Preparación y Análisis de datos

Capítulo 2: Análisis inferencial de datos

Mg. J. Eduardo Gamboa U.

2025-03-27

Introducción

¿Qué es la inferencia estadística?

¿Por qué es importante?

Estimación

Estimación puntual

Sea X_1,...,X_n una muestra de tamaño n de una población con parámetro θ. Se denomina estimador puntual de θ a cualquier estadístico $\hat{\Theta} = h(X_1, ..., X_n)$ cuyo valor $\hat{\theta} = h(x_1, ..., x_n)$ dará una estimación puntual de θ. En este caso, Θ es una variable aleatoria y$\hat{\theta}$ es un número.
Estimador puntual de la media:

$$\hat{\mu} = \overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Estimador puntual de la variancia:

$$\hat{\sigma^2} = s^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n-1}$$

Estimador puntual de la proporción:

$$\hat{\pi} = p = \frac{\text{Número de éxitos}}{n}$$

Ejemplo

En una población de madres adolescentes, la talla (en metros) sigue una distribución con media μ y variancia σ^2. Se extrae una muestra aleatoria 8 madres adolescentes, obteniendo las siguientes medidas de talla: 1.50, 1.60, 1.58, 1.45, 1.52, 1.68, 1.62, 1.55. Halle un estimador puntual para la media, la varianza y la desviación estándar poblacionales.

```
talla = c(1.60, 1.58, 1.45, 1.52, 1.68, 1.62, 1.55)
talla |> mean()
```

```
[1] 1.571429
```

```
talla |> var()
```

```
[1] 0.005480952
```

```
talla |> sd()
```

```
[1] 0.07403345
```

Estimación por intervalos de confianza

Sea X_1,...,X_n una muestra aleatoria de tamaño n de una población con parámetro θ, cuyos valores observados o datos respectivos son x_1,...,x_n.
Sea $a = h_1(x_1, ..., x_n)$ y $b = h_2(x_1, ..., x_n)$, valores numéricos calculados a partir de los datos de la muestra.
Entonces se dice que el intervalo $[a, b]$ tiene un nivel de confianza del $(1 - \alpha) \times 100\%$ de contener el parámetro θ, o que $\theta \in [a, b]$ con un nivel de confianza del $(1 - \alpha) \times 100\%$.
Interpretación: Con un nivel de confianza del $(1 - \alpha) \times 100\%$, se estima que el parámetro θ está contenido en el intervalo $[a, b]$.

Intervalo de confianza para la media

Si $X_1,\ldots,X_n$ es una muestra aleatoria de una población normal con media μ y σ^2 conocida, el intervalo con un nivel de confianza del $(1-\alpha)\times 100\%$ para la media μ se obtiene mediante:

$$\left(\bar{X} - Z_{(1-\alpha/2)}\frac{s}{\sqrt{n}} \leq \mu \leq \bar{X} + Z_{(1-\alpha/2)}\frac{s}{\sqrt{n}}\right)$$

Si σ^2 es desconocida, el intervalo con un nivel de confianza del $(1-\alpha)\times 100\%$ para la media μ se obtiene mediante:

$$\left(\bar{X} - t_{(1-\alpha/2;n-1)}\frac{s}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{(1-\alpha/2;n-1)}\frac{s}{\sqrt{n}}\right)$$

Ejemplo

En un estudio sobre hábitos de hidratación, se registró el consumo diario de agua en litros de 16 personas seleccionadas aleatoriamente en una comunidad. La cantidad de agua ingerida es un factor clave para la salud y el bienestar, y este análisis busca conocer la variabilidad en los patrones de hidratación.

2.75	2.43	1.59	2.82
3.26	2.38	1.79	2.08
3.29	2.88	2.27	2.72
2.62	1.54	1.64	2.22

a. Suponiendo que σ es conocida, $\sigma = 0.5$

```
agua = c(2.75, 2.43, 1.59, 2.82, 3.26, 2.38, 1.79, 2.08, 3.29, 2.88,
         2.27, 2.72, 2.62, 1.54, 1.64, 2.22)

media = agua |> mean()
sigma = 0.5
n     = agua |> length()
conf  = 0.95
z     = qnorm((1+conf)/2)
LI    = media - z*sigma/sqrt(n)
LS    = media + z*sigma/sqrt(n)
c(LI,LS)
```

```
[1] 2.147505 2.637495
```

```r
library(BSDA)
zsum.test(mean.x = media, sigma.x = sigma, n.x = n, conf.level = conf)
```

```
    One-sample z-Test

data:  Summarized x
z = 19.14, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 2.147505 2.637495
sample estimates:
mean of x
   2.3925
```

