Métodos Clásicos y Bayesianos para determinar el tamaño muestral en investigación clínica

Alejandro Sellés Pérez

13 de Junio de 2024

Universitat d'Alacant
Universidad de Alicante

Introducción

- El tamaño muestral es el **número de sujetos** que componen una muestra extraída de una población.

Introducción

- El tamaño muestral es el **número de sujetos** que componen una muestra extraída de una población.
- Se decide de manera **previa al ensayo**.

Introducción

- El tamaño muestral es el **número de sujetos** que componen una muestra extraída de una población.
- Se decide de manera **previa al ensayo**.
- Problemas económicos y de **salud**.

Introducción

- El tamaño muestral es el **número de sujetos** que componen una muestra extraída de una población.
- Se decide de manera **previa al ensayo**.
- Problemas económicos y de **salud**.

SALUD · Ensayo clínico fallido

90 voluntarios sanos habían iniciado la terapia con un fármaco que ha causado una muerte cerebral en Francia

MARÍA VALDERRAMA | París

ACTUALIZADO 15/01/2016 **16:46**

La ministra de Sanidad francesa, Marisol Touraine, ha aclarado que el medicamento que ha causado graves daños a seis voluntarios (uno de ellos en muerte cerebral) no contenía cannabis. Touraine ha apuntado al laboratorio portugués Bial Lab, como el fabricante del producto que se estaba investigando en Biotrial (un centro privado especializado en realizar ensayos clínicos pagados por la industria).

Comparación de medias

Para realizar una comparación de medias entre dos grupos, los autores frecuentistas proponen realizar **contrastes de hipótesis**.

$$\left.\begin{array}{lcccc} H_0 & : & \theta & \in & \Theta_0 \\ H_1 & : & \theta & \in & \Theta_1 \end{array}\right\} .$$

Comparación de medias

Para realizar una comparación de medias entre dos grupos, los autores frecuentistas proponen realizar **contrastes de hipótesis**.

$$\left.\begin{array}{lcccc} H_0 & : & \theta & \in & \Theta_0 \\ H_1 & : & \theta & \in & \Theta_1 \end{array}\right\} .$$

El objetivo es buscar la región de rechazo de H_0. Por ejemplo, en un test de una cola podemos encontrar una **región crítica** de la forma:

$$R_\alpha \quad = \quad]c_\alpha, \infty[\ .$$

Errores

Se deben controlar los 2 tipos de errores posibles a la hora de tomar una decisión:

	Aceptar H_0	**Rechazar** H_0
Verdadero H_0	Decisión correcta	Error tipo I
Verdadero H_1	Error tipo II	Decisión correcta

Cuadro: Tipos de errores en pruebas de hipótesis.

- α = Probabilidad de cometer un error Tipo I.
- β = Probabilidad de cometer un error Tipo II.

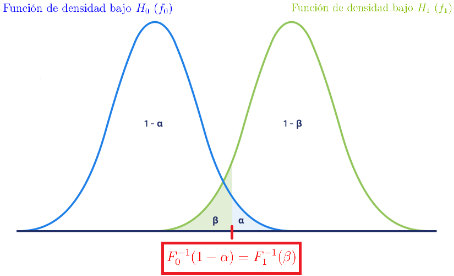
Método de potencia

F_0 y F_1 funciones de distribución bajo H_0 y H_1, obtenemos:

$$\left.\begin{array}{rcl} F_0(c_\alpha) & = & 1-\alpha \\ F_1(c_\alpha) & = & \beta \end{array}\right\} .$$

Despejando c_α, obtenemos:

$$F_1\left(F_0^{-1}(1-\alpha)\right) \quad = \quad \beta \ .$$



Ejemplo de comparación de medias

- **Objetivo**: Analizar si el consumo de leche agria es beneficioso a la hora de combatir la presión arterial.
- **Método**: Tomar 2 grupos (placebo/experimental) y estudiar las diferencias entre sus medias.

Ejemplo de comparación de medias

- **Objetivo**: Analizar si el consumo de leche agria es beneficioso a la hora de combatir la presión arterial.
- **Método**: Tomar 2 grupos (placebo/experimental) y estudiar las diferencias entre sus medias.

