XS3310 Teoría Estadística

I Semestre 2021

Escuela de Estadística

2021-06-14

class: center, middle

¿Qué hemos visto hasta ahora?

Todo sobre estimadores puntuales + pivotes e intervalos de confianza. Bootstrap y contrastes de hipótesis (función de potencia, tamaño del contraste, el valor p, contrastes más potentes, uniformemente más potentes, cocientes de verosimilitud y razón de verosimiliud).

¿Qué vamos a discutir hoy?

Contrastes de hipótesis: Razón de verosimilitudes para muestras grandes. Bootstrap para contrastes.

Prueba de comparación de medias en 2 poblaciones

Comparación de medias normales

Asuma que $X_1, \ldots, X_n \overset{i.i.d}{\sim} N(\mu_1, \sigma^2)$ y $Y_1, \ldots, Y_n \overset{i.i.d}{\sim} N(\mu_2, \sigma^2)$. Los parámetros desconocidos son μ_1, μ_2, σ^2. Asuma que (X_i, Y_i) son independientes y la varianza es la misma (homocedasticidad).

Hipótesis: $H_0 : \mu_1 \leq \mu_2$ vs $H_1 : \mu_1 > \mu_2$.

Notación: $\bar{X}_m, \bar{Y}_n$, $S_X^2 = \sum_{i=1}^{m}(X_i - \bar{X}_m)^2$, $S_Y^2 = \sum_{i=1}^{m}(Y_i - \bar{Y}_n)^2$.

Teorema. Considere

$$U = \frac{(m+n-2)^{1/2}(\bar{X}_m - \bar{Y}_n)}{\left(\frac{1}{m} + \frac{1}{n}\right)^{1/2}(S_X^2 + S_Y^2)^{1/2}}.$$

Si $\mu_1 = \mu_2 \implies U \sim t_{m+n-2}$.

Prueba t de dos muestras

Para el caso de una prueba de una cola, se define la región de rechazo como $U \geq c$,

$$\begin{aligned}\sup_{\mu_1 \leq \mu_2} \mathbb{P}[U \geq c|\mu_1, \mu_2, \sigma^2] \leq \alpha_0 &\implies \mathbb{P}[U \geq c|\mu_1 = \mu_2, \sigma^2] = 1 - \mathbb{P}[U \leq c|\mu_1 = \mu_2, \sigma^2] \leq \alpha_0 \\ &\implies P[U \leq c|\mu_1 = \mu_2, \sigma^2] \geq 1 - \alpha_0 \\ &\implies c = t_{1-\alpha_0, n+m-2}\end{aligned}$$

Rechazo H_0 si $U > t_{1-\alpha_0, n+m-2}$.

si observamos $U = u$, los p-valores son:

- Si $H_1 : \mu_1 - \mu_2 > 0$ entonces $1 - P[U \leq c|\mu_1 = \mu_2, \sigma^2]$.
- Si $H_1 : \mu_1 - \mu_2 < 0$ entonces $P[U \leq c|\mu_1 = \mu_2, \sigma^2]$.

Ejemplo: En el caso de las lluvia suponga que queremos probar

$$H_0 : \mu_{\text{con trat.}} \leq \mu_{\text{sin trat.}} \quad vs \quad H_1 : \mu_{\text{con trat.}} > \mu_{\text{sin trat.}}$$

```
nubes <- read.table(
    file = "./data/clouds.txt",
    sep = "\t", header = TRUE
)
log_lluvia <- log(nubes)

n <- nrow(nubes)

con_tratamiento <- log_lluvia$Seeded.Clouds
sin_tratamiento <- log_lluvia$Unseeded.Clouds

(Xbar <- mean(con_tratamiento))
```

```
## [1] 5.134187
```

```
(Ybar <- mean(sin_tratamiento))
```

```
## [1] 3.990406
```

```
(S2_X <- (n - 1) * var(con_tratamiento))
```

```
## [1] 63.96109
```

```
(S2_Y <- (n - 1) * var(sin_tratamiento))
```

```
## [1] 67.39158
```

