Análisis de Pruebas de Hipótesis

Ciencia de Datos - Maestria IA

Juan Carlos Vega Rueda

Mayo 2025

Introducción

El conjunto de datos Smokers Health Data ofrece una base sólida para analizar los efectos del tabaquismo sobre diversas variables fisiológicas y clínicas en adultos. A través de este análisis, se busca identificar diferencias significativas entre personas fumadoras y no fumadoras, utilizando herramientas estadísticas inferenciales implementadas en R. Este enfoque permite respaldar decisiones en salud pública y medicina preventiva con evidencia cuantitativa.

Objetivos

- Aplicar pruebas de hipótesis para comparar medias y proporciones entre fumadores y no fumadores.
- Determinar si existen diferencias estadísticamente significativas en variables como colesterol y frecuencia cardíaca.
- Evaluar el impacto del tabaquismo sobre indicadores clínicos clave.
- Promover el uso de análisis estadístico como herramienta de apoyo en los análisis.

Definiciones

- **Prueba de hipótesis:** procedimiento estadístico para evaluar afirmaciones sobre parámetros poblacionales a partir de datos muestrales.
- **Hipótesis nula (H_0):** suposición inicial que se pone a prueba, generalmente indicando ausencia de efecto o diferencia.
- **Hipótesis alternativa (H_1):** plantea una diferencia o efecto significativo respecto a H_0.
- **p-valor*:** probabilidad de obtener un resultado igual o más extremo que el observado, bajo la suposición de que H_0 es verdadera.
- **Nivel de significancia (α):** umbral (comúnmente 0.05) para decidir si se rechaza H_0.
- **Prueba t de Welch:** compara medias de dos grupos con varianzas desiguales.
- **Prueba de proporciones:** compara proporciones entre dos grupos independientes.

Cargar paquetes y datos

Las siguientes son las librerias en *R* que serán utilizadas en el desarrollo de los análisis y el reporte de resultados.

```
library(readr)
library(dplyr)
library(ggplot2)
library(stringr)
library(tidyr)
library(tigerstats)
library(BSDA)
library(lattice)
library(knitr)
library(kableExtra)
```

Las librerias permiten el análisis de datos, el manejo de tablas, análisis estadísticos, análisis de Componentes principales y modelos de agrupamiento como la clusterización k-means.

Cuadro 1: Primeras 10 filas de los datos efectos del tabaquismo en la salud

age	sex	current_smoker	heart_rate	blood_pressure	cigs_per_day	chol
54	male	yes	95	110/72	NA	219
45	male	yes	64	121/72	NA	248
58	male	yes	81	127.5/76	NA	235
42	male	yes	90	122.5/80	NA	225
42	male	yes	62	119/80	NA	226
57	male	yes	62	107.5/72.5	NA	223
43	male	yes	75	109.5/69	NA	222
42	male	yes	66	123/73	NA	196
37	male	yes	65	123.5/77	NA	188
49	male	yes	93	127.5/81.5	NA	256

Prueba de hipótesis para una media (frecuencia cardíaca = 75 Latidos por minuto)

Estas son las hipótesis planteadas:

- Hipótesis nula (H_0): la media poblacional es igual a 75 (μ = 75).
- Hipótesis alternativa (H_1): la media poblacional es distinta de 75 ($\mu \neq 75$).

```
z.test(x = df_smoking$heart_rate,
       sigma.x = sd(df_smoking$heart_rate, na.rm = TRUE),
       mu = 75,
       alternative = "two.sided",
       conf.level = 0.95)
```

```
##
##  One-sample z-Test
##
## data:  df_smoking$heart_rate
## z = 3.5809, p-value = 0.0003423
## alternative hypothesis: true mean is not equal to 75
## 95 percent confidence interval:
##  75.31188 76.06607
## sample estimates:
## mean of x
##  75.68897
```

De acuerdo con lo anterior:

- Si el p-valor < 0.05, se rechaza la hipótesis nula: la frecuencia cardíaca promedio es diferente de 75.

- Si el p-valor ≥ 0.05, no hay evidencia suficiente para rechazar que la media sea 75.

Los resultados indican que con un 95% de confianza, la verdadera media poblacional está entre 75.31 y 76.07, y además en este caso p-value es 0.0003423, se rechaza la hipótesis nula, por lo que la frecuencia cardiaca es diferente de 75, de acuerdo con el resultado del test la media es **75.68897** latidos por minuto.