Con un nivel de confianza del 95%, se estima que el consumo medio diario de agua está contenido en el intervalo $[2.15, 2.64]$ litros.

b. Suponiendo que σ es desconocida

```
agua = c(2.75, 2.43, 1.59, 2.82, 3.26, 2.38, 1.79, 2.08, 3.29, 2.88,
         2.27, 2.72, 2.62, 1.54, 1.64, 2.22)

media = agua |> mean()
s     = agua |> sd()
n     = agua |> length()
conf  = 0.95
vt    = qt((1+conf)/2, n-1)
LI    = media - vt*s/sqrt(n)
LS    = media + vt*s/sqrt(n)
c(LI,LS)
```

```
[1] 2.093929 2.691071
```

```r
agua |> t.test(conf.level = 0.95)
```

```
    One Sample t-test

data:  agua
t = 17.08, df = 15, p-value = 3.065e-11
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 2.093929 2.691071
sample estimates:
mean of x
   2.3925
```

Con un nivel de confianza del 95%, se estima que el consumo medio diario de agua está contenido en el intervalo $[2.09, 2.69]$ litros.

Intervalo de confianza para la variancia

Si $X_1,\ldots,X_n$ es una muestra aleatoria de una población normal con y σ^2 desconocida, el intervalo con un nivel de confianza del $(1-\alpha)\times 100\%$ para la variancia σ^2 se obtiene mediante

$$\frac{(n-1)s^2}{\chi^2_{(1-\alpha/2;n-1)}} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{(\alpha/2;n-1)}}$$

Interpretación: Con un nivel de confianza del $(1-\alpha)\times 100\%$, se estima que la variancia poblacional σ^2 esté contenida en el intervalo $[a,b]$.
Si se desea obtener los límites de confianza para la desviación estándar se obtiene la raíz cuadrada en la expresión anterior obteniéndose:

$$\sqrt{\frac{(n-1)s^2}{\chi^2_{(1-\alpha/2;n-1)}}} \leq \sigma \leq \sqrt{\frac{(n-1)s^2}{\chi^2_{(\alpha/2;n-1)}}}$$

Ejemplo (cont.)

```
agua = c(2.75, 2.43, 1.59, 2.82, 3.26, 2.38, 1.79, 2.08, 3.29, 2.88,
         2.27, 2.72, 2.62, 1.54, 1.64, 2.22)

varian = agua |> var()
n      = agua |> length()
conf   = 0.95
chi1   = qchisq((1-conf)/2, n-1)
chi2   = qchisq((1+conf)/2, n-1)
LI     = (n-1)*varian/chi2
LS     = (n-1)*varian/chi1
c(LI,LS)
```

```
[1] 0.1713196 0.7520275
```

```r
library(EnvStats)
varTest(agua, conf.level = 0.95)$conf.int |> as.numeric()
```

```
[1] 0.1713196 0.7520275
```

Con un nivel de confianza del 95%, se estima que la varianza del consumo diario de agua está contenida en el intervalo $[0.17, 0.75]$ litros^2.

```r
varTest(agua, conf.level = 0.95)$conf.int |> sqrt() |> as.numeric()
```

```
[1] 0.4139077 0.8671952
```

Con un nivel de confianza del 95%, se estima que la desviación estándar del consumo diario de agua está contenida en el intervalo $[0.41, 0.87]$ litros.

Intervalo de confianza para la proporción

Si X_1,...,X_n es una muestra aleatoria donde cada X_i indica la presencia (1) o ausencia (0) de una característica y $n > 30$, el intervalo con un nivel de confianza del $(1-\alpha) \times 100\%$ para la proporción π se obtiene mediante

$$p - Z_{(1-\alpha/2)}\sqrt{\frac{p(1-p)}{n}} \leq \pi \leq p + Z_{(1-\alpha/2)}\sqrt{\frac{p(1-p)}{n}}$$