Suponemos que tenemos dos grupos, donde

$$x_{ij} \overset{\text{i.i.d.}}{\sim} N(\mu_i, \sigma^2); \quad j = 1, ..., n_i; \quad i = 1, 2$$

Superioridad o no inferioridad de la nueva media resolviendo:

$$\left.\begin{array}{llll} H_0: & \mu_2 - \mu_1 & \leq & \delta \\ H_1: & \mu_2 - \mu_1 & > & \delta \end{array}\right\},$$

donde δ es el margen de superioridad o no inferioridad.

Ejemplo de comparación de medias

Aplicando el método anterior obtenemos n_1 y n_2 resolviendo:

$$\left\{ \begin{array}{lcl} n_1 & = & kn_2 \\ n_2 & = & \dfrac{(z_{\alpha/2} + z_{\beta})^2 \sigma^2 \left(1 + \frac{1}{k}\right)}{((\mu_2 - \mu_1) - \delta)^2} \end{array} \right. . \qquad (1)$$

Ejemplo de comparación de medias

Aplicando el método anterior obtenemos n_1 y n_2 resolviendo:

$$\begin{cases} n_1 = kn_2 \\ n_2 = \dfrac{(z_{\alpha/2} + z_\beta)^2 \sigma^2 \left(1 + \frac{1}{k}\right)}{((\mu_2 - \mu_1) - \delta)^2} \end{cases} . \tag{1}$$

Analizamos si las personas que toman leche agria tienen de media $\delta = 2$mmHg menos de presión sistólica. Supongamos que tienen la misma media y $\sigma^2 = 25$:

Tamaño Muestral para diferentes valores de alpha y beta

	beta = 0.24	beta = 0.22	beta = 0.2	beta = 0.18	beta = 0.16	beta = 0.14	beta = 0.12	beta = 0.1
alpha = 0.01	136.32	141.78	147.65	154.02	161.00	168.76	177.52	187.66
alpha = 0.02	116.32	121.37	126.81	132.72	139.21	146.43	154.60	164.07
alpha = 0.03	104.61	109.40	114.57	120.19	126.37	133.25	141.05	150.11
alpha = 0.04	96.29	100.89	105.85	111.26	117.21	123.84	131.37	140.11
alpha = 0.05	89.83	94.28	99.08	104.31	110.08	116.51	123.82	132.31
alpha = 0.06	84.56	88.87	93.54	98.62	104.23	110.49	117.61	125.90
alpha = 0.07	80.10	84.30	88.84	93.80	99.27	105.39	112.35	120.45
alpha = 0.08	76.23	80.34	84.77	89.62	94.97	100.96	107.77	115.70
alpha = 0.09	72.83	76.84	81.18	85.93	91.17	97.03	103.71	111.50
alpha = 0.1	69.78	73.71	77.96	82.62	87.76	93.51	100.08	107.73

Figura: Tamaños muestrales con $n_1 = n_2$, $\sigma^2 = 25$ y $\delta = 2$

Estimación de una media

Una vez comprobada esta diferencia de medias, otro estudio de interés puede ser **estimar la media** de la presión sistólica del grupo que consumirá leche agria.

Estimación de una media

Una vez comprobada esta diferencia de medias, otro estudio de interés puede ser **estimar la media** de la presión sistólica del grupo que consumirá leche agria.

- Los investigadores frecuentistas proponen obtener un **intervalo de confianza**.

Estimación de una media

Una vez comprobada esta diferencia de medias, otro estudio de interés puede ser **estimar la media** de la presión sistólica del grupo que consumirá leche agria.

- Los investigadores frecuentistas proponen obtener un **intervalo de confianza**.
- Para obtener el tamaño muestral, la idea es **despejar el valor de** n **en función de la longitud** del intervalo.

