Inferencia

Estadística Computacional

Juan Zamora Osorio

juan.zamora@pucv.cl

Instituto de Estadística
Pontificia Universidad Católica de Valparaíso

6 de noviembre de 2023

Inferencia

Hemos aprendido sobre...

- Describir datos.
- Probabilidades.
- Variables aleatorias.

Objetivo

- Contrastar datos reales con modelos basados en probabilidades.

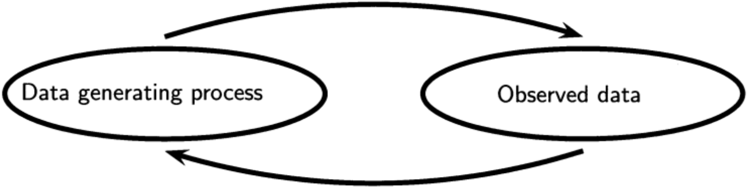
¿Qué necesitamos?

- Inferencia estadística.
- Contrastar hipótesis.

Recordar

Probabilidades

- ¿Dado un proceso que genera datos, cuáles son las propiedades que observaremos?



Inferencia estadística

- ¿Dadas las observaciones, qué podemos decir sobre el proceso que genera los datos?

Recuerdo

Media muestral

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{\sum_{i=1}^{n} x_i}{n}.$$

- Busca estimar la *media* de la población, denotada como μ.

Ejemplo: temperaturas máximas durante enero

- 22, 24, 21, 22, 25, 26, 25, 24, 23, 25, 25, 26, 27, 25, 26, 25, 26, 27, 27, 28, 29, 29, 29, 28, 30, 29, 30, 31, 30, 28, 29.
- $\bar{x} = \frac{22+24+21+\cdots+28+29}{31} = 26{,}48$ ºC.

Muestra

Resultados obtenidos de experimentos aleatorios

- Cantidad es finita.
- Generalmente se supone independencia.
- Generalmente se supone misma distribución en cada experimento.

Estadístico

- Función calculada a partir de la muestra.

Ejemplo: 10 lanzamientos de un dado

- Cantidad de datos: 10.
- Cada lanzamiento independiente del anterior.
- Cada lanzamiento posee la misma distribución: multinomial.

Muestra

Ejemplo: 10 tomas de temperatura a medio día en días distintos

- Cantidad de datos: 10.
- ¿Cada medición es independiente de la anterior?
- ¿Cada medición posee misma distribución?

Ejemplo: 10 nombres de estudiantes del curso

- Cantidad de datos: 10.
- ¿Cada nombre es independiente del anterior?
- ¿Cada nombre posee la misma distribución?

- Queremos que la muestra sea representativa.

Técnicas de muestreo

Sesgo

- Una técnica es sesgada si el estadístico calculado con la muestra obtenida es mayor o menor, en promedio, que el parámetro estimado.

Sesgo de selección

- La manera en que se construye la muestra introduce sesgo.

Sesgo de respuesta

- La técnica para obtener la respuesta introduce sesgo.

Sesgo de selección

Ejemplo – tamaño

- Los pacientes que pasan más días en un hospital son más propensos a ser elegidos para una muestra.

Ejemplo – respuesta voluntaria

- Las opiniones recolectadas por llamados a un programa de televisión sobre representan a quienes les importa el asunto y no representan a quienes no les interesa.

Ejemplo – conveniencia

- Selecciono a mis amigo/as como muestra para estudiar la opinión de la población.

Sesgo de selección

Ejemplo – juicio experto

- Se intenta recolectar un grupo de personas con ciertas características: tantos hombres, tantas mujeres, tantos sobre 40, tantos empleados, etc. creyendo que se mejora representatividad, pero se agrega sesgo.

Ejemplo – marco

- Se selecciona a partir de una lista que debería corresponder a la población.

Sesgo de respuesta

No hay respuesta

- Alguien que se niega a participar en una encuesta podría ser diferente a los demás.

