FACULTAD DE MATEMÁTICAS
PONTIFICIA UNIVERSIDAD
CATÓLICA DE CHILE

EYP1113 - Probabilidad y Estadística

Taller R - 09: Inferencia Estadística

Ricardo Aravena C. - Cristian Capetillo C. - Ingrid Guevara R.
Ricardo Olea O. - Bladimir Morales T., Daniel Saavedra M.

Facultad de Matemáticas
Departamento de Estadística
Pontificia Universidad Católica de Chile

Segundo Semestre 2024

Contenidos

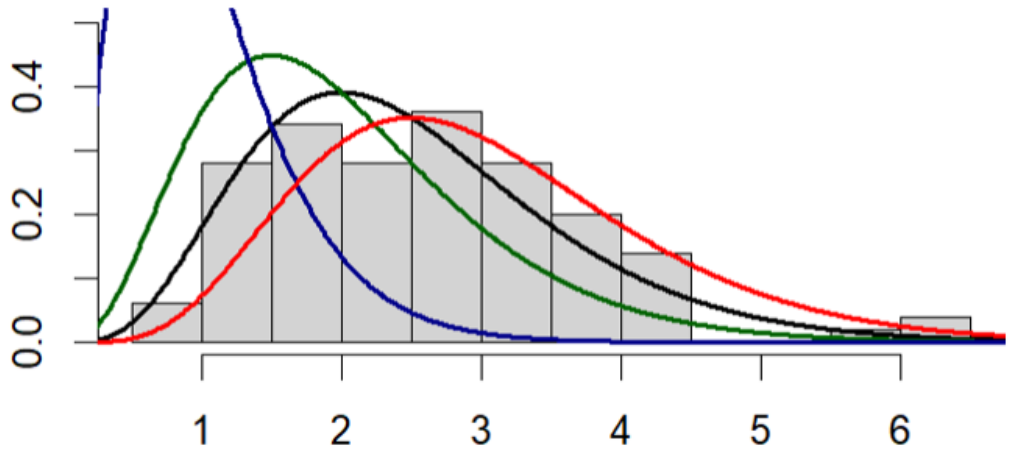
Motivación

- En el análisis de datos, nos enfrentamos con el problema de **aprender sobre la población** teniendo acceso a sólo una parte de ésta (la muestra).
- La inferencia estadística es la aplicación de un conjunto de técnicas que permiten **extrapolar los resultados** basados en la muestra hacia la población.
- Si creemos que nuestros datos provienen de cierta distribución parametrizada por θ, por ejemplo la distribución $\mathcal{N}(\theta, 1)$, entonces podemos utilizar los datos para "adivinar" el valor real de θ.
- La inferencia estadística puede dividirse en 3 tópicos: estimación puntual, estimación intervalar y pruebas de hipótesis.

Estimación puntual

Sea una muestra aleatoria $X_1, ..., X_n \overset{iid}{\sim} F_\theta$, con $\theta \in \Theta$. El proceso de estimación puntual, o simplemente estimación, tiene por objetivo hayar un **valor específico** en Θ, denotado por $\widehat{\theta}$, utilizando los **datos observados**.

Manualmente, podríamos ajustar curvas sobre el histograma y hacernos una idea de los valores reales de los parámetros del modelo.



En la práctica, debemos optar por un método que nos garantice estar lo más cerca posible del valor real de θ.

Estimador de momentos

El estimador de momentos es en teoría simple. Tenemos los **momentos muestrales** (promedio de los datos, promedios de los datos al cuadrado, promedio de los datos al cubo, etc.) y los igualamos a los **momentos teóricos** (Esperanza de la variable, esperanza de la variable al cuadrado, esperanza de la variable al cubo, etc.).

Esto nos genera un **sistema de ecuaciones**, que al momento de resolver nos dará valores posibles para los parámetros (lo que serían las estimaciones de éstos).