Entonces el estadístico que queremos construir para comparar la medias es (OJO en este caso $m = n$ porque tienen la misma cantidad de datos:)

```
(U <- sqrt(n + n - 2) * (Xbar - Ybar) /
    (sqrt(1 / n + 1 / n) * sqrt(S2_X + S2_Y)))
```

```
## [1] 2.544369
```

Por tanto se debe comparar con una t-student con $26 + 26 - 2 = 50$ grados de libertad. Asuma un $\alpha = 0.01$

```
(qnt <- qt(p = 1 - 0.01, df = n + n - 2))
```

```
## [1] 2.403272
```

¿ Rechazamos H_0?

```
U > qnt
```

```
## [1] TRUE
```

¿Cuál es el p-valor?

```
1 - pt(q = U, df = n + n - 2)
```

```
## [1] 0.007041329
```

Interpretación: rechazamos al nivel 1% de significancia la hipótesis de que las nubes irradiadas tienen una log-precipitación media menor a la de las nubes no irradiadas.

Prueba de 2 colas

Hipótesis. $H_0 : \mu_1 = \mu_2$ vs $H_1 : \mu_1 \neq \mu_2$ (Prueba ANOVA).

- Prueba. δ : Rechazo H_0 si $|U| \geq t_{1-\frac{\alpha_0}{2}, n+m-2}$.
- Valor-p: $2[1 - P[U \leq |u||\mu_1 = \mu_2, \sigma^2])]$ donde $U = u$.

Ejemplo. Minas de cobre. Sean $X_1, \ldots, X_8$ la cantidad de cobre (gramos) en 8 minas en un lugar 1, y $X_1, \ldots, X_{10}$ en 10 minas en un lugar 2. Después de recolectar los datos se obtiene lo siguiente

- $\bar{X}_8 = 2.6$
- $\bar{Y}_{10} = 2.3$
- $S_X^2 = 0.32$ y
- $S_Y^2 = 0.22$

El ingeniero de la mina se pregunta: ¿Las dos localizaciones generan el mismo nivel de cobre?

Entonces plantea hacer la prueba de hipótesis

$$H_0 : \mu_1 = \mu_2 \quad H_1 : \mu_1 \neq \mu_2$$

Con el supuesto que $X_i \sim N(\mu_1, \sigma^2)$, $Y_j \sim N(\mu_2, \sigma^2)$.

```r
n <- 8
m <- 10

n + m - 2
```

```
## [1] 16
```

```r
Xbar <- 2.6
Ybar <- 2.3

S2_X <- 0.32
S2_Y <- 0.22

(U <- sqrt(n + m - 2) * (Xbar - Ybar) /
    (sqrt(1 / n + 1 / m) * sqrt(S2_X + S2_Y)))
```

```
## [1] 3.442652
```

Si $\alpha_0 = 1\%$

```r
(qnt <- qt(p = 1 - 0.01 / 2, df = n + m - 2))
```

```
## [1] 2.920782
```

Entonces, ¿Rechazamos H_0?

```r
abs(U) > qnt
```

```
## [1] TRUE
```

El valor p es $2[1 - T_{16}(|3.442|)]$

```r
2 * (1 - pt(q = U, df = n + m - 2))
```

```
## [1] 0.003345064
```

Interpretación: Rechazamos al 1% de significancia la hipótesis de una diferencia no significativa entre las cantidades medias de cobre en cada localización.

Ejercicio. La prueba t de 2 muestras es un LRT.

Prueba F

Definición Si Y y W son variables aleatorias independientes, $Y \sim \chi^2_m$ y $W \sim \chi^2_n$, $m, n \in \mathbb{Z}^+$. Defina

$$X = \frac{Y/m}{W/n} \sim F_{m,n}$$

X tiene una distribución F con m y n grados de libertad.

