Prueba de Hipótesis: Estudio antes de un Examen

Julio César Velázquez Corona

Jose Manuel Herrera Sanchez

Jorge Emiliano Ávila Correa

3 de noviembre del 2025

Índice

1. Introducción

Los examenes son herramientas creadas para probar el conocimiento de una persona respecto a algo y como estudiantes, esa herramienta ha sido gran parte de nuestra vida. Cuando queremos hacer algo bien, solemos buscar la forma más óptima de hacerlo y en el caso de los examenes, lo más común es preguntarse si comer algo antes del examen o no, cuánto tiempo debería estudiar, cuánto debo dormir, etc.

En este proyecto, realizamos una estimación al número de horas promedio que un estudiante estudia antes de un examen, con el fin de obtener una referencia plausible y posiblemente aplicarla a nuestra carrera y futuro. Para llevar a cabo lo anterior, se realizará una prueba de hipótesis z cuando $n > 40$ y σ es desconocida, conocer σ implicaría reducir la dimensión del estudio cambiando el enfoque a una población con más datos conocidos, sin embargo, esto no se llevará a cabo en este momento.

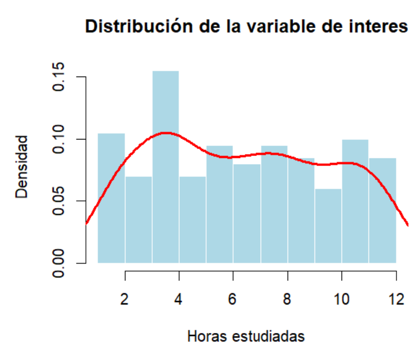
2. Datos

Los datos obtenidos para este proyecto provienen del dataset llamado *Student Academic Trends*, el cual contiene información de 200 estudiantes sobre sus hábitos de estudio, duración de sueño, cumplimiento de asistencia, desempeño académico, etc. Nuestra variable de interes es `hours_studied`. A continuación se presentará la estadística descriptiva:

```
mean(data$hours_studied)     # 6.3255
median(data$hours_studied)   # 6.15
sd(data$hours_studied)       # 3.2273
min(data$hours_studied)      # 1
max(data$hours_studied)      # 12
```

Podemos ver que $\bar{x}$ es de 6.3255 y que s es de 3.2273, lo que indica que con confianza se puede decir que los estudiantes muestrales estudian en promedio alrededor de 6 horas. En términos visuales, lo que más nos interesa respecto a nuestra variable de interes es conocer la densidad de cada dato. Además, es posible mostrar la distribución a la que corresponden nuestros datos, el histograma se muestra a continuación:

```
hist(data$hours_studied,
    main = 'Distribución de la variable de
    ↪  interes',
    xlab = 'Horas estudiadas',
    ylab = 'Densidad',
    col = 'lightblue',
    border = 'white',
    freq = FALSE)
lines(density(data$hours_studied), col =
↪  'red', lwd = 3)
```



Podemos ver un pico en cantidad de datos entre las 4 y 5 horas, lo cual puede indicar una tendencia a una reducción de tiempo respecto a la media muestral.

A pesar de que nuestros datos no son normales, el Teorema de Límite Central es invocado ya que $n > 40$, tampoco conviene normalizarlos manualmente porque no se cuentan con valores altamente asimétricos o con valores atípicos problemáticos.

3. Inferencia Estadística

Nuestra $\bar{x}$ es igual a $6.3255 \approx 6.33$, sin embargo, existe la probabilidad de que nuestra muestra contenga a un grupo de estudiantes bastante estudiosos, ¿qué sucede cuando por causas externas, una persona debe de estudiar menos?

Como equipo, nuestro tiempo de estudio promedio antes de un examen es de 5.33, ¿será posible que la media de tiempo real de estudio sea equivalente a la nuestra? Lo anterior sugiere que 5.33 sea el valor de μ_0. Al presentarse evidencia en contra de H_0, se determina que las horas de estudio reales son mayores a 5.33. Se utilizará el nivel de significancia estandar $\alpha = 0.05$, el desarrollo se plantea a continuación:

Hipótesis nula $H_0 : \mu = 5.33$

Estadístico de prueba: $z = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$

Hipótesis alternativa $H_a : \mu > 5.33$

$$\bar{x} = 6.33, \quad n = 200$$

$$s = 3.2273, \quad \mu_0 = 5.33$$

$$z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{6.33 - 5.33}{3.2273/\sqrt{200}} = 4.3820$$

$$p - value = 1 - \Phi(z) = 1 - \Phi(4.3820) = 5.87 \cdot 10^{-6} \approx 0$$

Concluimos que $p - value < 0.05$ y por lo tanto se rechaza la hipótesis nula, concluimos que las horas de estudio poblacionales son mayores a 5.33 a un nivel de significancia de 0.05. Lo que concluye que estudiamos menos que la población.