Respuesta incorrecta o error de medición

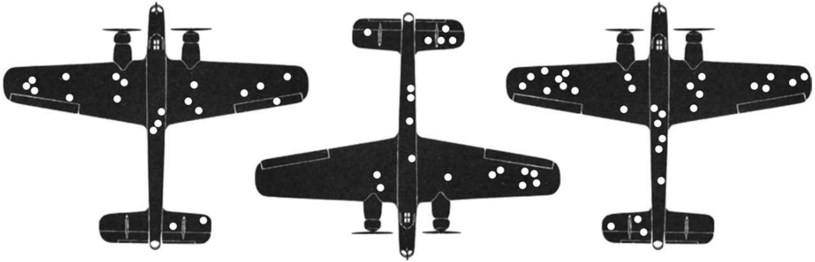
- Mentira intencional.
- Memoria imprecisa.
- Medición imprecisa.
- Ejemplos:
  - Muchas personas no admiten ver un programa de televisión.
  - Pacientes que dicen que siguen indicaciones médicas.
  - ¿Cuánto tiempo pasan en el celular al día?

Sesgo de respuesta

Cuestionario

- La respuesta depende de la pregunta, del tono de voz del entrevistador, el orden de las preguntas, etc.

Ejemplo: muestra de localización de daños en bombarderos



Muestra aleatoria

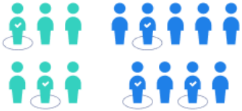
- No introduce sesgo.

Aleatoria simple

- Todas las observaciones son igual de probables.

Aleatoria estratificada

- Se divide la población en grupos que no se traslapan.



Ejemplo: Plaza Pública CADEM

Metodología

_Técnica
Encuestas Telefónicas aplicadas a través de sistema Cati a celulares de prepago y postpago.

_Universo
Hombres y mujeres de 18 años o más, habitantes en las 16 regiones del país.

_Muestreo
Muestreo probabilístico con selección aleatoria de individuo y estratificado previamente por región.

_Muestra y cobertura semanal
703 casos. Margen de error de ±3,7 puntos porcentuales al 95% de confianza.
Se alcanzó una cobertura total de 190 comunas. El 90% de la muestra fue aplicada en población urbana y el 10% en población rural.

_Tasa de logro
Para lograr los 703 casos efectivos se realizaron un total de 4.059 llamados, lo que representa una tasa de éxito del 17,3%.

_Ponderación
Los datos fueron ponderados a nivel de sujetos por zona, género y edad, obteniendo una muestra de representación nacional para el universo en estudio.

_Fecha de terreno
Jueves 23 al viernes 24 de septiembre de 2021.

Para mayor información y detalle sobre metodología visita el sitio cadem.cl/plaza-publica/

Plaza PÚBLICA CADEM

Muestra aleatoria

Independientes e idénticamente distribuidos (*iid*)

- Cantidad de datos finita.
- Cada dato es independiente del anterior.
- Cada dato posee misma distribución.

Inferencia

- Como todos tienen misma distribución, podemos inferir sobre ella.
- Modelos estadísticos que suponen cantidad finita de parámetros se llaman *paramétricos*.
- Ejemplo: $\mathcal{N}\left(\mu, \sigma^2\right)$ tiene solo dos parámetros.
- Un estadístico que busca estimar un parámetro de la población se llama *estimador*.

Muestra aleatoria

Independientes e idénticamente distribuidos (*iid*)

- Cada dato puede modelarse como una variable aleatoria $X^{(i)}$.
- Cada dato podría ser multivariado, $\mathrm{X}^{(i)} = \left(X_1^{(i)}, \ldots, X_J^{(i)}\right)$.
- Matriz de datos:

$$\mathrm{X} = \begin{bmatrix} \mathrm{X}^{(1)\,T} \\ \vdots \\ \mathrm{X}^{(i)\,T} \\ \vdots \\ \mathrm{X}^{(n)\,T} \end{bmatrix} = \begin{bmatrix} X_1^{(1)} & \cdots & X_j^{(1)} & \cdots & X_J^{(1)} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ X_1^{(i)} & \cdots & X_j^{(i)} & \cdots & X_J^{(i)} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ X_1^{(n)} & \cdots & X_j^{(n)} & \cdots & X_J^{(n)} \end{bmatrix}$$