Estimador de momentos

El estimador de momentos es en teoría simple. Tenemos los **momentos muestrales** (promedio de los datos, promedios de los datos al cuadrado, promedio de los datos al cubo, etc.) y los igualamos a los **momentos teóricos** (Esperanza de la variable, esperanza de la variable al cuadrado, esperanza de la variable al cubo, etc.).

Esto nos genera un **sistema de ecuaciones**, que al momento de resolver nos dará valores posibles para los parámetros (lo que serían las estimaciones de éstos).

Si $X_1, ..., X_n \overset{iid}{\sim} \mathcal{G}amma(\alpha, \beta)$, entonces

$$\mathbb{E}[X] = \frac{\alpha}{\beta} \Rightarrow \overline{X} = \frac{\widehat{\alpha}}{\widehat{\beta}},$$

$$\mathbb{V}[X] = \frac{\alpha}{\beta^2} \Rightarrow \overline{X^2} - \overline{X}^2 = \frac{\widehat{\alpha}}{\widehat{\beta}^2}.$$

Estimador de máximo verosímil

Otra opción para encontrar un estimador para los parámetros es el famoso estimador **máximo** verosímil.

Estimador de máximo verosímil

Otra opción para encontrar un estimador para los parámetros es el famoso estimador **máximo** verosímil.

Si la densidad o cuantía de los datos está dada por $f_\theta(x)$, la **función de verosimilitud** para θ basado en las observaciones $\boldsymbol{x} = (x_1, ..., x_n)$ se define mediante

$$L(\theta|\boldsymbol{x}) = \prod_{i=1}^{n} f_\theta(x_i).$$

Estimador de máximo verosímil

Otra opción para encontrar un estimador para los parámetros es el famoso estimador **máximo** verosímil.

Si la densidad o cuantía de los datos está dada por $f_\theta(x)$, la **función de verosimilitud** para θ basado en las observaciones $\boldsymbol{x} = (x_1, ..., x_n)$ se define mediante

$$L(\theta|\boldsymbol{x}) = \prod_{i=1}^{n} f_\theta(x_i).$$

Dado que la **función log** es monótona, el máximo de la verosimilitud y el máximo de la log-verosimilitud es el mismo. Luego, se hace más simple maximizar esta última función, que denotamos por $l(\theta|\boldsymbol{x})$ o simplemente $l(\theta)$.

Típicamente se obtienen las **derivadas de la log-verosimilitud**, **se igualan a 0**, y se resuelve el **sistema de ecuaciones** resultante.

Estimador de máximo verosímil

- $\mathcal{B}inomial(n, p)$:

$$l(p) = \log \binom{n}{x} + x \log(p) + (n - x) \log(1 - p).$$

- $\mathcal{P}oisson(\lambda)$:

$$l(\lambda) = \log(\lambda) \sum_{i=1}^{n} x_i - n\lambda - \sum_{i=1}^{n} \log(x_i!).$$

- $\mathcal{N}(\mu, \sigma^2)$:

$$l(\mu, \sigma^2) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{n}{2\sigma^2}(\overline{x^2} - 2\mu\bar{x} + \mu^2).$$

- $\mathcal{G}amma(\alpha, \beta)$:

$$l(\alpha, \beta) = n\alpha \log(\beta) - \log(\Gamma(\alpha)) + (\alpha - 1) \sum_{i=1}^{n} \log(x_i) - \beta \sum_{i=1}^{n} x_i.$$

Estimador de máximo verosímil

La **maximización** puede tener algunas complicaciones (por ejemplo, con la distribución Gamma). Pero esto puede ser resuelto **computacionalmente**. Una posible función que podemos utilizar es optim().

optim(par, fn)

En *par* entregamos un vector con valores iniciales, mientras que en *fn* entregamos la función a optimizar.