Ejemplo

Una organización de salud realizó una encuesta en una comunidad para conocer la proporción de personas que consumen agua filtrada en sus hogares. Se entrevistó a 150 personas, registrando si consumen (1) o no consumen (0) agua filtrada. En total, 103 personas indicaron que sí consumen agua filtrada.

```
# Intervalo bajo la aproximación de Wald
# Para valores grandes de n o proporciones lejanas a 0 o 1
p    = 103/150
n    = 150
conf = 0.95
z    = qnorm((1 + conf)/2)
LI   = p - z*sqrt(p*(1-p)/n)
LS   = p + z*sqrt(p*(1-p)/n)
c(LI,LS)
```

```
[1] 0.6124368 0.7608965
```

```r
# Intervalo de Wilson sin corrección de Yates (n grande)
prop.test(x = 103, n = 150, conf.level = 0.95, correct = FALSE)$conf.int |>
  as.numeric()
```

```
[1] 0.6085602 0.7554509
```

```r
# Intervalo de Wilson con corrección de Yates (n pequeño)
# Más conservador que su versión sin corrección
prop.test(x = 103, n = 150, conf.level = 0.95, correct = TRUE)$conf.int |>
  as.numeric()
```

```
[1] 0.6051060 0.7584926
```

```r
# Intervalo exacto de Clopper-Pearson
# Para valores pequeños de n o proporciones cercanas a 0 o 1
# Suele ser el más amplio
binom.test(x = 103, n = 150, conf.level = 0.95)$conf.int |>
  as.numeric()
```

```
[1] 0.6059383 0.7598481
```

Pruebas de hipótesis

Hipótesis

Una hipótesis estadística es una afirmación sobre la distribución de probabilidad de una población o sobre el valor o valores de uno o más parámetros, como la media (μ), la variancia (σ^2) o la proporción (π).

Esta afirmación debe estar basada en la comprensión del fenómeno y sus variables. Una buena hipótesis permite hacer predicciones específicas y, si es rechazada, ayuda a revelar la complejidad del fenómeno.

Tipos de hipótesis estadísticas

Hipótesis nula (H_0 o H_p): Es la hipótesis que es aceptada provisionalmente como verdadera y cuya validez será sometida a verificación experimental. Los resultados experimentales nos permitirán seguir aceptándola como verdadera o si debemos rechazarla como tal.

Hipótesis alterna (H_1 o H_a): Es la hipótesis que se acepta en caso de que la hipótesis nula sea rechazada. La H_1 es la suposición contraria a H_0.

Prueba de hipótesis

Una prueba de hipótesis es un proceso estructurado para tomar decisiones basadas en datos. Se fundamenta en el método hipotético-deductivo, donde las hipótesis se contrastan con la evidencia en lugar de verificarse directamente.

Una prueba de hipótesis estadística es el proceso mediante el cual se toma la decisión de aceptar o rechazar la hipótesis nula.

El proceso de prueba de hipótesis determina si se rechaza o no H_0, pero **no se prueba su veracidad absoluta**.

Plantear hipótesis **antes del análisis de datos** (hipótesis a priori) mejora la solidez del estudio, enfocando la prueba en relaciones específicas y reduciendo sesgos.

Tipos de pruebas de hipótesis

En principio, se pueden formular hasta tres tipos de prueba, la cual dependerá de la forma de la hipótesis alterna que se planteé en el estudio:

Hipótesis unilateral con cola a la derecha	Hipótesis bilateral o de dos colas	Hipótesis unilateral con cola a la izquierda
$H_0 : \theta \leq \theta_0$ $H_1 : \theta > \theta_0$	$H_0 : \theta = \theta_0$ $H_1 : \theta \neq \theta_0$	$H_0 : \theta \geq \theta_0$ $H_1 : \theta < \theta_0$

donde θ es el parámetro de interés a probarse, pudiendo ser μ, σ^2, π (o algún otro que, por cuestiones de tiempo y/o complejidad no es abordado en el curso), y θ_0 es el valor o los valores supuestos que puede tomar el parámetro.

Aplicaciones

Una empresa de manufactura especializada en la producción de componentes electrónicos ha decidido evaluar el desempeño de dos equipos de trabajo en su planta principal. La gerencia está interesada en analizar la productividad y la calidad de los productos generados por cada equipo con el fin de mejorar la eficiencia operativa y reducir costos.

Los directivos han identificado diferencias percibidas en los tiempos de producción, costos y tasas de defectos entre los equipos. Sin embargo, antes de tomar decisiones estratégicas, desean realizar un análisis estadístico riguroso para determinar si estas diferencias son significativas o si podrían deberse al azar.

Se ha tomado una muestra aleatoria de los tiempos de producción (en minutos), costos de producción (en dólares) y tasas de defectos (1 = producto defectuoso, 0 = producto correcto) de dos equipos de trabajo en la planta.

Los datos están disponibles en el archivo **Produccion.csv**.

Aplicación 1

La gerencia de producción ha establecido que, para mantener la competitividad en el mercado, el tiempo promedio de producción por unidad debe ser de 45 minutos o menos. Sin embargo, han recibido reportes de que alguno de los equipos podría estar tardando más de lo esperado, lo que afectaría los tiempos de entrega y la satisfacción del cliente.