Ejemplo de estimación de una media

Supongamos

$$x_i \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2) \quad i = 1, ..., n \ .$$

Obtenemos el siguiente intervalo de confianza:

$$IC_{1-\alpha}(\mu) \quad = \quad \left]\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \ , \ \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right[\ .$$

Ejemplo de estimación de una media

Supongamos

$$x_i \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2) \quad i = 1, ..., n \ .$$

Obtenemos el siguiente intervalo de confianza:

$$IC_{1-\alpha}(\mu) \quad = \quad \left]\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \ , \ \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right[\ .$$

Tomamos la longitud:

$$l \quad = \quad \left(\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) - \left(\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) \quad = \quad 2z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \ .$$

Ejemplo de estimación de una media

Supongamos

$$x_i \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2) \quad i = 1, ..., n \ .$$

Obtenemos el siguiente intervalo de confianza:

$$IC_{1-\alpha}(\mu) \quad = \quad \left] \bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \ , \ \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \right[\ .$$

Tomamos la longitud:

$$l \quad = \quad \left(\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \right) - \left(\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \right) \quad = \quad 2z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \ .$$

Finalmente, despejando n:

$$n \quad = \quad \frac{4z_{\alpha/2}^2\sigma^2}{l^2} \ . \qquad (2)$$

Ejemplo de estimación de una media

Tamaño Muestral para Estimación de una Media

	alpha = 0.025	alpha = 0.05	alpha = 0.075	alpha = 0.1
l 1 = 1	503.00	385.00	318.00	271.00
l 2 = 2	126.00	97.00	80.00	68.00
l 3 = 3	56.00	43.00	36.00	31.00

Figura: Tamaño muestral para la estimación de una media para distintos α y l con $\sigma^2 = 25$

σ^2 desconocido

Si el valor de σ^2 es desconocido, se puede proceder de la misma manera que antes pero tomando s^2.

σ^2 desconocido

Si el valor de σ^2 es desconocido, se puede proceder de la misma manera que antes pero tomando s^2.

Probemas:

- El valor de s^2 **será desconocido** también.
- Obtenemos *t de student* con $n-1$ grados de libertad y **no podemos despejar** n.

σ^2 desconocido

Si el valor de σ^2 es desconocido, se puede proceder de la misma manera que antes pero tomando s^2.

Probemas:

- El valor de s^2 **será desconocido** también.
- Obtenemos *t de student* con $n-1$ grados de libertad y **no podemos despejar** n.

Solución: Estimar el valor de σ^2 y trabajar como si fuera conocido.

Motivación de la visión bayesiana

Con esta manera de proceder podemos encontrar distintos problemas:

- Los métodos frecuentistas están pensados para **trabajar de manera posterior** a la obtención de los datos. Mientras que el valor de σ^2 será obtenido con información de estudios previos o predicciones de los investigadores.
- Las fórmulas son sensibles al valor de σ^2 escogido.
- Puede haber **información a priori** interesante con respecto al parámetro a estudiar que no se incluye.

Motivación de la visión bayesiana

Con esta manera de proceder podemos encontrar distintos problemas:

- Los métodos frecuentistas están pensados para **trabajar de manera posterior** a la obtención de los datos. Mientras que el valor de σ^2 será obtenido con información de estudios previos o predicciones de los investigadores.
- Las fórmulas son sensibles al valor de σ^2 escogido.
- Puede haber **información a priori** interesante con respecto al parámetro a estudiar que no se incluye.

Por ello, parece interesante abordar el problema del tamaño muestral desde el punto de vista de la estadística bayesiana.

Teorema de Bayes

Tomamos la distribución marginal de x haciendo:

$$f(x) = \int_{\Theta} f(x|\theta)f(\theta)d\theta \ .$$

Teorema de Bayes

Sean x y θ variables aleatorias con funciones de densidad $f(x|\theta)$ y $f(\theta)$. Tenemos

$$f(\theta|x) = \frac{f(x|\theta)f(\theta)}{f(x)} \propto f(x|\theta)f(\theta) \ ,$$

donde

- *x: vector o matriz de datos.*
- *θ: parámetro desconocido.*
- *$f(x_1, ..., x_n|\theta)$: verosimilitud de los datos dado θ.*
- *$f(\theta)$: distribución a priori del parámetro θ.*

Diferencias entre teorias

Idea: Partir de una distribución a priori $f(\theta)$ y aplicando el teorema vamos a obtener una actualización $f(\theta|x)$ con la que trabajaremos.

Diferencias entre teorias

Idea: Partir de una distribución a priori $f(\theta)$ y aplicando el teorema vamos a obtener una actualización $f(\theta|x)$ con la que trabajaremos.