Media muestral

Estadístico

- Sea X una muestra iid univariada.
- Sea $T_n = \sum_{i=1}^{n} X^{(i)}$, un estadístico.
- Sea $\bar{X}_n = \frac{1}{n} T_n = \frac{1}{n} \sum_{i=1}^{n} X^{(i)}$, un estadístico.

Propiedades

- Si cada observación sigue una distribución con media μ y varianza σ^2:
- $\mathbb{E}\left[\bar{X}_n\right] = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[X^{(i)}\right] = \frac{1}{n} n\mu = \mu$.
- $\mathbb{V}\left[\bar{X}_n\right] = \frac{1}{n^2} \sum_{i=1}^{n} \mathbb{V}\left[X^{(i)}\right] = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$.

- ¿Qué significa que $\bar{X}_n$ sea una variable aleatoria con media μ y varianza $\frac{\sigma^2}{n}$?

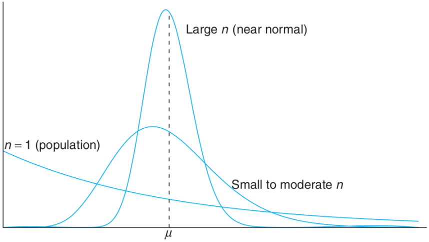
Sobre los estimadores

Funciones de la muestra aleatoria

- $\hat{\theta}_n = g\left(X^{(1)}, X^{(2)}, \ldots, X^{(n)}\right)$.
- Son estimadores *puntuales*, nos entregan un valor para una muestra.

¡Son variables aleatorias!

- Su valor depende del resultado de un experimento aleatorio.



Media muestral

Ejemplo: tragamonedas

- Cada tirada tiene valor esperado $-\$1000$ y desviación estándar de $\$10000$.
- ¿Qué se espera en promedio luego de jugar...
  - 1 vez?
  - 10 veces?
  - 100 veces?
  - 1000 veces?

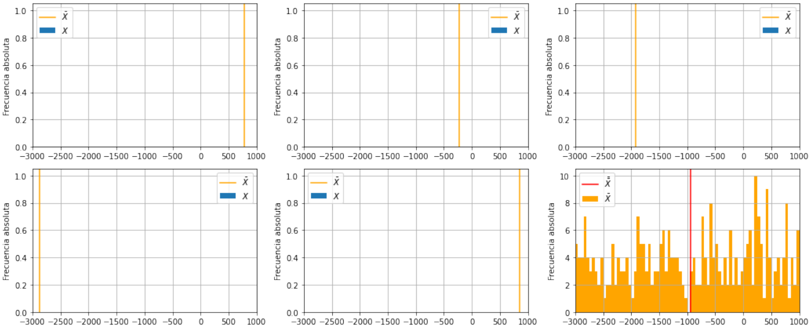
Sabemos que

- $\mathbb{E}\left[X^{(i)}\right] = -1000.$
- $\mathbb{V}\left[X^{(i)}\right] = 10^8.$
- $\mathbb{E}\left[\bar{X}_n\right] = -1000.$
- $\mathbb{V}\left[\bar{X}_n\right] = \frac{10^8}{n}.$

Media muestral – simulación

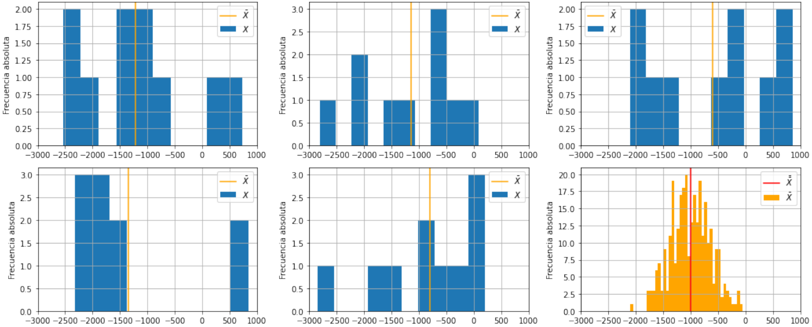
Ejemplo: tragamonedas

- Suponiendo $X^{(i)} \sim U(-3000, 1000)$.
- 1 juego.



