Práctica 1: Modelado Estadístico de Datos

Alex Herrerías Ramírez

Noviembre 2026

Ejercicio 1

Se pide proporcionar un código en R que ilustre la diferencia entre clínicamente relevante y estadísticamente significativo.

Para ver esta diferencia, simularemos dos escenarios diferentes, en uno usaremos una prueba en 10000 personas de un medicamento que baja la presión arterial en 0.4mmHg, y en el otro solamente 10 personas pero la presión alterial cambiara 5mmHg

Código en R

```r
# Fijamos la semilla para obtener siempre los mismos resultados
set.seed(123)

# Creamos un tamaño muestral grande para el primer escenario
n1 <- 10000
# Generamos los datos del grupo control con media 120 y sd 10
control_1 <- rnorm(n1, mean = 120, sd = 10)
# Generamos los datos del grupo tratamiento con una media
   ligeramente mayor
tratamiento_1 <- rnorm(n1, mean = 120.4, sd = 10)
# Realizamos los t-test entre el tratamiento y el control
test_1 <- t.test(tratamiento_1, control_1)

# Mostramos la diferencia de medias estimada
cat("Dif. de medias:", round(diff(test_1$estimate)*-1, 4), "\n")
# Resultado: 0.3327

# Mostramos el p-valor del test
cat("P-valor:", format.pval(test_1$p.value, eps=0.001), "\n")
# Resultado: 0.018684

# Creamos un tamaño muestral pequeño para el segundo escenario
n2 <- 10
# Generamos los datos del grupo control
control_2 <- rnorm(n2, mean = 120, sd = 10)
# Generamos los datos del grupo tratamiento con una diferencia clí
   nica notable
tratamiento_2 <- rnorm(n2, mean = 125, sd = 10)
# Realizamos el t-test entre ambos grupos
test_2 <- t.test(tratamiento_2, control_2)

# Mostramos la diferencia de medias
cat("Dif. de medias:", round(diff(test_2$estimate)*-1, 4), "\n")
# Resultado: 6.34

# Mostramos el p-valor
cat("P-valor:", round(test_2$p.value, 4), "\n")
# Resultado: 0.1722
```

Interpretación de los resultados

- **Escenario 1: Estadísticamente significativo, pero clínicamente irrelevante.**
  - **Datos:** Con un tamaño muestral alto ($N = 10,000$), observamos una diferencia de medias de solo **0.3327 mmHg**.
  - **Análisis:** Esta diferencia es minima en la práctica médica. Sin embargo, debido al gran tamaño de la muestra, el error estándar es minúsculo, resultando en un **p-valor de 0.0187** ($< 0{,}05$).
  - **Conclusión:** El resultado es estadísticamente significativo, pero carece de interés clínico. Esto demuestra que con suficiente muestra, cualquier diferencia distinta de cero será detectada como significativa.
- **Escenario 2: Clínicamente relevante, pero no estadísticamente.**
  - **Datos:** Con un tamaño muestral muy reducido ($N = 10$), observamos una diferencia de medias de **6.34 mmHg**.
  - **Análisis:** Esta reducción de presión es clínicamente muy importante. No obstante, la varianza y el bajo N provocan que el test no tenga potencia suficiente, arrojando un **p-valor de 0.1722** ($> 0{,}05$).
  - **Conclusión:** No se alcanza la significación estadística a pesar de que el tratamiento parece tener un gran efecto real.

Uso de IA en este ejercicio

He usado la IA para formatear de forma correctamente visual la interpretación de resultados. El prompt que le pasé es:

> "Tras ejecutar el código de R {insertar código R}, he obtenido los resultados {adjuntar resultados}, quiero realizar una interpretación de resultados que sea visual en R."

IA utilizada: Gemini 3 Pro

Ejercicio 2

La varianza de la transformación arcoseno raíz de una proporción sólo depende del tamaño muestral.

a) **Verdadero.**

b) Falso.

Justificación

La afirmación es **VERDADERA**.

Para justificarlo, usamos el **Método Delta** para poder aproximar la varianza de una transformación de una variable aleatoria.