Hay diferencias significativas entre ambas teorías a la hora de realizar la estimación de un parámetro:

	Teoría Frecuentista	**Teoría Bayesiana**
Parámetro de interés	Es una constante desconocida	Es una variable aleatoria
Distribución a priori	No se tiene en cuenta	Existe y es necesaria para los cálculos
Distribución a posteriori	No existe	Existe y se deriva
Tipo de razonamiento	Inductivo	Deductivo
Estimación de un parámetro	Intervalo de confianza	Intervalo de credibilidad posterior

Cuadro: Diferencias entre teorías.

Diferencias entre teorias

Idea: Partir de una distribución a priori $f(\theta)$ y aplicando el teorema vamos a obtener una actualización $f(\theta|x)$ con la que trabajaremos.

Hay diferencias significativas entre ambas teorías a la hora de realizar la estimación de un parámetro:

	Teoría Frecuentista	**Teoría Bayesiana**
Parámetro de interés	Es una constante desconocida	Es una variable aleatoria
Distribución a priori	No se tiene en cuenta	Existe y es necesaria para los cálculos
Distribución a posteriori	No existe	Existe y se deriva
Tipo de razonamiento	Inductivo	Deductivo
Estimación de un parámetro	Intervalo de confianza	Intervalo de credibilidad posterior

Cuadro: Diferencias entre teorías.

Vamos a denominar con el nombre de ***precisión*** a τ, donde

$$\tau = \frac{1}{\sigma^2} .$$

3 criterios bayesianos para la estimación de una media

- **Average Coverage Criterion (ACC):** Fijando l, el objetivo será que el promedio de las probabilidades de que el parámetro se encuentre dentro de los intervalos sea al menos $1 - \alpha$.

3 criterios bayesianos para la estimación de una media

- **Average Coverage Criterion (ACC):** Fijando *l*, el objetivo será que el promedio de las probabilidades de que el parámetro se encuentre dentro de los intervalos sea al menos $1 - \alpha$.
- **Average Length Criterion (ALC):** Fijando α, el objetivo será seleccionar el tamaño muestral para el cual los intervalos tienen una longitud promedio *l*, que se tratará de minimizar, ya que cuánto más cortos sean estos intervalos, más precisas serán las estimaciones.

3 criterios bayesianos para la estimación de una media

- **Average Coverage Criterion (ACC):** Fijando l, el objetivo será que el promedio de las probabilidades de que el parámetro se encuentre dentro de los intervalos sea al menos $1 - \alpha$.
- **Average Length Criterion (ALC):** Fijando α, el objetivo será seleccionar el tamaño muestral para el cual los intervalos tienen una longitud promedio l, que se tratará de minimizar, ya que cuánto más cortos sean estos intervalos, más precisas serán las estimaciones.
- **Worst Outcome Criterion (WOC):** Controlamos los valores de α y l simultáneamente. Tomaremos un % del espacio muestral donde los intervalos obtenidos a partir de estos datos no tengan una longitud mayor a l ni una probabilidad inferior a $1 - \alpha$ de encontrar el parámetro.

Estimación de una media con σ^2 conocida

Suponemos

$$x_i \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2) \quad i = 1, ..., n \quad .$$

Y, a priori:

$$\mu \sim N(\mu_0, n_0 \tau) \quad .$$

Aplicando el Teorema de Bayes llegamos a que la distribución a posteriori cumple:

$$\mu | x \sim N(\mu_n, \tau_n) \quad ,$$

donde

$$\begin{aligned} \tau_n &= (n + n_0)\tau \quad , \\ \mu_n &= \frac{n_0 \mu_0 + n \bar{x}}{n_0 + n} \quad . \end{aligned}$$

Desarrollando los 3 criterios obtenemos la misma solución:

$$n \geq \frac{4 z^2_{1-\alpha/2}}{\tau l^2} - n_0 \quad . \tag{3}$$

Estimación de una media con σ^2 desconocida

Suponemos a priori:

$$\tau \sim \Gamma(v, \rho) \ ,$$

$$\mu|\tau \sim N(\mu_0, n_0\tau) \ .$$

Esta vez, aplicando el Teorema de Bayes y después obteniendo la distribución marginal de μ:

$$\mu|x \quad \sim \quad t_{2v+n}\sqrt{\frac{\rho_n}{(n+n_0)\left(v+\frac{n}{s}\right)}} + \mu_n \ ,$$

donde

$$\mu \quad = \quad \frac{n_0\mu_0 + n\bar{x}}{n+n_0} \ ,$$

$$\rho_n \quad = \quad \rho + \frac{1}{2}ns^2 + \frac{nn_0}{2(n+n_0)}(\bar{x}-\mu_0)^2 \ .$$