Ejemplo: tragamonedas

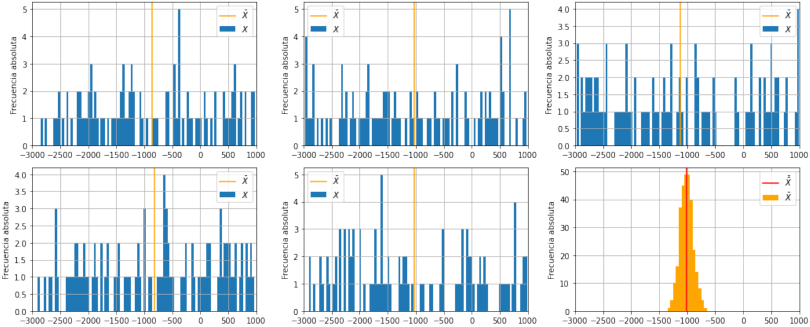
- Suponiendo $X^{(i)} \sim U(-3000, 1000)$.
- 10 juegos.



Media muestral – simulación

Ejemplo: tragamonedas

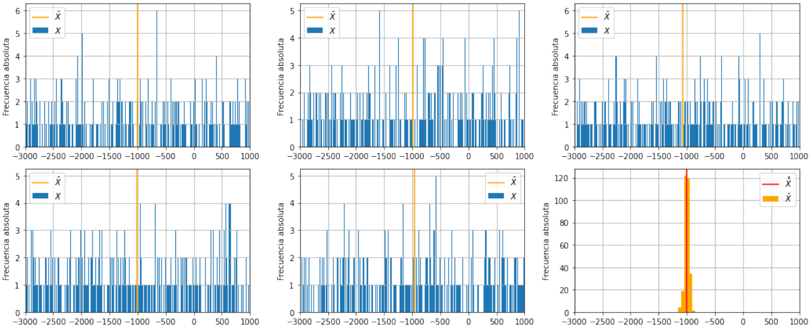
- Suponiendo $X^{(i)} \sim U(-3000, 1000)$.
- 100 juegos.



Media muestral – simulación

Ejemplo: tragamonedas

- Suponiendo $X^{(i)} \sim U(-3000, 1000)$.
- 1000 juegos.



Ley de los grandes números

Media muestral

- Sea X una variable aleatoria con media μ y varianza σ^2.
- Sea $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X^{(i)}$ media de $n \in \mathbb{N}$ observaciones.

Teorema: Ley *débil* de los grandes números

$$\lim_{n\to\infty} P\left(\left|\bar{X}_n - \mu\right| \geq \epsilon\right) = 0.$$

- Para todo $\epsilon > 0 \in \mathbb{R}$ elegido se cumple.
- Para todo $\epsilon, \delta > 0 \in \mathbb{R}$, existe $n \in \mathbb{N}$ tal que $P\left(\left|\bar{X}_n - \mu\right| \geq \epsilon\right) < \delta$.