Estimador de máximo verosímil: Ejemplo

1. Genere una muestra de $n = 50$ datos provenientes de la distribución $\mathcal{P}oisson(\lambda)$, con $\lambda = 2$.

2. Programe la función log-verosimilitud de λ y grafíquela.

3. Obtenga la estimación máximo verosímil para λ. ¿Por cuánto se diferencian el valor real y la estimación?

4. Realice lo anterior para las distribuciones $\mathcal{N}(\mu, 1)$, $\mathcal{G}amma(1, \beta)$ y $\mathcal{B}ernoulli(p)$ con $\mu = 5$, $\beta = 1/3$, $p = 0.3$.

Estimación intervalar

Claramente, cuando estimamos una característica poblacional será **prácticamente imposible obtener el valor real** de ésta. Sin embargo, entendemos que estaremos relativamente cerca.

Para medir esta **incertidumbre**, típicamente se opta por calcular los denominados **intervalos de confianza**. Estos intervalos dan una idea de dónde podría estar realmente el parámetro poblacional, aunque su interpretación es algo más complicada que esto.

Ayudándonos del **Teorema del Límite Central**, podemos calcular intervalos de confianza para la media de una población. En general, cualquier intervalo de confianza construido a partir de un **estimador asintóticamente normal** puede construirse de manera análoga.

Estimación intervalar: Intervalo para la media

Si $X_1, ..., X_n$ es una muestra i.i.d. de una población con media μ y varianza σ^2, entonces

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \overset{aprox}{\sim} \mathcal{N}(0,1).$$

Luego, este resultado nos permite calcular un intervalo basado en la distribución normal estándar. Así, un intervalo de $100(1-\alpha)\%$ de confianza para μ está dado por

$$\left[\overline{X} + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \; ; \overline{X} + z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right].$$

Valores grandes de α implican más confianza y un intervalo que abarca más valores. Por el contrario, menos confianza nos entregará intervalos más acotados.

Estimación intervalar: Intervalo para la media

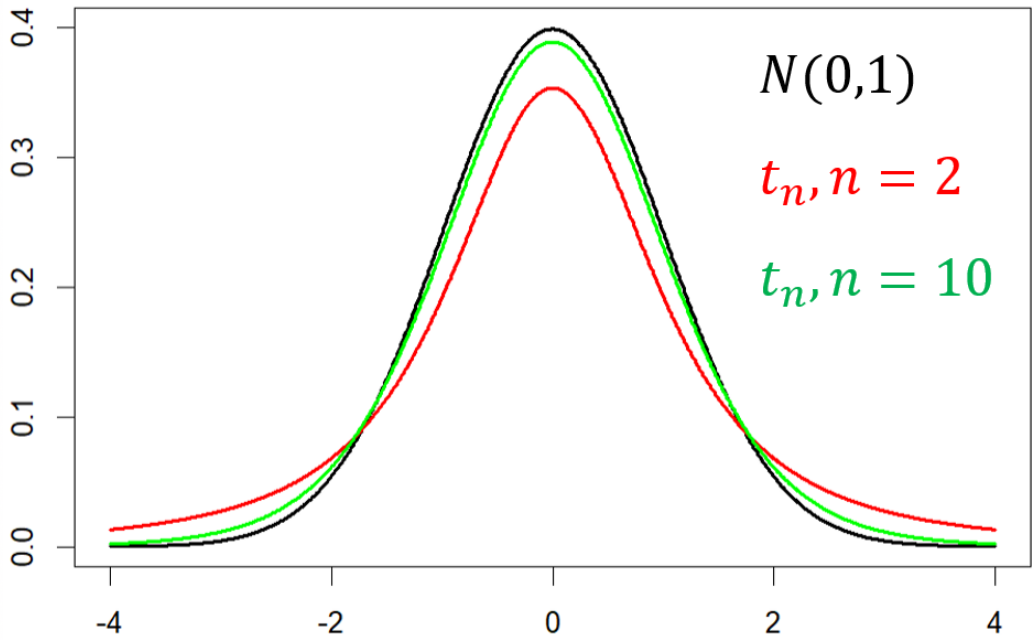
En general, σ^2 es desconocido y reemplazado por la varianza muestral S_x^2. Este cambio sugiere agregar **más incertidumbre** al intervalo de confianza, utilizando la distribución t_{n-1}.