La función de densidad es

$$f(x) = \begin{cases} \dfrac{\Gamma\left[\frac{1}{2}(m+n)\right] m^{m/2} n^{n/2}}{\Gamma\left(\frac{1}{2}m\right)\Gamma\left(\frac{1}{2}n\right)} \cdot \dfrac{x^{(m/2)-1}}{(mx+n)^{(m+n)/2}} & x > 0 \\ 0 & x \leq 0. \end{cases} \tag{1}$$

Propiedades:

1. Si $X \sim F_{m,n} \implies 1/X \sim F_{n,m}$.
2. Si $Y \sim t_n \implies Y^2 \sim F_{1,n}$.

Sean $X_1, \ldots, X_n \overset{i.i.d}{\sim} N(\mu_1, \sigma_1^2)$ y $Y_1, \ldots, Y_n \overset{i.i.d}{\sim} N(\mu_2, \sigma_2^2)$.
Considere el esquema

$$\begin{array}{ccc} U \sim t_{n-1} & & U^2 \sim F_{1,n-1} \\ H_0 : \mu = \mu_0 & \Leftrightarrow & H_0 : \mu = \mu_0 \\ |U| \geq |c| & & U^2 \geq c^* \end{array}$$

Bajo el esquema anterior y si (X, Y) son independientes, considere:

$$H_0 : \sigma_1^2 \leq \sigma_2^2 \text{ vs } H_1 : \sigma_1^2 > \sigma_2^2$$

y tome $\alpha_0 \in (0, 1)$.

La lógica de esta prueba es, como $\dfrac{S_X^2}{\sigma_1^2} \sim \chi_{m-1}^2$ y $\dfrac{S_Y^2}{\sigma_2^2} \sim \chi_{n-1}^2$, calculamos

$V^* = \dfrac{\dfrac{S_X^2/\sigma_1^2}{m-1}}{\dfrac{S_Y^2/\sigma_2^2}{n-1}} \sim F_{m-1,n-1}$. Bajo el supuesto de homocedasticidad,

$V = \dfrac{\dfrac{S_X^2}{m-1}}{\dfrac{S_Y^2}{n-1}} \sim F_{m-1,n-1}$.

δ : Rechazo H_0 si $V \geq c$.

Usando el δ anterior

$$\sup_{\sigma_1^2 \leq \sigma_2^2} \mathbb{P}[V \geq c | \mu_1 \mu_2, \sigma_1^2, \sigma_2^2] \leq \alpha_0,$$

resuelve

$$\mathbb{P}[V \geq c | \mu_1, \mu_2, \sigma_1^2, \sigma_2^2] = \alpha_0 \implies c = F_{1-\alpha_0, m-1, n-1} = F_{1-\alpha_0, m-1, n-1}.$$

El valor-p es $1 - \mathbb{P}_{m,n}(V \leq v | \sigma_1 = \sigma_2)$ y con $V = v$.

Ejemplo. $X_1, \ldots, X_6 \sim N(\mu_1, \sigma_1^2)$, $S_X^2 = 30$, $Y_1, \ldots, Y_{21} \sim N(\mu_2, \sigma_2^2)$, $S_Y^2 = 30$.
La hipótesis nula es $H_0 : \sigma_1^2 \leq \sigma_2^2$.

Se calcula $V = \dfrac{30/5}{40/20} = 3$ y $F1 - 0.05, 5, 20(1 - 0.05)$.

El valor-p corresponde a $1 - \mathbb{P}_{5,20}(V < 3|\sigma_1 = \sigma_2) = 0.035$.

Si $\alpha_0 = 1\%$, no rechazo. Si $\alpha_0 = 5\%$ rechazo.