¿Los equipos realmente están cumpliendo con el estándar de 45 minutos, o en alguno (o ambos) los tiempos son mayores y se requiere una intervención?

$$H_0 : \mu \leq 45 \qquad H_1 : \mu > 45 \qquad \alpha = 0.05$$

```
datos = read.csv('Produccion.csv')
library(dplyr)
```

```r
datos |>
  filter(Grupo == "Equipo 1") |>
  pull(Tiempo) |>
  t.test(alternative = "greater", mu = 45)
```

```
    One Sample t-test

data:  pull(filter(datos, Grupo == "Equipo 1"), Tiempo)
t = -1.1487, df = 29, p-value = 0.87
alternative hypothesis: true mean is greater than 45
95 percent confidence interval:
 42.66133      Inf
sample estimates:
mean of x
 44.05667
```

No se rechaza H_0, ya que $pv > \alpha$

```r
datos |>
  filter(Grupo == "Equipo 2") |>
  pull(Tiempo) |>
  t.test(alternative = "greater", mu = 45)
```

```
    One Sample t-test

data:  pull(filter(datos, Grupo == "Equipo 2"), Tiempo)
t = 2.17, df = 34, p-value = 0.01855
alternative hypothesis: true mean is greater than 45
95 percent confidence interval:
 45.44277      Inf
sample estimates:
mean of x
 47.00571
```

Se rechaza H_0, ya que $pv < \alpha$ ¿Cuál es la respuesta a la problemática planteada?

Aplicación 2

La gerencia ha establecido que la desviación estándar debe ser menor a 3 dólares para ser considerada aceptable. Sin embargo, hay indicios de que el equipo 2 está experimentando una variabilidad en costos fuera de lo esperado, lo que puede afectar la planificación financiera y el control de presupuesto. Verificar esta afirmación.

$$H_0 : \sigma^2 \geq 9 \qquad H_1 : \sigma^2 < 9 \qquad \alpha = 0.05$$

```r
library(EnvStats)
datos |>
  filter(Grupo == "Equipo 2") |>
  pull(Costo) |>
  varTest(alternative = "less", sigma.squared = 9)
```

```
Results of Hypothesis Test
--------------------------

Null Hypothesis:                 variance = 9

Alternative Hypothesis:          True variance is less than 9

Test Name:                       Chi-Squared Test on Variance

Estimated Parameter(s):          variance = 9.097422

Data:                            pull(filter(datos, Grupo == "Equipo 2"), Costo)

Test Statistic:                  Chi-Squared = 34.36804

Test Statistic Parameter:        df = 34

P-value:                         0.549879

95% Confidence Interval:         LCL =  0.00000
```

Aplicación 3

La empresa de manufactura tiene un estándar de calidad que establece que el porcentaje de productos defectuosos debe ser menor al 3%. Sin embargo, hay sospechas de que se podría estar generando una tasa de defectos diferente a la esperada, lo que podría afectar la satisfacción del cliente y aumentar los costos de reproceso.

$$H_0 : \pi \geq 0.03 \qquad H_1 : \pi < 0.03 \qquad \alpha = 0.05$$

```r
datos |> nrow() -> n
datos |> filter(Defectuoso == 1) |> nrow() -> x
prop.test(x, n, p = 0.03, alternative = "less", correct = FALSE)
```

```
    1-sample proportions test without continuity correction

data:  x out of n, null probability 0.03
X-squared = 0.58287, df = 1, p-value = 0.7774
alternative hypothesis: true p is less than 0.03
95 percent confidence interval:
 0.0000000 0.1099856
sample estimates:
         p
0.04615385
```

Aplicación 4

En la empresa de manufactura, ambos equipos de producción que fabrican el mismo producto. La gerencia de calidad ha detectado que uno de los equipos podría estar generando más productos defectuosos que el otro, lo que podría afectar la rentabilidad y la satisfacción del cliente.
¿Existe una diferencia significativa en la tasa de defectos entre los dos equipos o las diferencias muestrales se deben al azar?

$$H_0 : \pi_1 - \pi_2 = 0 \qquad H_1 : \pi_1 - \pi_2 \neq 0 \qquad \alpha = 0.05$$

```
datos |> group_by(Grupo) |> count(Defectuoso) |>
  filter(Defectuoso == 1) |> pull(n) -> x

x
```

```
[1] 1 2
```

```
datos |> count(Grupo) |>
  pull(n) -> n

n
```

```
[1] 30 35
```

```r
prop.test(x, n, alternative = "two.sided")
```

```
    2-sample test for equality of proportions with continuity correction

data:  x out of n
X-squared = 8.8709e-31, df = 1, p-value = 1
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.1478158  0.1001968
sample estimates:
    prop 1     prop 2
0.03333333 0.05714286
```

Decisión: No se rechaza H_0

Aplicación 5

Existen indicios de que uno de los equipos podría estar tardando más en completar sus tareas, lo que impactaría la eficiencia y los tiempos de entrega. ¿El tiempo promedio de producción por unidad es el mismo en ambos equipos, o hay una diferencia significativa?