Nota: Tipos de convergencia

En probabilidad (ley *débil*)

$$\lim_{n\to\infty} P\left(\left|\bar{X}_n - \mu\right| \geq \epsilon\right) = 0.$$

- Para todo $\epsilon, \delta > 0 \in \mathbb{R}$, existe $n \in \mathbb{N}$ tal que $P\left(\left|\bar{X}_n - \mu\right| \geq \epsilon\right) < \delta$.

$$\bar{X}_n \xrightarrow{P} \mu.$$

Casi segura o casi en todas partes (ley *fuerte*)

$$P\left(\lim_{n\to\infty} \bar{X}_n = \mu\right) = 1.$$

- Para toda secuencia infinita observada ω, la media $\bar{X}_n$ converge a μ, exceptuando, a lo más, un conjunto de probabilidad 0.

$$\bar{X}_n \xrightarrow{\text{c.s.}} \mu.$$

Ley débil de los grandes números – demostración

Teorema: Ley *débil* de los grandes números

$$\lim_{n\to\infty} P\left(\left|\bar{X}_n - \mu\right| \geq \epsilon\right) = 0.$$

- Para todo $\epsilon, \delta > 0 \in \mathbb{R}$, existe $n \in \mathbb{N}$ tal que $P\left(\left|\bar{X}_n - \mu\right| \geq \epsilon\right) < \delta$.

Desigualdad de Chebyschev

- Sea $k > 0 \in \mathbb{R}$ y X una variable aleatoria con media μ y varianza σ^2.
- Entonces:

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}.$$

Ley débil de los grandes números – demostración

Demostración desigualdad de Chebyschev

$$
\begin{aligned}
P(|X-\mu| \geq k\sigma) &= P(X \leq \mu - k\sigma) + P(X \geq \mu + k\sigma) \\
&= \int_{-\infty}^{\mu-k\sigma} p_X(x)dx + \int_{\mu+k\sigma}^{+\infty} p_X(x)dx \\
&\leq \int_{-\infty}^{\mu-k\sigma} \frac{(x-\mu)^2}{k^2\sigma^2} p_X(x)dx + \int_{\mu+k\sigma}^{+\infty} \frac{(x-\mu)^2}{k^2\sigma^2} p_X(x)dx \\
&= \frac{1}{k^2\sigma^2}\left(\int_{-\infty}^{\mu-k\sigma} (x-\mu)^2 p_X(x)dx + \int_{\mu+k\sigma}^{+\infty} (x-\mu)^2 p_X(x)dx\right) \\
&\leq \frac{1}{k^2\sigma^2} \int_{-\infty}^{+\infty} (x-\mu)^2 p_X(x)dx = \frac{1}{k^2\sigma^2}\sigma^2 = \frac{1}{k^2} \\
&\Rightarrow P(|X-\mu| \geq k\sigma) \leq \frac{1}{k^2}
\end{aligned}
$$

Ley débil de los grandes números – demostración

Teorema: Ley *débil* de los grandes números

- Para todo $\epsilon, \delta > 0 \in \mathbb{R}$, existe $n \in \mathbb{N}$ tal que $P\left(\left|\bar{X}_n - \mu\right| \geq \epsilon\right) < \delta$:
$$\lim_{n\to\infty} P\left(\left|\bar{X}_n - \mu\right| \geq \epsilon\right) = 0.$$

Demostración

- Sabemos que $\mathbb{V}\left[\bar{X}_n\right] = \frac{\sigma^2}{n}$.
- Entonces, para todo $k > 0 \in \mathbb{R}$:
$$P\left(\left|\bar{X}_n - \mu\right| \geq \frac{k\sigma}{\sqrt{n}}\right) \leq \frac{1}{k^2}.$$
- Basta elegir $k = \frac{\epsilon\sqrt{n}}{\sigma}$ y se tiene:
$$P\left(\left|\bar{X}_n - \mu\right| \geq \epsilon\right) \leq \frac{\sigma^2}{\epsilon^2 n}.$$
- Finalmente, basta tomar $n > \frac{\sigma^2}{\epsilon^2\delta}$ y se tiene que $P\left(\left|\bar{X}_n - \mu\right| \geq \epsilon\right) < \delta$.