Sea $\hat{p}$ la proporción muestral estimada a partir de una distribución Binomial $B(n, p)$. Sabemos que los momentos de $\hat{p}$ son:

$$E[\hat{p}] = p \quad \text{y} \quad Var(\hat{p}) = \frac{p(1-p)}{n}$$

Buscamos una función de transformación $g(\hat{p})$ tal que su varianza sea constante (independiente del parámetro p). La transformación propuesta es el arcoseno de la raíz cuadrada:

$$g(\hat{p}) = \arcsin(\sqrt{\hat{p}})$$

Aplicando la aproximación de primer orden del Método Delta, la varianza de la variable transformada es:

$$Var(g(\hat{p})) \approx [g'(p)]^2 \cdot Var(\hat{p})$$

Paso 1: Calcular la derivada $g'(p)$

Aplicamos la regla de la cadena a $g(p) = \arcsin(\sqrt{p})$:

$$\frac{d}{dp} \arcsin(u) = \frac{1}{\sqrt{1-u^2}} \cdot u'$$

Donde $u = \sqrt{p} \implies u' = \frac{1}{2\sqrt{p}}$.

Sustituyendo:

$$g'(p) = \frac{1}{\sqrt{1-(\sqrt{p})^2}} \cdot \frac{1}{2\sqrt{p}} = \frac{1}{\sqrt{1-p}} \cdot \frac{1}{2\sqrt{p}}$$

$$g'(p) = \frac{1}{2\sqrt{p(1-p)}}$$

Paso 2: Calcular la varianza transformada

Sustituimos la derivada y la varianza original en la fórmula del Método Delta:

$$Var(g(\hat{p})) \approx \left(\frac{1}{2\sqrt{p(1-p)}}\right)^2 \cdot \frac{p(1-p)}{n}$$

Elevamos al cuadrado el término de la derivada:

$$Var(g(\hat{p})) \approx \frac{1}{4p(1-p)} \cdot \frac{p(1-p)}{n}$$

Simplificamos los términos $p(1-p)$ del numerador y denominador:

$$Var(g(\hat{p})) \approx \frac{1}{4n}$$

Conclusión: Como se observa en el resultado final, la varianza asintótica de la transformación arcoseno raíz es $\frac{1}{4n}$. Este valor depende exclusivamente del tamaño muestral n y no de la proporción poblacional p, logrando así la estabilización de la varianza.

Uso de IA en este ejercicio

En este ejercicio he decidido usar la IA como guía para desarrollar la demostración matemática paso a paso y poder usar las librerías necesarias para que vea bien visualmente en LaTeX.

- **Prompt utilizado:** "Ayudame paso a paso a demostrar matemáticamente usando el Método Delta que la varianza de la transformación arcoseno raíz de una proporción es independiente de p y dame las librerias y los comandos para que se vea correctamente en LaTeX".
- **Integración:** He verificado con los materiales de la asignatura que la respuesta proporcionada por la IA es correcta y he copiado la formulas matematicas en el documento.

Ejercicio 3

En el modelo binormal, la curva ROC nunca puede estar por debajo de la diagonal principal.

La afirmación es **FALSA**.

El modelo binormal asume que la variable de decisión sigue una distribución normal tanto en la población sana ($X \sim N(\mu_0, \sigma_0^2)$) como en la enferma ($Y \sim N(\mu_1, \sigma_1^2)$). La forma de la curva ROC en este modelo viene dada por la ecuación:

$$ROC(t) = \Phi(a + b \cdot \Phi^{-1}(t))$$

Donde Φ es la función de distribución acumulada de la normal estándar, y los parámetros son:

$$a = \frac{\mu_1 - \mu_0}{\sigma_1}, \quad b = \frac{\sigma_0}{\sigma_1}$$

La curva ROC binormal es convexa (y por tanto siempre por encima de la diagonal, asumiendo $\mu_1 > \mu_0$) **únicamente** cuando las varianzas son iguales ($\sigma_0 = \sigma_1$, es decir, $b = 1$).