Ejemplo: Suponga que se tienen los siguientes datos

```
m <- 20
X <- rnorm(n = m, mean = 0, sd = sqrt(6))
head(X)
```

```
## [1]  0.9242208 -2.5393911 -0.9310742  3.1244809 -1.4094041  2.6060968
```

```
n <- 40
Y <- rnorm(n = n, mean = 10, sd = sqrt(2))
head(Y)
```

```
## [1]  8.785633 10.508503  8.151788 10.079816  9.189266 10.942824
```

Es decir tener 20 datos normales con $\sigma_1^2 = 6$ y 40 datos normales con $\sigma_2^2 = 2$.

En todo caso asuma que σ es desconocidos para cada caso y solo tenemos los datos. Además queremos hacer la prueba de hipótesis

$$\begin{array}{ll} H_0: & \sigma_1^2 \leq \sigma_2^2 \\ H_1: & \sigma_1^2 > \sigma_2^2 \end{array}$$

OJO: Según la forma que planteamos el ejercicio, deberíamos de rechazar H_0 ya que $\sigma_1^2 = 6 > 2 = \sigma_2^2$

Calculamos el estadístico V

```
(S2_X_divido_m_1 <- var(X))
```

```
## [1] 5.2134
```

```
(S2_Y_divido_n_1 <- var(Y))
```

```
## [1] 1.887867
```

```
(V <- S2_X_divido_m_1 / S2_Y_divido_n_1)
```

```
## [1] 2.761529
```

Para calcular un cuantil te tamaño $1 - \alpha = 0.95$ se usa la siguiente función

```
(qnt <- qf(p = 1 - 0.05, df1 = m - 1, df2 = n - 1))
```

```
## [1] 1.85992
```

¿Rechazamos H_0?

```r
V > qnt
```

```
## [1] TRUE
```

y el valor-p de la prueba es

```r
1 - pf(q = V, df1 = m - 1, df2 = n - 1)
```

```
## [1] 0.003570364
```

Interpretación: Rechazamos la hipótesis que $\sigma_1^2 \leq \sigma_2^2$ con un valor-p de 0.02.

Prueba de 2 colas (prueba de homocedasticidad)

Bajo las hipótesis $H_0 : \sigma_1^2 = \sigma_2^2$ vs $H_1 : \sigma_1^2 \neq \sigma_2^2$, se rechaza si $V \geq c_2$ o $V \leq c_1$ con c_1, c_2 tales que

$$\mathbb{P}[V \leq c_1] = \frac{\alpha_0}{2} \text{ y } \mathbb{P}[V \geq c_2] = \frac{\alpha_0}{2} \implies c_1 = F_{\frac{\alpha_0}{2},m-1,n-1} \text{ y } c_2 = F_{1-\frac{\alpha_0}{2},m-1,n-1}$$

Ejemplo. Mismo ejemplo de las nubes.

$$H_0 : \sigma^2_{\text{con trat.}} = \sigma^2_{\text{sin trat.}} \quad vs \quad H_1 : \sigma^2_{\text{con trat.}} \neq \sigma^2_{\text{sin trat.}}$$

```r
(m <- length(con_tratamiento))
```

```
## [1] 26
```

```r
(n <- length(sin_tratamiento))
```

```
## [1] 26
```

```r
(S2_X_divido_m_1 <- var(con_tratamiento))
```

```
## [1] 2.558444
```

```r
(S2_Y_divido_n_1 <- var(sin_tratamiento))
```

```
## [1] 2.695663
```

```r
(V <- S2_X_divido_m_1 / S2_Y_divido_n_1)
```

```
## [1] 0.9490963
```

$$V = \frac{\frac{63.96}{25}}{\frac{67.39}{25}} = 0.9491$$

Se tiene que $c_1 = F_{0.025,25,25} = 0.4484$ y $c_2 = F_{0.975,25,25} = 2.23$.

```r
(c1 <- qf(0.025, df1 = m - 1, df2 = n - 1))
```

```
## [1] 0.4483698
```

```r
(c2 <- qf(0.975, df1 = m - 1, df2 = n - 1))
```

```
## [1] 2.230302
```

¿Rechazamos H_0?

```r
V < c1
```

```
## [1] FALSE
```

```r
V > c2
```

```
## [1] FALSE
```

No rechazamos la hipótesis nula.