Ley de los grandes números

Ejemplo

- Sea $X \sim \text{Bernoulli}\left(\frac{\pi}{4}\right)$.
- Sabemos que $\mathbb{E}[X] = \mu = \frac{\pi}{4}$ y $\mathbb{V}[X] = \sigma^2 = \frac{\pi}{4}\left(1 - \frac{\pi}{4}\right)$.
- Por ley de los grandes números, sabemos que $\bar{X}_n$ converge a $\mu = \frac{\pi}{4}$.

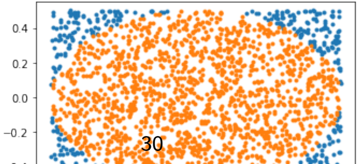
Cálculo de π

- Obtenemos n observaciones de X y calculamos la media $\bar{X}_n$.
- Sabemos que $\bar{X}_n \to \frac{\pi}{4}$.
- Entonces, aproximamos $\pi \approx 4\bar{X}_n$.
- ¿Y cómo obtenemos observaciones de X?

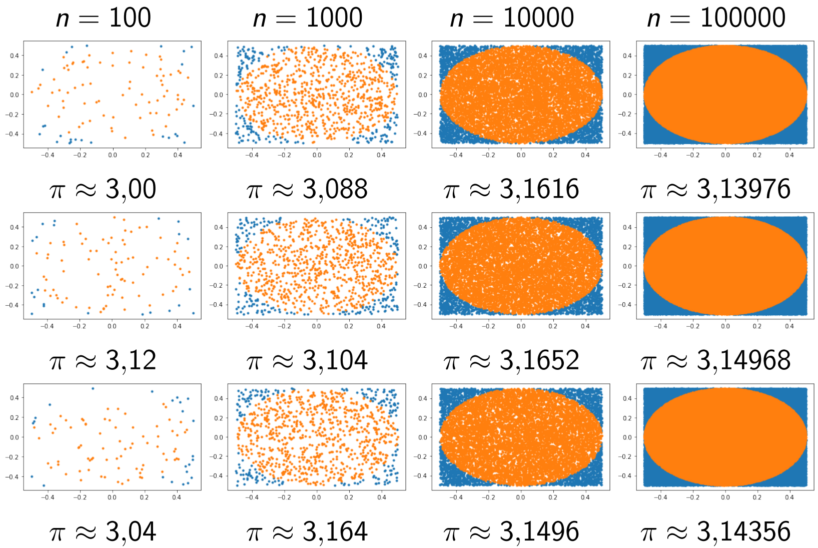
Ley de los grandes números

Simulando $X \sim \text{Bernoulli}\left(\frac{\pi}{4}\right)$

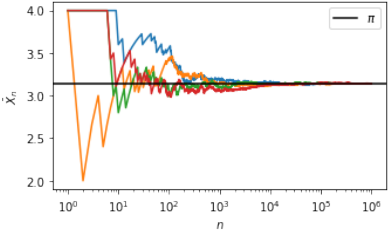
- Sean $X_1, X_2 \sim U\left(-\frac{1}{2}, \frac{1}{2}\right)$.
- Sea $X = \begin{cases} 1 \text{ si } (X_1, X_2) \text{ dentro de círculo radio } \frac{1}{2} \text{ y centro } (0,0) \\ 0 \text{ si no} \end{cases}$.
- Notar que el espacio muestral $\Omega = X_1 \times X_2$ es un cuadrado de lado 1.
- El área que corresponde a $X = 1$ es el área del cículo de radio $\frac{1}{2}$.
- $P(X = 1) = \frac{\pi}{4}$ y $P(X = 0) = 1 - \frac{\pi}{4}$. Es decir, $X \sim \text{Bernoulli}\left(\frac{\pi}{4}\right)$.

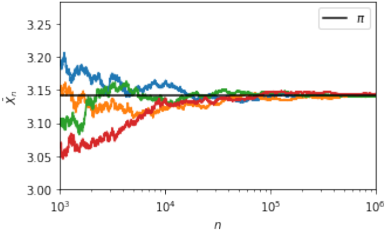


Aproximando π



Aproximando π





- Esta metodología se llama aproximación de Monte Carlo.