Sin embargo, cuando las varianzas son diferentes ($b \neq 1$), ocurren los siguientes fenómenos que contradicen la afirmación:

1. **Curvas ROC impropias ("Hooks"):** Si $b \neq 1$, la curva ROC cruza la diagonal principal en algún punto. Aunque el área bajo la curva (AUC) sea alta, la curva puede presentar un "gancho" o *hook* en los extremos, cruzando y situándose por debajo de la línea de no discriminación (diagonal principal) en regiones de alta o baja especificidad. Esto implica que, para ciertos puntos de corte, el test podría rendir peor que el azar.
2. **Inversión de poblaciones:** Si la media de la población enferma fuera menor que la de la sana ($\mu_1 < \mu_0$) y no ajustamos la dirección del test, la curva ROC completa estaría por debajo de la diagonal principal (AUC $< 0{,}5$), indicando que el test predice sistemáticamente lo contrario a la realidad.

Por tanto, afirmar que "nunca" puede estar por debajo es incorrecto, ya que depende estrictamente de la relación entre las varianzas y las medias de ambas distribuciones.

Uso de IA en este ejercicio

He utilizado la IA para estructurar la justificación teórica basada en las propiedades matemáticas del modelo binormal.

- **Prompt utilizado:** "Explicame por qué en el modelo binormal la curva ROC puede cruzar la diagonal principal si las varianzas son distintas (improper ROC curves)".
- **Integración:** La IA me ha proporcionado las formulas para el desarrollo del ejercicio y me ha explicado el fenómeno de los "hooks". He revisado la respuesta con los materiales de la asignatura y he redactado la respuesta final.

Ejercicio 4

A partir del siguiente código en R se pide completarlo para: a) Proporcionar un modelo final. b) Realizar un análisis de residuos de dicho modelo final.

a) Estrategia de Modelización y Modelo Final

En el enunciado se cargan las librerías `faraway`, `MASS` y `car`, y se cargan los datos `mtcars`. Como objetivo tenemos que predecir el consumo (`mpg`).

Dado que tenemos muchas variables predictoras (10) para un tamaño muestral pequeño ($n = 32$), es muy probable que exista multicolinealidad. La estrategia a seguir será:

1. Ajustar el modelo completo con todas las variables.
2. Evaluar la multicolinealidad inicial (VIF).
3. Utilizar un procedimiento de selección de variables *Stepwise* (basado en el criterio AIC) para eliminar variables redundantes y obtener un modelo parsimonioso.

Código en R para la selección del modelo:

```r
# Limpieza y carga de librerías
rm(list=ls())
library(faraway)
library(MASS)
library(car)
data(mtcars)

modelo_completo <- lm(mpg ~ cyl + disp + hp + drat + wt + qsec + vs
    + am + gear + carb, data = mtcars)

# Stepwise basado en AIC (ambas direcciones)
modelo_final <- stepAIC(modelo_completo, direction = "both", trace
   = 0)
summary(modelo_final)
```

Interpretación del Modelo Final:

El algoritmo `stepAIC` ha seleccionado un modelo con tres predictores: **wt** (peso), **qsec** (tiempo en 1/4 milla) y **am** (transmisión).

A partir de la salida proporcionada (`summary`), la ecuación del modelo ajustado es:

$$\widehat{mpg} = 9{,}618 - 3{,}917 \cdot wt + 1{,}226 \cdot qsec + 2{,}936 \cdot am$$

Análisis de los resultados:

- **Significatividad individual:** Las variables `wt` ($p < 0{,}001$) y `qsec` ($p < 0{,}001$) son altamente significativas. La variable `am` es significativa al nivel del 5 % ($p = 0{,}0467$). El intercepto no resulta estadísticamente significativo ($p > 0{,}05$), pero se mantiene por coherencia del modelo.

- **Interpretación de coeficientes:**
  - El peso (`wt`) tiene un efecto negativo: por cada 1000 lbs adicionales, el rendimiento baja en $\approx 3{,}9$ millas/galón.
  - La transmisión manual (`am`=1) incrementa el rendimiento en $\approx 2{,}9$ millas/galón frente a la automática, manteniendo el resto constante.
- **Bondad de ajuste:** El modelo presenta un **R-cuadrado ajustado de 0.8336**. Esto indica que el modelo es capaz de explicar el **83.36 %** de la variabilidad observada en el consumo de combustible, lo cual es un ajuste muy alto considerando la simplicidad del modelo (solo 3 variables).
- **Significatividad global:** El estadístico F (52,75) con un p-valor extremadamente bajo ($1{,}21 \times 10^{-11}$) confirma que el modelo en su conjunto es válido y superior al modelo nulo.

b) Análisis de Residuos

Para validar el modelo, debemos comprobar las hipótesis del modelo de regresión lineal: Linealidad, Homocedasticidad, Normalidad e Independencia.