Estimación de una media con σ^2 desconocida. Criterios

ACC

$$n = \frac{4\rho}{vl^2} t^2_{2v,1-\alpha/2} - n_0 \tag{4}$$

ALC

$$2t_{n+2v,1-\alpha/2}\sqrt{\frac{2\rho}{(n+2v)(n+n_0)}}\frac{\Gamma\left(\frac{n+2v}{2}\right)\Gamma\left(\frac{2v-1}{2}\right)}{\Gamma\left(\frac{n+2v-1}{2}\right)\Gamma(v)} \leq l \tag{5}$$

WOC

$$\frac{l^2(n+2v)(n+n_0)}{8\rho\left(1+\left(\frac{n}{2v}\right)F_{n,2v,1-w}\right)} \leq t^2_{n+2v,1-\alpha/2} \tag{6}$$

Comparación de las fórmulas para la estimación de una media

Cálculo del Tamaño Muestral para estimación de una media

v:

2

rho:

50

longitud del intervalo:

1

n0:

10

Alpha:

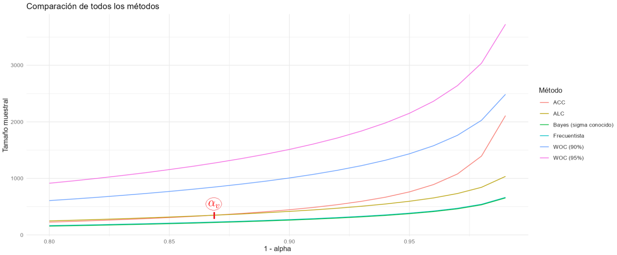
0 — 0.05 — 0.2

0 0.02 0.04 0.06 0.08 0.1 0.12 0.14 0.16 0.18 0.2

Calcular

Tamaño Muestral (Alpha: 0.05)

	Criterio	n
1	Frecuentista	385.00
2	Bayes (sigma conocido)	375.00
3	ACC	761.00
4	ALC	595.00
5	WOC(90%)	1435.00
6	WOC(95%)	2152.00



Conclusiones

- Para varianza conocida las fórmulas obtenidas son **similares**.

Conclusiones

- Para varianza conocida las fórmulas obtenidas son **similares**.
- Para varianza desconocida es más natural la manera de trabajar desde el punto de vista bayesiano, ya que se basa en la **distribución que seguirá este parámetro, en lugar de tomar una estimación de él**. Esto puede ser más sencillo de deducir a priori.

Conclusiones

- Para varianza conocida las fórmulas obtenidas son **similares**.
- Para varianza desconocida es más natural la manera de trabajar desde el punto de vista bayesiano, ya que se basa en la **distribución que seguirá este parámetro, en lugar de tomar una estimación de él**. Esto puede ser más sencillo de deducir a priori.
- **Analizar cada ensayo de manera independiente**. Debido a que, por ejemplo, si se conoce mucha información previa sobre la muestra, quizás los métodos bayesianos aporten un número de personas excesivamente elevado.

Referencias

[1] Adcock, C.J. (1988) A Bayesian Approach to Calculating Sample Sizes. *Journal of the Royal Statistical Society. Series D (The Statistician)*. 37(4/5): 433-439

[2] Adcock, C.J. (1997) Sample Size Determination: A Review. *Journal of the Royal Statistical Society. Series D (The Statistician)*. 46(2): 261-283.

[3] Cao, J., Lee, J.J., and Albert, S. (2009) Comparison of Bayesian sample size criteria: ACC, ALC, and WOC. *J Stat Plan Inference*. 139(12): 4111-4122.

[4] Chow, S.C., Shao, J., Wang, H., y Lokhnygina, Y. (2020). *Sample Size Calculations in Clinical Research*. CRC Press.

[5] Joseph, L., y Belisle, P. (1997) Bayesian sample size determination for normal means and differences between normal means. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 46(2): 209-226.