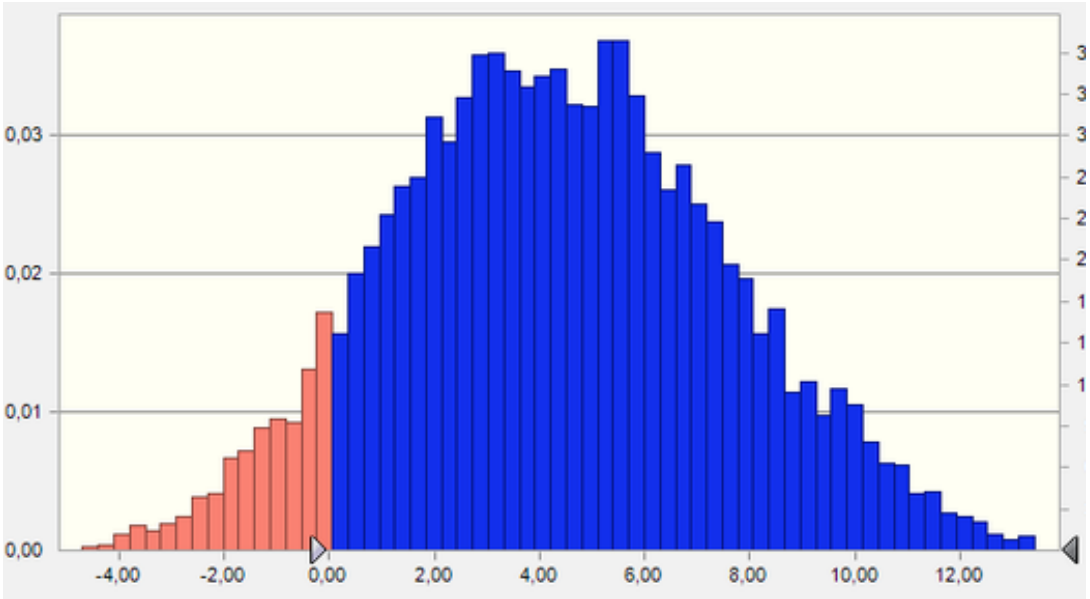
Problemas de dos muestras



Mike Wiper

Departamento de Estadística

Universidad Carlos III de Madrid

Grado en Estadística y Empresa

Objetivo

		$\nu = 45°$			
	ν_2	75%	90%	95%	97.5%
	6	0.71	1.39	1.84	2.27
$\nu_1 = 24$	8	0.71	1.38	1.80	2.18
	12	0.70	1.35	1.75	2.11
	24	0.69	1.33	1.71	2.06
	∞	0.68	1.30	1.68	2.01

Mostrar como resolver un famoso problema de dos muestras utilizando el método Monte Carlo.

Problemas de dos muestras

Queremos estimar la diferencia entre las medias: $\delta = \mu_1 - \mu_2$ para dos poblaciones normales: Normal $\left(\mu_i, \sigma_i^2\right)$ para $i = 1, 2$.

Tomamos dos muestras independientes de tamaños n_i, and medias $\bar{y}_i$ y varianzas muestrales s_i^2, para $i = 1, 2$.

¿Cómo podemos hacer inferencia clásica y bayesiana para este problema?

Problemas de dos muestras

Queremos estimar la diferencia entre las medias: $\delta = \mu_1 - \mu_2$ para dos poblaciones normales: Normal $\left(\mu_i, \sigma_i^2\right)$ para $i = 1, 2$.

Tomamos dos muestras independientes de tamaños n_i, and medias $\bar{y}_i$ y varianzas muestrales s_i^2, para $i = 1, 2$.

¿Cómo podemos hacer inferencia clásica y bayesiana para este problema?

Depende de lo que suponemos sobre las varianzas.