Código en R para el diagnóstico:

```r
# Analisis grafico
par(mfrow = c(2, 2))
plot(modelo_final)
par(mfrow = c(1, 1))

# Analisis de multicolinealidad
print(vif(modelo_final))

# Test de normalidad de residuos (Shapiro-Wilk)
shapiro_test <- shapiro.test(residuals(modelo_final))
print(shapiro_test)

# Test de homocedasticidad (Breusch-Pagan, requiere library car)
print(ncvTest(modelo_final))
```

Interpretación Numérica y Gráfica:

Para validar el modelo, hemos realizado tanto una inspección visual de los gráficos (Figura 1) como los contrastes de hipótesis pertinentes:

1. **Linealidad (Residuals vs Fitted):** En el primer gráfico, la línea roja de suavizado muestra una ligera curvatura, pero se mantiene relativamente cercana a la línea horizontal de cero. Aunque hay residuos algo altos etiquetados (Chrysler Imperial, Fiat 128), no se observa una estructura sistemática grave (como una parábola perfecta) que invalide el modelo lineal.

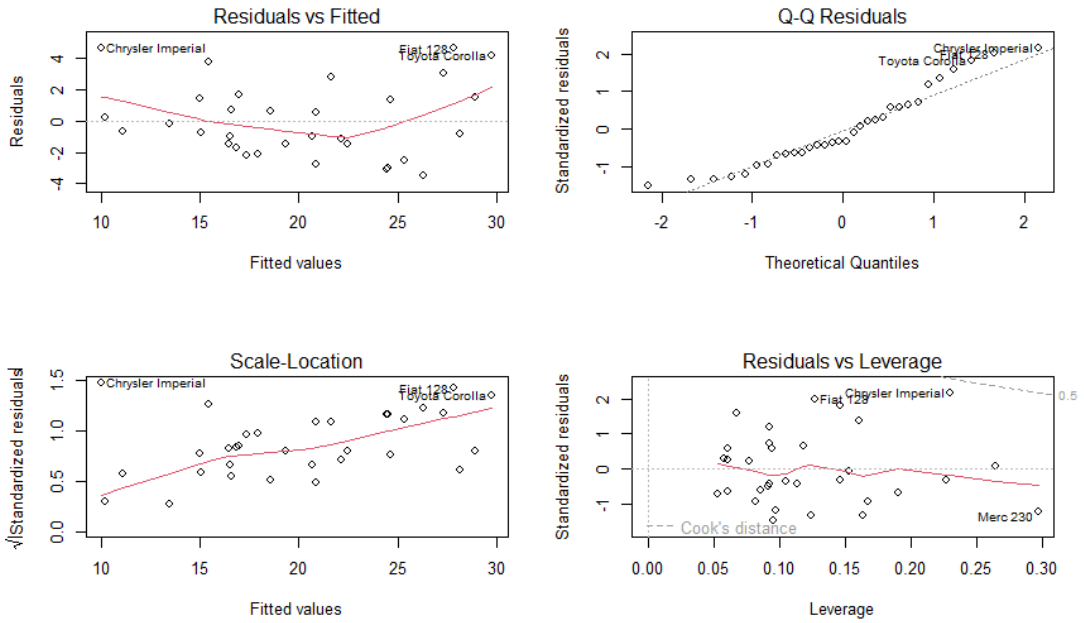

Figura 1: Gráficos de diagnóstico del modelo final (`mpg ~ wt + qsec + am`).

2. **Normalidad (Q-Q y Shapiro-Wilk):** Los residuos siguen la diagonal salvo ligeras desviaciones (Chrysler Imperial, Fiat 128). El test de Shapiro-Wilk ($W = 0{,}9411$, $p = 0{,}0804 > 0{,}05$) confirma la asunción de normalidad.
3. **Homocedasticidad (Scale-Location y ncvTest):** A pesar de una leve pendiente visual, el test de Breusch-Pagan ($p = 0{,}2119 > 0{,}05$) confirma estadísticamente la varianza constante de los errores.
4. **Diagnóstico Estructural:**
   - **Influyentes:** El "Merc 230"tiene alto apalancamiento, pero ningún punto supera la Distancia de Cook crítica, descartando observaciones influyentes.
   - **Multicolinealidad:** Los VIF son bajos (`wt`: 2.48, `qsec`: 1.36, `am`: 2.54), confirmando ausencia de redundancia entre predictores.