Finalmente, un intervalo de $100(1-\alpha)\%$ de confianza para la media μ estará dado por

$$\left[\overline{X} + t_{n-1,\frac{\alpha}{2}} \cdot \frac{S_x}{\sqrt{n}} \; ; \overline{X} + t_{n-1,1-\frac{\alpha}{2}} \cdot \frac{S_x}{\sqrt{n}}\right].$$

Estimación intervalar: Intervalo para la media

La corrección anterior puede obviarse si $n \geq 30$, puesto que t_{n-1} tiende a la distribución normal estándar cuando $n \to \infty$.



Estimación intervalar: Ejemplo

Se hacen ocho mediciones independientes del diámetro de pistones. Las mediciones (en pulgadas) son 3.236, 3.223, 3.242, 3.244, 3.228, 3.253, 3.253 y 3.230.

1. Relice un diagrama de cajas para estos datos.
2. ¿Se debe utilizar la distribución t para encontrar un intervalo de confianza para el diámetro medio de estos pistones?
3. Calcule un intervalo de confianza para el diametro medio de los pistones utilizando la distribución t y utilizando la distribución normal. ¿Qué diferencia hay entres ambos?

Prueba de hipótesis

Es común que la recopilación de datos tenga por objetivo **confirmar o refutar una teoría**. ¿Será que el salario medio de las mujeres es menor que el de los hombres?, ¿será verdad que cierto medicamento tiene efecto positivo en la evolución de una enfermedad?

En este taller, nuevamente nos basamos en el **TLC** para realizar este procedimiento, aunque esto puede ser extendido.

Prueba de hipótesis

Sea una muestra aleatoria $X_1, ..., X_n \overset{iid}{\sim} F_\theta$. Una **hipótesis estadística** es una afirmación sobre los parámetros de la distribución. Por ejemplo,

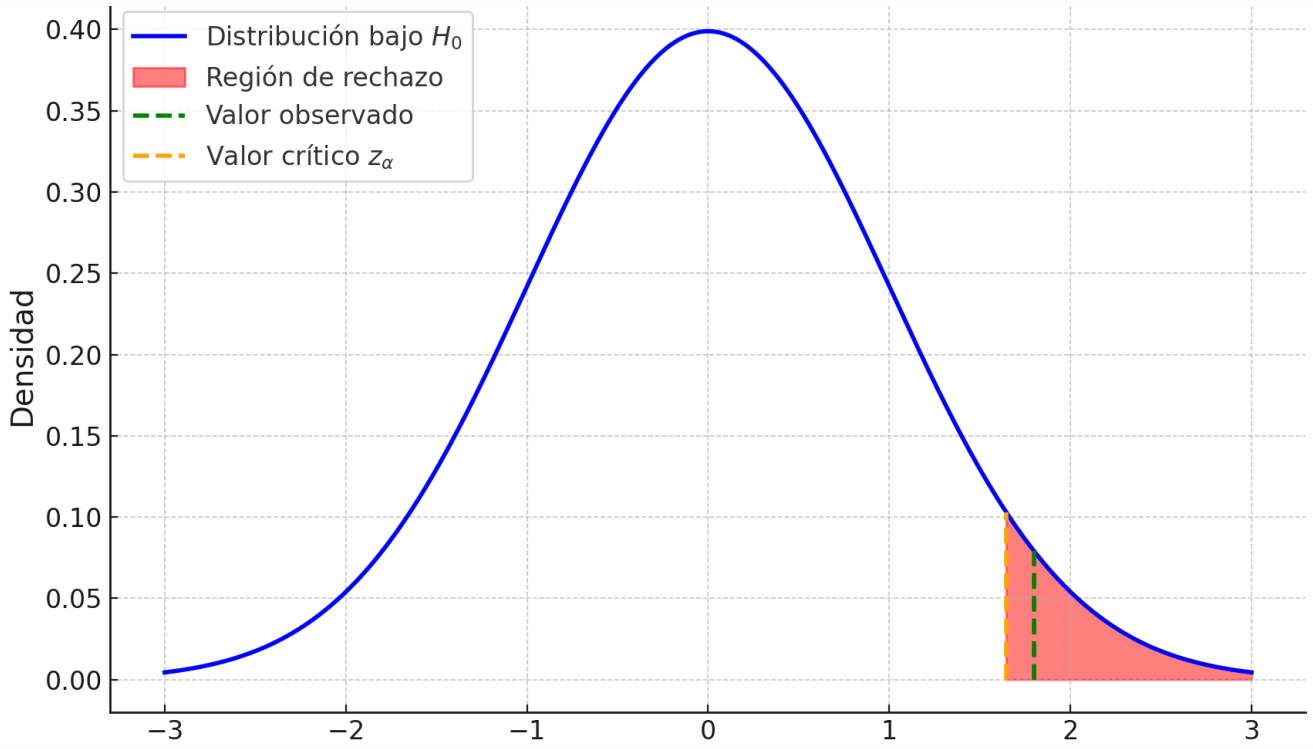
$$H_0 : \theta \in \Theta_0 \quad vs \quad H_1 : \theta \in \Theta_1,$$