Varianzas conocidas

Usando la inferencia frecuentista, tenemos

$$\bar{Y}_1 - \bar{Y}_2 \sim \text{Normal}\left(\delta, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

y un intervalo de $100(1-\alpha)\,\%$ de confianza para δ es:

$$\bar{y}_1 - \bar{y}_2 \pm z\left(1 - \frac{\alpha}{2}\right)\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}.$$

Desde el enfoque bayesiano, con distribuciones a priori "no informativas" $f(\mu_i) \propto 1$ para $i = 1, 2$, tenemos:

$$\mu_i|\text{datos} \sim \text{Normal}\left(\bar{y}_i, \frac{\sigma_i^2}{n_i}\right) \quad \text{para } i = 1, 2 \text{ y luego,}$$

$$\delta|\text{datos} \sim \text{Normal}\left(\bar{y}_1 - \bar{y}_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

y el intervalo bayesiano coincide con el clásico.

Varianzas desconocidas pero iguales: enfoque frecuentista

En la inferencia frecuentista, es fácil ver que

$$S_c^2 = \frac{1}{n_1 + n_2 - 2}\left[(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2\right]$$

es un estimador insesgado de la varianza $\sigma^2 = \sigma_1^2 = \sigma_2^2$ y luego tememos que el intervalo es

$$\bar{y}_1 - \bar{y}_2 \pm t_{n_1+n_2-2}\left(1 - \frac{\alpha}{2}\right) s_c \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}.$$

Varianzas desconocidas pero iguales: enfoque bayesiano

Desde el enfoque bayesiano, escribiendo $\phi = \frac{1}{\sigma^2}$ y usando la distribución a priori "no informativa" $f(\mu_1, \mu_2, \phi) \propto \frac{1}{\phi}$, tenemos

$$\mu_i|\phi, \text{datos} \sim \text{Normal}\left(\bar{y}_i, \frac{1}{n_i\phi}\right) \quad \text{para } i = 1, 2,$$

$$\phi|\text{datos} \sim \text{Gamma}\left(\frac{n_1 + n_2 - 2}{2}, \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{2}\right) \quad \text{y luego}$$

$$\delta|\phi, \text{datos} \sim \text{Normal}\left(\bar{y}_1 - \bar{y}_2, \left(\frac{1}{n_1} + \frac{1}{n_2}\right)\frac{1}{\phi}\right)$$

y usando los resultados para la distribución normal gamma, tenemos, otra vez, que el intervalo bayesiano coincide con el intervalo clásico.

Varianzas desconocidas pero iguales: el problema de Behrens y Fisher

Recordamos que para σ_1^2, σ_2^2 conocidos, se tiene:

$$\bar{Y}_1 - \bar{Y}_2 \sim \text{Normal}\left(\delta, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right).$$

Luego, si n_1 y n_2 son muy grandes, para implementar la inferencia frecuentista, sería suficiente sustituir la varianza de $\bar{Y}_1 - \bar{Y}_2$ con

$$\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}$$

en la fórmula para el intervalo con varianzas conocidas.

¿Qué hacemos si los tamaños muestrales no son tan grandes?

La "solución" frecuentista

La aproximación de Welch sugiere estimar la distribución de $\bar{Y}_1 - \bar{Y}_2$ con una distribución t de Student (escalada y no centrada) con grados de libertad

$$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{s_1^4}{n_1^2(n_1-1)} + \frac{s_2^4}{n_2^2(n_2-1)}}.$$

Luego, el intervalo es:

$$\bar{y}_1 - \bar{y}_2 \pm t_\nu \left(1 - \frac{\alpha}{2}\right) \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

La "solución" frecuentista

La aproximación de Welch sugiere estimar la distribución de $\bar{Y}_1 - \bar{Y}_2$ con una distribución t de Student (escalada y no centrada) con grados de libertad

$$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{s_1^4}{n_1^2(n_1-1)} + \frac{s_2^4}{n_2^2(n_2-1)}}.$$

Luego, el intervalo es:

$$\bar{y}_1 - \bar{y}_2 \pm t_\nu \left(1 - \frac{\alpha}{2}\right) \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

Problema: la verdadera nivel de confianza depende de la razón de varianzas desconocidas $\frac{\sigma_1^2}{\sigma_2^2}$. ¡No es $100(1-\alpha)\,\%$!