Para reforzar nuestra hipótesis, se construirá un intervalo de confianza bilateral para la media poblacional. Dado que los resultados muestran una fuerte evidencia en contra de H_0, es apropiado emplear un nivel de confianza más estricto ($\alpha = 0.01$), ya que el valor de μ_0 quedará excluido incluso utilizando un nivel de confianza estándar.

$$\bar{x} = 6.33, \quad n = 200, \quad s = 3.2273, \quad \alpha = 0.01$$

$$IC = \bar{x} \pm z_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

$$IC = 6.33 \pm z_{0.005} \cdot \frac{3.2273}{\sqrt{200}}$$

$$IC = 6.33 \pm 2.5758 \cdot 0.2282$$

$$IC = 6.33 \pm 0.5877 = (5.7421, 6.9178)$$

Con una nivel de confianza del 99 %, podemos darnos cuenta que de hecho, debemos rechazar la hipótesis nula y que el valor mínimo posible del tiempo de estudio en horas real es de $5.7421 \approx 5.74$.

Aunque hayamos obtenido un $p - value$ bastante pequeño, aún existe la posibilidad de que nuestro estudio no refleje un resultado 100 % preciso, por lo que se propone realizar el cálculo de la potencia, la cual nos indicará la sensibilidad de nuestra prueba.

Para eso, necesitamos una metrica importante: la probabilidad de no detectar que las horas de estudio poblacionales son mayores a 5.33, en otras palabras, la probabilidad de cometer un error tipo II (β). Se utilizará el mismo nivel de confianza utilizado en la prueba de hipótesis ($\alpha = 0.05$).

$$\alpha = 0.05, \quad \mu_0 = 5.33, \quad \mu' = 6.33, \quad s = 3.2273, \quad n = 200$$

$$\beta(\mu') = \Phi\left(z_\alpha + \frac{\mu_0 - \mu'}{s/\sqrt{n}}\right)$$

$$\beta(6.33) = \Phi\left(z_{0.05} + \frac{5.33 - 6.33}{3.2273/\sqrt{200}}\right)$$

$$\beta(6.33) = \Phi\left(1.6448 + \frac{5.33 - 6.33}{3.2273/\sqrt{200}}\right)$$

$$\beta(6.33) = \Phi(-2.7371)$$

$$\beta(6.33) = 0.0030 \to 100 \cdot (0.0030) = 0.30\,\%$$

Concluimos que la probabilidad de no rechazar la hipótesis nula cuando es falsa es de $0.30\,\%$. La potencia se obtiene facilmente: $1 - \beta(6.33) = 0.9969$, eso quiere decir que la probabilidad de detectar correctamente que la media real es mayor que 5.33 es $99.69\,\%$.

4. Conclusiones

Todo lo calculado indica que hay evidencia muy firme de que la probabilidad de horas de estudio promedio reales son mayores a las nuestras. Se concluye que como referencia, para el mayor desempeño posible, estudiemos un mínimo de 6 horas antes de un examen.

Es importante considerar que no se cuenta con mucha información respecto a nuestra muestra y eso nos puede generar bastantes incógnitas. Por mencionar algunas, no se sabe de qué institución se obtuvieron los datos, así que no es posible comprobar si los estudiantes son realmente estudiosos o no. Además, no se sabe de qué país provienen, asi que pueden haber influencias económicas y geográficas desconocidas que puedan influir con el tiempo de estudio. Si se deseara llevar a cabo un análisis más extensivo, se necesitarían más datos.

5. Referencias

S. F. Eslava. *Apuntes de Estadística Inferencial.* Universidad Anáhuac México Campus Norte, 2025.

Yaseen. S. A. *Analyzing Student Academic Trends.* Kaggle.
https://www.kaggle.com/datasets/saadaliyaseen/analyzing-student-academic-trends

6. Apéndice

```
# Codigo en R del proyecto

## Lectura de datos
setwd("C:\\Personal\\semestre13\\proyectoEI1\\latex-homework-template-master\\data")
data <- read.csv("student_exam_scores.csv", header=TRUE)
View(data)
str(data$hours_studied)
```

```
## Estadistica descriptiva
summary(data)
length(data$hours_studied)
mean(data$hours_studied)     # 6.3255
median(data$hours_studied)   # 6.15
sd(data$hours_studied)       # 3.2273
min(data$hours_studied)      # 1
max(data$hours_studied)      # 12

## Analisis de distribucion
hist(data$hours_studied,
     main = 'Distribución de la variable de interes',
     xlab = 'Horas estudiadas',
     ylab = 'Densidad',
     col = 'lightblue',
     border = 'white',
     freq = FALSE)
lines(density(data$hours_studied), col = 'red', lwd = 3)

## Prueba de hipotesis
xbar <- 6.33
n <- 200
s <- 3.2273
mu_0 <- 5.33
z <- (xbar - mu_0) / (s / sqrt(n))
z
pvalue <- 1 - pnorm(z)
pvalue

## Calculo del intervalo de confianza
z <- abs(qnorm(0.005))
z
var <- z * (3.2273/sqrt(200))
(IC_inf <- xbar - var)
(IC_sup <- xbar + var)

## Calculo del error tipo II
alpha <- 0.05
mu_0 <- 5.33
mu_indice <- 6.33
s <- 3.2273
n <- 200
(z_alpha <- qnorm(1 - alpha))
(betaCalc1 <- z_alpha + (mu_0 - mu_indice) / (s / sqrt(n)))
(betaCalc2 <- pnorm(betaCalc1))
(betaCalc3 <- 100 * betaCalc2)
(potencia = 1 - betaCalc2)
```