Si observamos $V = v$, podemos rechazar si

$$v \leq F_{\frac{\alpha_0}{2},m-1,n-1} \implies 2\mathbb{P}_{m-1,n-1}(V \leq v|\sigma_1 = \sigma_2) \leq \alpha_0$$

o tambien si

$$v \geq F_{1-\frac{\alpha_0}{2},m-1,n-1} \implies P_{m-1,n-1}(V \leq v|\sigma_1 = \sigma_2) \geq 1 - \frac{\alpha_0}{2} \implies \alpha_0 \geq 2(1 - P_{m-1,n-1}(V \leq v|\sigma_1 = \sigma_2))$$

El p-valor es

$$2\min[1 - \mathbb{P}_{m-1,n-1}(V \leq v|\sigma_1 = \sigma_2), \mathbb{P}_{m,n}(V \leq v|\sigma_1 = \sigma_2)]$$

```r
2 * min(1 - pf(q = V, df1 = m - 1, df2 = n - 1),
        pf(q = V, df1 = m - 1, df2 = n - 1))
```

```
## [1] 0.8971154
```

Interpretación: La prueba de hipótesis no rechaza la hipótesis de homocedasticidad con un nivel de confianza del 5%.

Propiedad. La prueba F es un LRT.

Bootstrap

¿Qué pasa cuando una aproximación no es suficiente, o cuando queremos una segunda opinión?

Idea: podemos remuestrar el estadístico T para construir la distribución empírica y así calcular el valor p de una manera empírica.

Sea X y Y dos muestras de dos poblaciones distribuidas como P y Q, dos distribuciones posiblemente distintas y desconocidas. Nos interesa contrastar la hipótesis nula de igualdad de distribuciones:

$H_0 : P = Q \quad \text{vs} \quad H_0 : P \neq Q$

Asuma que existe un estadístico de prueba adecuado T para construir el contraste para este problema, en ese caso cuando observamos $T = t$ para el estadístico de prueba, y tenemos evidencia para rechazar hipótesis nula con un tamaño de contraste de α si:

$$P(T \geq t) \leq \alpha$$

bajo la hipótesis nula.

Bootstrap

En muchas aplicaciones, la distribución muestral del estadístico de prueba T no es conocido (o exactamente conocido), y el valor p no se puede calcular. Esto sugiere el uso de bootstrap para estimar el valor p con:

$$\hat{P}(T \geq t) = P^*(T^* \geq t)$$

Un asunto importante de aclarar en este punto, es que para calcular el valor p, SIEMPRE vamos a muestrar asumiendo que la hipótesis nula es cierta.

Por ejemplo, para el problema anterior, se remuestrea $X^{*(b)}$ y $Y^{*(b)}$ de una muestra conjunta (X, Y). De estas muestras bootstrap, podemos calcular las iteraciones bootstrap del estadístico T:

$$T^{*(b)} = s(X^{*(b)}, Y^{*(b)})$$

y luego estimar el valor p con:

$$\hat{P}(T \geq t) = \frac{1}{B} \sum_{b=1}^{B} 1\{T^{*(b)} \geq t\}$$

.

Bootstrap

Ejemplo 1: Datos de ratones.

Tengo datos de sobrevivencia de 16 ratones luego de una cirugía de prueba: hay 9 ratones en el grupo control y 7 ratones en el grupo de tratamiento. La pregunta de investigación es si el nuevo tratamiento prolonga el tiempo de sobrevivencia.