Prueba de hipótesis para una media (colesterol > 200 mg/dL)

Estas son las hipótesis planteadas:

- Hipótesis nula (H_0): El nivel medio de colesterol en la población es menor o igual a 200 mg/dL.
- Hipótesis alternativa (H_1): El nivel medio de colesterol en la población es mayor a 200 mg/dL.

```
z.test(x = df_smoking$chol,
       sigma.x = sd(df_smoking$chol, na.rm = TRUE),
       mu = 200,
       alternative = "greater",
       conf.level = 0.95)
```

```
##
##  One-sample z-Test
##
## data:  df_smoking$chol
## z = 51.456, p-value < 2.2e-16
## alternative hypothesis: true mean is greater than 200
## 95 percent confidence interval:
##  235.4261       NA
## sample estimates:
## mean of x
##  236.5959
```

De acuerdo con lo anterior:

- Si el p-valor < 0.05, se concluye que el colesterol medio es mayor a 200 mg/dL.
- Si el p-valor ≥ 0.05, no hay evidencia suficiente para afirmar que es mayor mg/dL.

Los resultados indican que con un 95% de confianza, el colesterol medio está por encima de 235.42 mg/dL, y además en este caso p-value es menor a 0.05, por lo que se rechaza la hipótesis nula.

Prueba de hipótesis para una proporción (colesterol alto > 20%)

Estas son las hipótesis planteadas:

- Hipótesis nula (H_0): La proporción de personas con colesterol alto es igual al 20%.
- Hipótesis alternativa (H_1): La proporción de personas con colesterol alto es mayor al 20%.

```
df_smoking$chol_alto <- ifelse(df_smoking$chol > 240, 1, 0)
prop.test(x = sum(df_smoking$chol_alto, na.rm = TRUE),
          n = sum(!is.na(df_smoking$chol_alto)),
          p = 0.20,
          alternative = "greater",
          conf.level = 0.95)
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  sum out of sumdf_smoking$chol_alto out of !is.na(df_smoking$chol_alto)TRUE out of
## X-squared = 1271.4, df = 1, p-value < 2.2e-16
## alternative hypothesis: true p is greater than 0.2
## 95 percent confidence interval:
##  0.4155977 1.0000000
## sample estimates:
##         p
## 0.4287182
```

De acuerdo con lo anterior:

- Si el p-valor < 0.05, la proporción de personas con colesterol alto es mayor al 20%.
- Si el p-valor ≥ 0.05, no hay evidencia suficiente para afirmarlo.

Los resultados indican que con un 95% de confianza, el porcentaje de personas con colesterol alto es mayor a 41.56%, y además en este caso p-value es menor a 0.05, por lo que se acepta la hipótesis nula, en donde la proporción de personas con colesterol alto es de 42.8%.

Prueba de hipótesis para una proporción (taquicardia ≠ 5%)

Estas son las hipótesis planteadas:

- Hipótesis nula (H_0): La proporción de personas con taquicardia es igual al 5%.
- Hipótesis alternativa (H_1): La proporción de personas con taquicardia es diferente del 5%.

```
df_smoking$taquicardia <- ifelse(df_smoking$heart_rate > 100, 1, 0)
prop.test(x = sum(df_smoking$taquicardia, na.rm = TRUE),
          n = sum(!is.na(df_smoking$taquicardia)),
          p = 0.05,
          alternative = "two.sided",
          conf.level = 0.95)
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  sum out of sumdf_smoking$taquicardia out of !is.na(df_smoking$taquicardia)TRUE out
## X-squared = 55.613, df = 1, p-value = 8.825e-14
## alternative hypothesis: true p is not equal to 0.05
## 95 percent confidence interval:
##  0.01939026 0.02926409
## sample estimates:
##          p
## 0.02384615
```

De acuerdo con lo anterior:

- Si el p-valor < 0.05, la proporción de taquicardia es diferente al 5%.
- Si el p-valor ≥ 0.05, no hay evidencia suficiente para afirmar una diferencia.