Existen métodos alternativos: bootstrap, Mann Whitney, ... pero todos sufren problemas.

Ejemplo

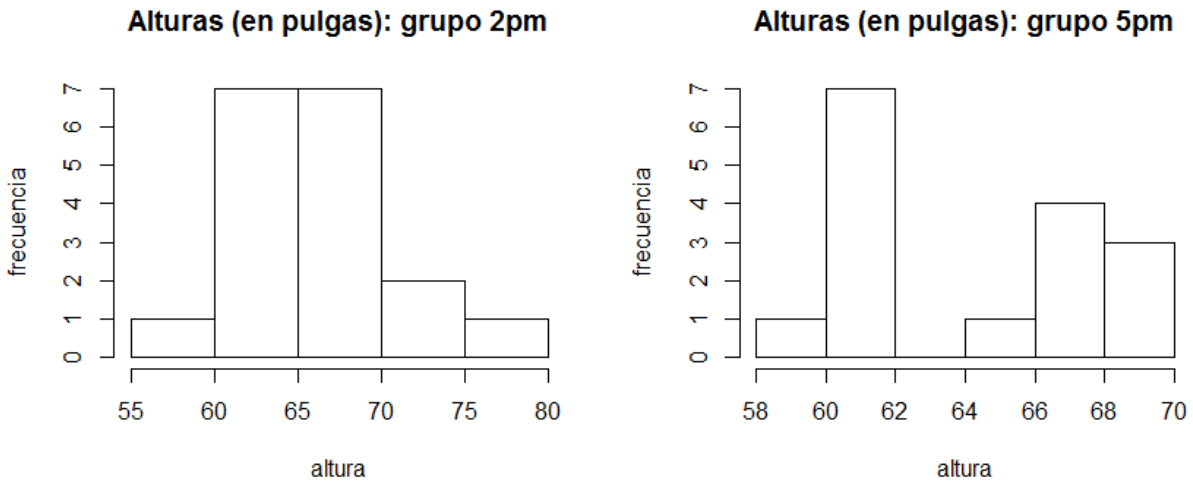
Hacemos 100000 simulaciones de dos muestras de datos normales con $n_1 = n_2 = 10$, $\mu_1 = 0$, $\mu_2 = 5$ y $\sigma_1^2 = \sigma_2^2 = 10$.

Con un valor nominal de $\alpha = 0{,}05$, aproximadamente 48 % de los intervalos contienen el verdadero valor, $\delta = -5$.

```
n1 <- 10; n2 <- 10
mu1 <- 0; mu2 <- 5
delta <- mu1-mu2
sigma21 <- 10; sigma22 <- 10
alpha <- 0.05
simuls <- 100000
nfuera <- 0
for (i in 1:simuls){
    y1 <- rnorm(n1,mu1,sqrt(sigma21))
    y2 <- rnorm(n2,mu2,sqrt(sigma22))
    ci <- t.test(y1,y2)$conf.int
    if (delta<ci[1]|delta>ci[2]){
        nfuera <- nfuera+1
    }
}
nfuera/simuls
```

Ejemplo

Datos de alturas de estudiantes de dos grupos (tomados del *Handbook of Biological Statistics por John Macdonald*).



El intervalo de un 95 % de confianza nominal es $(-1{,}072, 4{,}933)$.