Group	Survival time (in days)	Mean
Treatment	94,197,16,38,99,141,23	86.86
Control	52,104,146,10,51,30,40,27,46	56.22

Es decir, queremos contrastar la siguiente hipótesis:

$$H_0 : \mu_X = \mu_Y \quad \text{vs} \quad H_0 : \mu_X \neq \mu_Y$$

A diferencia de la hipótesis general que vimos antes, en este caso la hipótesis nula requiere únicamente la igualdad de las medias, pero no por ejemplo de las variancias.

Bootstrap

Como las medias son distintas y necesitamos generar datos asumiendo que la hipótesis nula es cierta, entonces podemos hacer una pequeña transformación a los datos originales (para poder generar datos bajo el supuesto de que la hipótesis nula es cierta):

$$\tilde{X}_i = X_i - \bar{X} + \bar{Z}$$
$$\tilde{Y}_i = Y_i - \bar{Y} + \bar{Z}$$

donde:

$$\bar{Z} = \frac{1}{n_X + n_Y}[\sum_{i=1}^{n_X} X_i + \sum_{i=1}^{n_Y} Y_i]$$

.

Ahora, con esa transformación, la distribución empírica de las dos variables transformadas tendrá iguales medias y por ende, satisface la condición de que la hipótesis nula es verdadera.

Algoritmo de Bootstrap

1. Remuestree $X_1^{*(b)}, \dots X_{n_X}^{*(b)}$ independientemente de $\tilde{X}$.
2. Remuestree $Y_1^{*(b)}, \dots Y_{n_X}^{*(b)}$ independientemente de $\tilde{Y}$.
3. Evalúe las iteraciones de bootstrap:

$$T^{*(b)} = \frac{\bar{X}^{*(b)} - \bar{Y}^{*(b)}}{\sqrt{\frac{s^2_{X^{*(b)}}}{n_X} + \frac{s^2_{Y^{*(b)}}}{n_Y}}}$$

4. Estime el valor de p:

$$\hat{P}(T \geq t) = \frac{1}{B}\sum_{b=1}^{B} 1\{T^{*(b)} \geq t\}$$

donde t es el valor observado del contraste usando el estadístico t.

Específicamente, en el ejemplo de los ratones, el valor observado de T era $t = 1.06$. Con $B = 1000$ iteraciones bootstrap de T, 133 eran mayores o iguales a t, entonces $\hat{P}(T \geq t) = 0.133$ y no encontramos evidencia para rechazar la hipótesis nula.

Bootstrap

- Útil cuando no tenemos la distribución empírica del estadístico de prueba.
- Puede ser difícil encontrar la transformación que permita remuestrear de una muestra asumiendo la hipótesis nula como cierta.
-

Recuerden que tanto en este caso como en los ejercicios de simulación, estamos calculando el valor p con aproximaciones empíricas.

Ejercicio para la siguiente clase:

Problema 4. Contrastes de hipótesis clásicos (20 puntos). Sea $X_1, X_2, \ldots, X_m$ una muestra aleatoria de la densidad exponencial con media θ_1 y sea $Y_1, Y_2, \ldots, Y_n$ una muestra aleatoria independiente de una densidad exponencial con media θ_2.

a. Encuentre el criterio de razón de Verosimilitud para probar $H_0 : \theta_1 = \theta_2$ contra $H_a : \theta_1 \neq \theta_2$. (7 pts)

b. Demuestre que la prueba del inciso a es equivalente a una prueba F exacta. Sugerencia: transforme $\sum X_i$ y $\sum Y_j$ en variables aleatorias distribuidas como Chi Cuadradas y construya el estadístico F. (7 pts)

c. Encuentre el contraste aproximado utilizando el teorema de Wilks. Defina claramente el estadístico de prueba, la regla de decisión y la distribución correspondiente. (No hace falta concluir en términos del contraste en este caso) (6 pts)

---

class: center, middle

¿Qué discutimos hoy?

Contrastes de hipótesis: Razón de verosimilitud y bootstrap.