Los resultados indican que con un 95% de confianza, la proporción de personas con taquicardia está entre 1.93% y 2.92%, y además en este caso p-value es menor a 0.05, por lo que se acepta la hipótesis nula, en donde la proporción de personas con taticardia es diferente al 5%. En este caso es de 2.38%

Diferencia de medias (colesterol entre fumadores y no fumadores)

Estas son las hipótesis planteadas:

- Hipótesis nula (H_0): No hay diferencia en los niveles medios de colesterol entre fumadores y no fumadores.
- Hipótesis alternativa (H_1): Existe una diferencia significativa entre los niveles medios de colesterol.

```r
smokers <- df_smoking[df_smoking$current_smoker == "yes", ]
non_smokers <- df_smoking[df_smoking$current_smoker == "no", ]

t.test(smokers$chol, non_smokers$chol,
       alternative = "two.sided",
       var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  smokers$chol and non_smokers$chol
## t = -2.9119, df = 3884.8, p-value = 0.003612
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -6.925837 -1.352281
## sample estimates:
## mean of x mean of y
##  234.5067  238.6458
```

De acuerdo con lo anterior:

- Si el p-valor < 0.05, hay una diferencia significativa en los niveles de colesterol entre fumadores y no fumadores.

De acuerdo con los resultados, el p-valor = 0.0036, este valor es menor que 0.05, por lo tanto, se rechaza la hipótesis nula. **Hay evidencia estadísticamente significativa para afirmar que existe una diferencia en los niveles medios de colesterol entre fumadores y no fumadores.**

Diferencia de medias (frecuencia cardíaca: fumadores > no fumadores)

Estas son las hipótesis planteadas:

- Hipótesis nula (H_0): No hay diferencia en la frecuencia cardíaca promedio entre fumadores y no fumadores.
- Hipótesis alternativa (H_1): La frecuencia cardíaca promedio de los fumadores es mayor que la de los no fumadores.

```
t.test(smokers$heart_rate, non_smokers$heart_rate,
       alternative = "greater",
       var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  smokers$heart_rate and non_smokers$heart_rate
## t = 3.5809, df = 3896.4, p-value = 0.0001733
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.7434658       Inf
## sample estimates:
## mean of x mean of y
##  76.38302  75.00762
```

De acuerdo con lo anterior:

- Si el p-valor < 0.05, la frecuencia cardíaca de los fumadores es mayor que la de los no fumadores.

los resultados muestran que:

- p-valor = 0.00017:

Este valor es mucho menor que 0.05, por lo tanto, se rechaza la hipótesis nula. Hay evidencia estadísticamente significativa para afirmar que la frecuencia cardíaca promedio de los fumadores es mayor que la de los no fumadores.

- Intervalo de confianza 95%:

El intervalo $[0.74, \infty)$ indica que la diferencia promedio entre fumadores y no fumadores es al menos 0.74 latidos por minuto, a favor de los fumadores.

Diferencia de proporciones (colesterol alto entre fumadores y no fumadores)

Estas son las hipótesis planteadas:

- Hipótesis nula (H_0): La proporción de personas con colesterol alto es la misma en fumadores y no fumadores.
- Hipótesis alternativa (H_1): La proporción de personas con colesterol alto es diferente entre fumadores y no fumadores.

```
table_chol <- table(df_smoking$current_smoker, df_smoking$chol_alto)
prop.test(x = c(table_chol["yes", "1"], table_chol["no", "1"]),
          n = c(sum(table_chol["yes", ]), sum(table_chol["no", ])),
          alternative = "two.sided",
          conf.level = 0.95)
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c out of ctable_chol["yes", "1"] out of sum(table_chol["yes", ])table_chol["no", "
## X-squared = 5.4583, df = 1, p-value = 0.01948
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.069158628 -0.005995786
## sample estimates:
##    prop 1    prop 2
## 0.4097510 0.4473282
```

De acuerdo con lo anterior:

- Si el p-valor < 0.05, la proporción de colesterol alto difiere entre fumadores y no fumadores.

los resultados muestran que:

- p-valor = 0.0195:

Es menor que 0.05, por lo tanto, rechazamos la hipótesis nula. Hay evidencia estadísticamente significativa para afirmar que la proporción de personas con colesterol alto difiere entre fumadores y no fumadores.

- Intervalo de confianza 95%:

El intervalo [−0.0692, −0.0060] no incluye el 0, lo que confirma que la diferencia es significativa. Como ambos extremos son negativos, indica que la proporción de colesterol alto es menor en fumadores que en no fumadores.

Conclusiones

- Se encontró evidencia significativa de que los fumadores tienen una frecuencia cardíaca promedio mayor que los no fumadores.
- Los niveles medios de colesterol fueron ligeramente menores en fumadores, aunque la diferencia fue estadísticamente significativa.
- La proporción de colesterol alto también fue significativamente diferente entre los grupos, siendo menor en fumadores.
- Las pruebas de hipótesis permiten validar diferencias entre grupos con base en evidencia estadística, lo cual es esencial para la toma de decisiones en salud pública y medicina preventiva.