. of  17 variables:
##  $ age       : int  58 44 33 47 33 35 28 42 58 43 ...
##  $ job       : chr  "management" "technician" "entrepreneur" "blue-collar" ...
##  $ marital   : chr  "married" "single" "married" "married" ...
```

```
##  $ education: chr  "tertiary" "secondary" "secondary" "unknown" ...
##  $ default  : chr  "no" "no" "no" "no" ...
##  $ balance  : int  2143 29 2 1506 1 231 447 2 121 593 ...
##  $ housing  : chr  "yes" "yes" "yes" "yes" ...
##  $ loan     : chr  "no" "no" "yes" "no" ...
##  $ contact  : chr  "unknown" "unknown" "unknown" "unknown" ...
##  $ day      : int  5 5 5 5 5 5 5 5 5 5 ...
##  $ month    : chr  "may" "may" "may" "may" ...
##  $ duration : int  261 151 76 92 198 139 217 380 50 55 ...
##  $ campaign : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ pdays    : int  -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 ...
##  $ previous : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ poutcome : chr  "unknown" "unknown" "unknown" "unknown" ...
##  $ y        : chr  "no" "no" "no" "no" ...
```

```r
# Definir tamaño de la muestra
sample_size <- 1000

# Seleccionar una muestra aleatoria de la edad de los clientes
sample <- bank$age[sample.int(nrow(bank), sample_size)]

# Calcular estadísticas muestrales
x_bar <- mean(sample)
s <- sd(sample)
n <- length(sample)

# Definir parámetros para el contraste de hipótesis
mu_0 <- 40
alpha <- 0.05

# Calcular el intervalo de confianza para la media
lower <- x_bar - qt(1 - alpha/2, df = n - 1) * (s / sqrt(n))
upper <- x_bar + qt(1 - alpha/2, df = n - 1) * (s / sqrt(n))

# Realizar el contraste de hipótesis
z <- (x_bar - mu_0) / (s / sqrt(n))
z_critical <- qnorm(1 - alpha)

# Mostrar los resultados
cat("Intervalo de confianza para la edad media de
    los clientes del banco:",
```

```
    lower, "-", upper, "\n")
```

```
## Intervalo de confianza para la edad media de
##     los clientes del banco: 40.09765 - 41.42435
```

```
if (z > z_critical) {
  cat("Rechazamos la hipótesis nula. Hay evidencia para sugerir que la
      edad media es mayor que 40 años.\n")
} else {
  cat("No rechazamos la hipótesis nula. No hay suficiente evidencia
      para sugerir que la edad media es mayor que 40 años.\n")
}
```

```
## Rechazamos la hipótesis nula. Hay evidencia para sugerir que la
##       edad media es mayor que 40 años.
```