Conclusión: El modelo es estadísticamente válido al cumplir las hipótesis de normalidad y homocedasticidad, y no presentar problemas de colinealidad.

Uso de IA en este ejercicio

He usado la IA para generar el código de selección y la estructura de interpretación.

- **Prompt:** "Completa el código con stepAIC, grafica residuos e interpreta Shapiro y ncvTest con los valores reales".
- **Integración:** Ejecuté el código en R, validé los p-valores (0,08 y 0,21) y modifique la respuesta generada por la IA para poder realizar un análisis de residuos con graficos.

Ejercicio 5

Se pide proporcionar un código en R que ilustre que el procedimiento stepwise en regresión lineal puede seleccionar variables irrelevantes.

Código en R: Simulación de Ruido Puro

Para demostrar este fenómeno, generaremos un conjunto de datos donde la variable respuesta Y y las variables predictoras $X_1, \ldots, X_{50}$ son generadas de forma totalmente aleatoria e independiente. No existe ninguna relación real entre ellas. Posteriormente, ejecutaremos una selección *stepwise*.

```r
set.seed(42) # Semilla para reproducibilidad

# 1. Generación de datos puramente aleatorios (Ruido)
n <- 100    # Tamaño muestral
p <- 50     # Número de variables predictoras irrelevantes

# Generamos una matriz de predictores X (ruido normal)
X <- matrix(rnorm(n * p), nrow = n, ncol = p)
colnames(X) <- paste0("Var", 1:p)

# Generamos la variable respuesta Y (ruido normal, independiente de X)
# En la realidad, ningún predictor debería ser seleccionado.
y <- rnorm(n)

datos <- data.frame(y, X)

# 2. Ajuste del modelo inicial
# Empezamos con un modelo nulo (solo el intercepto)
modelo_nulo <- lm(y ~ 1, data = datos)
# Definimos el modelo completo (con todas las variables)
modelo_maximo <- lm(y ~ ., data = datos)

# 3. Aplicación del método Stepwise (Forward Selection)
# Le pedimos que busque variables que "mejoren" el modelo según AIC
modelo_stepwise <- step(modelo_nulo,
                        scope = list(lower = modelo_nulo, upper = modelo_maximo),
                        direction = "forward",
                        trace = 0) # trace=0 para no llenar la consola

# 4. Resultados
cat("Número de variables seleccionadas siendo todas irrelevantes:",
    length(coef(modelo_stepwise)) - 1, "\n")
print(summary(modelo_stepwise))
```

Interpretación de los resultados

Al ejecutar el código anterior, nos encontramos con un resultado alarmante que ilustra perfectamente el peligro de la minería de datos ciega (*data dredging*):

1. Hemos construido los datos de tal manera que la hipótesis nula es cierta para todas las variables ($\beta_i = 0$). Ninguna variable `VarX` tiene capacidad predictiva real sobre Y.
2. A pesar de que los datos son ruido, el algoritmo ha seleccionado un total de **21 variables** como "importantes".
   - Variables como `Var39` y `Var29` presentan una significancia estadística altísima ($p < 0{,}001$), lo cual nos indica que es un falso positivo.
   - El modelo final obtiene un R^2 **de 0.5234**. Podríamos interpretar que el modelo explica más del 52 % de la variabilidad de los datos, cuando en realidad la capacidad real es 0 %.
   - El p-valor global del modelo es $2{,}839 \times 10^{-6}$, lo que a primera vista nos haría pensar que el modelo es extremadamente sólido.
3. Esto ocurre por el problema de las comparaciones múltiples y la alta dimensionalidad ($p = 50$) respecto al tamaño muestral ($n = 100$). La probabilidad de encontrar correlaciones espurias por puro azar es inmensa. El algoritmo *stepwise* capitaliza este azar, reteniendo cualquier variable que reduzca el AIC, creando un modelo que describe perfectamente el "ruido" de la muestra concreta pero que fallará al intentar predecir nuevos datos.