Sol.

Primero se verificará si las varianzas son iguales

$$H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1 \qquad H_0 : \frac{\sigma_1^2}{\sigma_2^2} \neq 1 \qquad \alpha = 0.05$$

```r
var.test(Tiempo ~ Grupo, datos, alternative = "two.sided")
```

```
    F test to compare two variances

data:  Tiempo by Grupo
F = 0.67659, num df = 29, denom df = 34, p-value = 0.2865
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.3348314 1.3952971
sample estimates:
ratio of variances
         0.6765865
```

```r
tiempo1 <- datos |> filter(Grupo == "Equipo 1") |> pull(Tiempo)
tiempo2 <- datos |> filter(Grupo == "Equipo 2") |> pull(Tiempo)
var.test(tiempo1, tiempo2, alternative = "two.sided", var.equal = TRUE)
```

```
    F test to compare two variances

data:  tiempo1 and tiempo2
F = 0.67659, num df = 29, denom df = 34, p-value = 0.2865
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.3348314 1.3952971
sample estimates:
ratio of variances
         0.6765865
```

No se rechaza H_0 ya que $pv > \alpha$, por lo tanto las varianzas son homogéneas.

Ahora, se prueba la diferencia de medias considerando que las varianzas son homogéneas:

$$H_0 : \mu_1 - \mu_2 = 0 \qquad H_1 : \mu_1 - \mu_2 \neq 0 \qquad \alpha = 0.05$$

```
    Two Sample t-test

data:  Tiempo by Grupo
t = -2.3495, df = 63, p-value = 0.02195
alternative hypothesis: true difference in means between group Equipo 1 and
95 percent confidence interval:
 -5.4573575 -0.4407378
sample estimates:
mean in group Equipo 1 mean in group Equipo 2
              44.05667               47.00571
```

```r
tiempo1 <- datos |> filter(Grupo == "Equipo 1") |> pull(Tiempo)
tiempo2 <- datos |> filter(Grupo == "Equipo 2") |> pull(Tiempo)
t.test(tiempo1, tiempo2, alternative = "two.sided", var.equal = TRUE)
```

```
    Two Sample t-test

data:  tiempo1 and tiempo2
t = -2.3495, df = 63, p-value = 0.02195
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -5.4573575 -0.4407378
sample estimates:
mean of x mean of y
 44.05667  47.00571
```

Se rechaza H_0, ya que $pv < \alpha$, lo que indica que hay diferencias significativas entre los tiempos promedio de producción de los equipos.

Ejercicio

Una empresa de logística busca mejorar la eficiencia de su operación y ha identificado dos centros de distribución que manejan envíos a distintas regiones. La gerencia ha recibido reportes sobre diferencias en los tiempos de entrega, costos de distribución y la cantidad de paquetes entregados con demora entre los centros. Antes de tomar decisiones estratégicas, desean realizar un análisis estadístico riguroso para determinar si estas diferencias son significativas o si pueden atribuirse al azar.

Se ha tomado una muestra aleatoria de los tiempos de entrega (en horas), costos de distribución (en dólares) y registros de entregas tardías ($1 =$ entrega tardía, $0 =$ entrega a tiempo) en cada centro de distribución.

Los datos están disponibles en el archivo Logistica.csv.

1. La empresa ha establecido que el tiempo promedio de entrega debe ser menor a 24 horas para cumplir con los estándares de servicio. Sin embargo, hay sospechas de que el centro 1 no está cumpliendo esta especificación, lo que podría generar retrasos en la entrega.
2. La empresa busca mantener estabilidad en los costos de distribución. Se ha establecido que la desviación estándar no debe ser mayor a 2 dólares. Hay indicios de que el centro 2 tiene costos con alta variabilidad. Verifique esta afirmación.
3. La empresa establece que la tasa máxima aceptable de entregas tardías es del 6%. Se sospecha que uno de los centros está superando esta tasa, lo que afectaría la percepción del servicio por parte de los clientes.
4. ¿Existe una diferencia significativa en la tasa de entregas tardías entre los dos centros, o las diferencias observadas son solo aleatorias?
5. ¿El tiempo promedio de entrega en del centro 1 es más de 3 horas mayor que en el otro, o la diferencia observada es solo variabilidad aleatoria?

Regresión lineal