con $\Theta_0 \cap \Theta_1 = \emptyset$.

Todo test se basa en un **estadístico pivote**, esto es, una función de la muestra y los parámetros testeados ($g(\boldsymbol{X}, \theta)$) cuya distribución bajo H_0 está totalmente determinada.

Finalmente, calculamos el estadístico pivote bajo H_0 y evaluamos **qué tan raro es** según su distribución. Si lo obtenido es raro, entonces la hipótesis es poco creíble y por ende la rechazamos.

Prueba de hipótesis



Prueba de hipótesis

Para automatizar la decisión, se establece una **región de rechazo** con un nivel de significancia α. Menor α, mayor probabilidad de **rechazar** correctamente H_0. No obstante, la probabilidad de **no rechazar** H_0 cuando esta es falsa tambien es alta, a no ser de que tengamos una basta cantidad de datos. Típicamente se utiliza un valor de $\alpha = 0.05$, pero valores como 0.1 o 0.01 también son comúnes.

Pruebas de hipótesis

Cuando no conocemos la distribución de la muestra, el TLC es siempre una buena opción. En particular,

- $\dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}} \overset{aprox}{\sim} \mathcal{N}(0,1)$.
- $\dfrac{\overline{X} - \mu}{S_x/\sqrt{n}} \overset{aprox}{\sim} t_{n-1}$.
- Si tenemos otra muestra $Y_1, ..., Y_n$ independiente de la anterior pero con igual varianza, entonces se sostiene que

$$\frac{(\overline{X} - \mu_x) - (\overline{Y} - \mu_y)}{S_{pool}\sqrt{2/n}} \overset{aprox}{\sim} t_{2n-2},$$

donde $S^2_{pool} = \frac{1}{2}(S^2_x + S^2_y)$.

Prueba de hipótesis: t-test

El **t-test** es una prueba de hipótesis sobre la media de una población, o la comparación de medias. Para este último caso, el t-test puede ser calculado suponiendo varianzas iguales, varianzas distintas, independencia entre poblaciones o muestras pareadas.

Cada uno de estos supone un estadístico distinto pero con una distribución t-student (de ahí el nombre de la prueba estadística).

La manera de implementarlo en `R` es a través de la función *t.test* (Ver ayuda de `R`).

Prueba de hipótesis: Ejemplo

1. Simule $n_1 = 100$ datos $\mathcal{G}amma(2, 2)$ y $n_2 = 80$ datos $\mathcal{G}amma(4, 3)$.
2. Grafique el histrograma del grupo 1 y superponga aquel del grupo 2. ¿Aprecia alguna diferencia en localización?, ¿y en la dispersión?
3. Realice el t-test verificando si la media de los grupos es igual o difieren.
4. ¿Rechaza la hipótesis de que la media de los grupos es la misma?