El enfoque bayesiano

Supongamos que utilizamos distribuciones a priori "no-informativas": $f(\mu_i, \phi_i) \propto \frac{1}{\phi_i}$. Entonces, sabemos que dados los datos,

$$\frac{\mu_i - \bar{y}_i}{s_i/\sqrt{n_i}} \sim t_{n_i - 1} \quad \text{para } i = 1, 2.$$

Luego, tenemos:

$$\delta - (\bar{y}_i - \bar{y}_2) = \frac{s_1}{\sqrt{n_1}} T_{n_1 - 1} - \frac{s_2}{\sqrt{n_2}} T_{n_2 - 1},$$

donde T_a representa una variable t de Student con a grados de libertad y:

$$\frac{\delta - (\bar{y}_i - \bar{y}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{\frac{s_1}{\sqrt{n_1}} T_{n_1 - 1} - \frac{s_2}{\sqrt{n_2}} T_{n_2 - 1}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

El enfoque bayesiano

Supongamos que utilizamos distribuciones a priori "no-informativas": $f(\mu_i, \phi_i) \propto \frac{1}{\phi_i}$. Entonces, sabemos que dados los datos,

$$\frac{\mu_i - \bar{y}_i}{s_i/\sqrt{n_i}} \sim t_{n_i - 1} \quad \text{para } i = 1, 2.$$

Luego, tenemos:

$$\delta - (\bar{y}_i - \bar{y}_2) = \frac{s_1}{\sqrt{n_1}} T_{n_1 - 1} - \frac{s_2}{\sqrt{n_2}} T_{n_2 - 1},$$

donde T_a representa una variable t de Student con a grados de libertad y:

$$\begin{aligned} \frac{\delta - (\bar{y}_i - \bar{y}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} &= \frac{\frac{s_1}{\sqrt{n_1}} T_{n_1 - 1} - \frac{s_2}{\sqrt{n_2}} T_{n_2 - 1}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \\ &= T_{n_1 - 1} \sin\phi - T_{n_2 - 1} \cos\phi \quad \text{donde } \phi = \tan^{-1} \frac{s_1/\sqrt{n_1}}{s_2/\sqrt{n_2}}. \end{aligned}$$

La distribución de Behrens y Fisher

Se dice que una variable $\delta = T_{\nu_1} \sin\phi - T_{\nu_2} \cos\phi$ tiene una distribución de Behrens y Fisher con parámetros ν_1, ν_2, ϕ.

No se puede expresar la densidad en una forma sencilla, pero sí existen tablas de la distribución para aproximar a la función de distribución.

Table A.1 Percentage points of the Behrens-Fisher distribution

		$\psi = 0°$					$\psi = 15°$			
	ν_2	75%	90%	95%	97.5%		75%	90%	95%	97.5%
$\nu_1 = 6$	6	0.72	1.44	1.94	2.45	$\nu_1 = 6$	0.72	1.45	1.95	2.45
	8	0.71	1.40	1.86	2.31		0.72	1.41	1.87	2.32
	12	0.70	1.36	1.78	2.18		0.71	1.37	1.80	2.19
	24	0.68	1.32	1.71	2.06		0.69	1.34	1.73	2.09
	∞	0.67	1.28	1.65	1.96		0.68	1.30	1.67	2.00

El método Monte Carlo

En lugar de usar integración numérica, otra manera de aproximación de las características de una distribución se basa en generar una muestra de la distribución y usar los valores muestrales para aproximar.

Para aproximar una esperanza $E[g(Y)]$ para una variable Y, donde $V[g(Y)] < \infty$, se genera una muestra $y_1, ..., y_N$ y se estima con

$$\hat{g} = \frac{1}{N} \sum_{i=1}^{N} g(y_i).$$

- Justificado por la ley de los números grandes.
- Se puede estimar la precisión de la aproximación usando un intervalo de confianza: $\hat{g} \pm 1{,}96\text{dt}(g)/\sqrt{N}$.