Uso de IA en este ejercicio

He utilizado la IA para generar el escenario de simulación en R.

- **Prompt utilizado:** "Como genero un código en R que cree un dataset de ruido (variables aleatorias independientes de Y) y usa step forward para demostrar que selecciona variables irrelevantes.".
- **Integración:** He ajustado el código generado con el número de predictores a 50 para asegurar que el fenómeno se produzca visualmente, y he interpretrado los resultados para asegurarme de que se cumple lo que pide el ejercicio.

Ejercicio 6

¿Cuáles son las características fundamentales de los Ensayos Clínicos? (Se pide dar al menos 4 características).

Un Ensayo Clínico es una evaluación experimental de un producto, sustancia, medicamento, técnica diagnóstica o terapéutica que, a través de su aplicación a seres humanos, pretende valorar su eficacia y seguridad. Para que sus resultados sean válidos y extrapolables (validez interna y externa), deben cumplir cuatro características fundamentales:

1. Carácter Prospectivo

El ensayo clínico se planifica *a priori*. Antes de reclutar al primer paciente, se debe redactar un protocolo detallado que especifique los objetivos, los criterios de inclusión y exclusión, la metodología estadística, el tamaño muestral necesario y las variables a medir.

- A diferencia de los estudios observacionales retrospectivos (donde se analizan datos ya existentes), en el ensayo clínico los datos se generan específicamente para responder a la pregunta de investigación.
- Esto permite controlar la calidad de los datos y asegurar que la secuencia temporal es correcta (la intervención precede al resultado), lo cual es indispensable para establecer relaciones de causalidad.

2. Presencia de Grupo Control (Comparabilidad)

Para afirmar que un nuevo tratamiento es eficaz, se debe comparar contra algo. No basta con observar mejoría en los pacientes tratados (efecto Hawthorne o evolución natural de la enfermedad), es necesario demostrar que mejoran *más* que un grupo similar que no recibe dicho tratamiento.

- El **grupo control** recibe el tratamiento estándar actual (gold standard) o un placebo (si no existe tratamiento eficaz y es éticamente aceptable).
- El objetivo es que la única diferencia entre el grupo experimental y el control sea la intervención estudiada, permitiendo atribuir las diferencias en los resultados exclusivamente al tratamiento.

3. Aleatorización (Randomización)

La asignación de los pacientes a los grupos de tratamiento (experimental o control) debe hacerse estrictamente al azar.

- La aleatorización es el único método que asegura que, en promedio, las variables basales (edad, sexo, gravedad de la enfermedad) y, crucialmente, las **variables de confusión desconocidas**, se distribuyan equitativamente entre ambos grupos.
- Esto evita el *sesgo de selección*, garantizando que los grupos sean comparables al inicio del estudio. Sin aleatorización, el médico podría (consciente o inconscientemente) asignar a los pacientes más graves al grupo control, invalidando los resultados.

4. Enmascaramiento (Cegamiento)

El enmascaramiento consiste en ocultar la asignación del tratamiento a los participantes para evitar sesgos de información y evaluación.

- **Simple ciego:** El paciente desconoce qué recibe (evita el efecto placebo psicológico).
- **Doble ciego:** Ni el paciente ni el médico/investigador conocen el tratamiento asignado (evita que el médico evalúe de forma distinta la mejoría según sus expectativas).
- **Triple ciego:** También se enmascara al analista de datos.
- Aunque no siempre es posible (ej. cirugía vs fármaco), es una característica deseable para mantener la objetividad en la medida de los resultados.

Uso de IA en este ejercicio

He utilizado la IA para estructurar y sintetizar las definiciones teóricas de las características clave.

- **Prompt utilizado:** "Enumera 4 características fundamentales de un ensayo clínico aleatorizado y dame referencias para que pueda investigarlas".
- **Integración:** La IA me propuso Aleatorización, Control, Cegamiento y Prospectivo. He revisado las definiciones proporcionadas para asegurar que las caracteristicas fueran las buscadas y he ajustado el formato a LaTeX.