Usando el método Monte Carlo para simular de una distribución de Behrens y Fisher

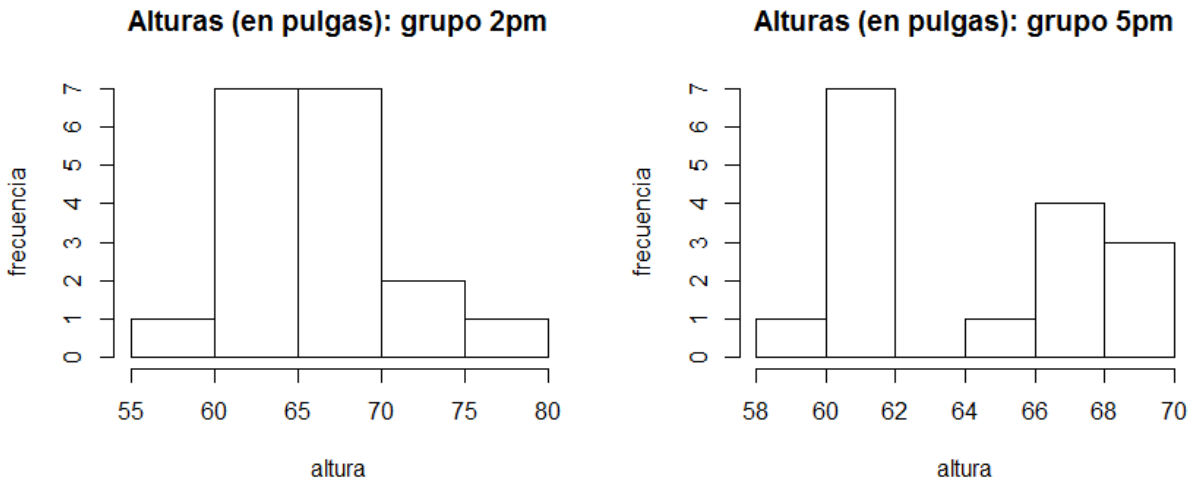
Utilizamos el siguiente algorítmo para generar una muestra de tamaño N de la distribución de Behrens con parámetros ν_1, ν_2, ϕ.

Para i en $1, ..., N$

- Generar $T_1^{(i)} \sim t_{\nu_1}$, $T_2^{(i)} \sim t_{\nu_2}$.
- Calcular $D_i = T_1^{(i)} \sin\theta - T_2^{(i)} \cos\phi$.

Se utilizan los percentiles de la muestra $d_1, ..., d_n$ para hallar un intervalo de credibilidad. Además, se puede aumentar la precisión de la estimación aumentando el tamaño de la muestra.

Ejemplo



- Intervalo frecuentista: $(-1{,}072, 4{,}933)$.
- Intervalo bayesiano "exacta": $(-1{,}175, 5{,}036)$.
- Intervalo MC con 100000 simulaciones: $(-1{,}175, 5{,}024)$.

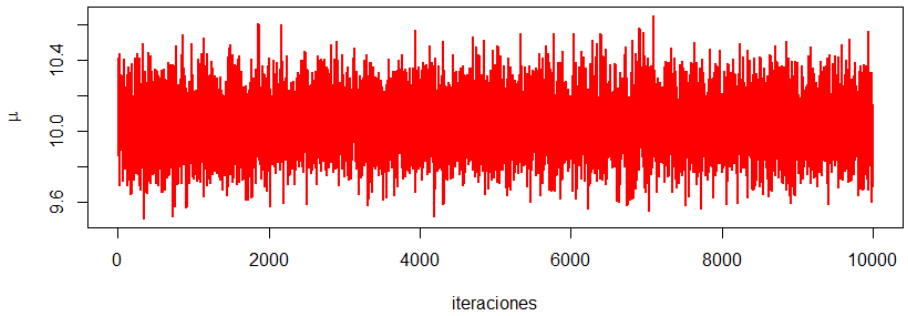
Una curiosidad sobre la solución bayesiana

Una curiosidad sobre la solución bayesiana

¡Fue introducido por Fisher como solución usando su método de inferencia fiducial!

Resumen y siguiente sesión

En esta sesión, vimos que en muchos, pero no todos, problemas de dos muestras, con una selección apropiada de la distribución a priori, resultados bayesianos coinciden numéricamente con los frecuentistas.



En la próxima sesión, comentamos una dificultad con la distribución a priori conjugada empleada en problemas normales y mostramos como hacer inferencia con una a priori más razonable mediante el muestreo de Gibbs.