Teorema del límite central

Media muestral

- Sea X una variable aleatoria con media μ y varianza σ^2.
- Sea $T_n = \sum_{i=1}^{n} X^{(i)}$, suma de $n \in \mathbb{N}$ observaciones.
- Sea $\bar{X}_n = \frac{1}{n} T_n = \frac{1}{n} \sum_{i=1}^{n} X^{(i)}$ media de $n \in \mathbb{N}$ observaciones.

Teorema

- La variable aleatoria Z_n siguiente converge a una distribución normal estándar $\mathcal{N}(0, 1)$:

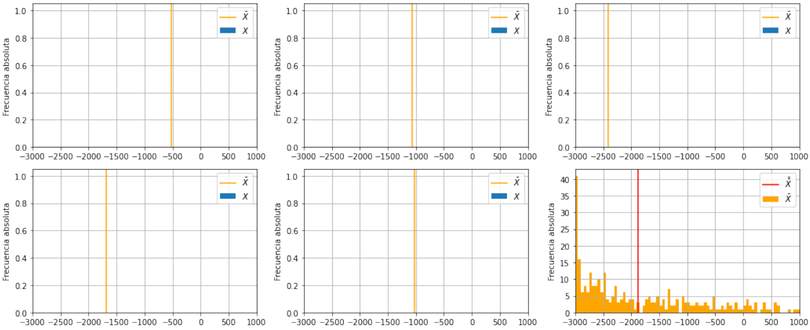
$$Z_n = \frac{T_n - n\mu}{\sigma\sqrt{n}} = \frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}}$$

- Equivalentemente, $\bar{X}_n \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$.

Media muestral – simulación

Ejemplo: tragamonedas

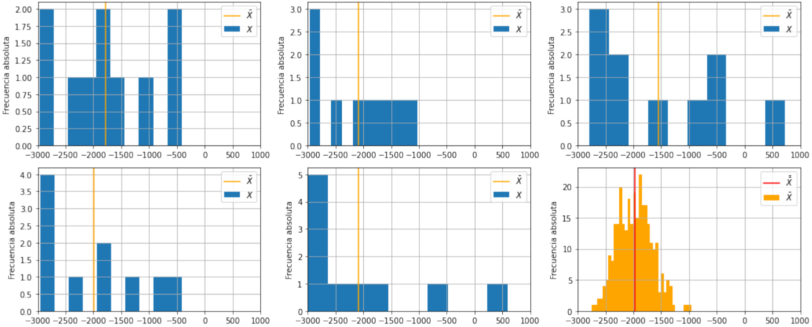
- Suponiendo $Y^{(i)} \sim \text{Beta}(0{,}5, 1{,}5)$ y $X^{(i)} = 4000\,Y^{(i)} - 3000$.
- 1 juego.



Media muestral – simulación

Ejemplo: tragamonedas

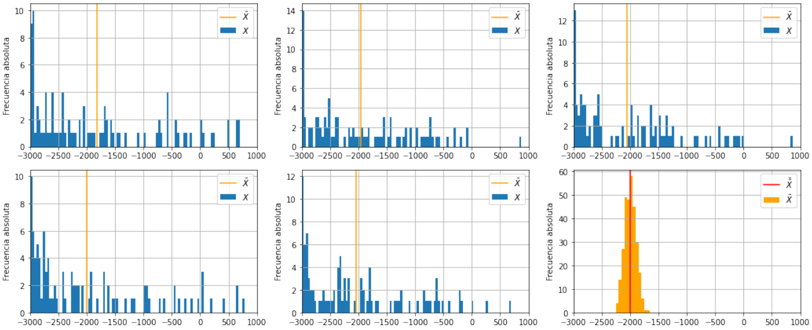
- Suponiendo $Y^{(i)} \sim \text{Beta}(0{,}5, 1{,}5)$ y $X^{(i)} = 4000\,Y^{(i)} - 3000$.
- 10 juegos.



Media muestral – simulación

Ejemplo: tragamonedas

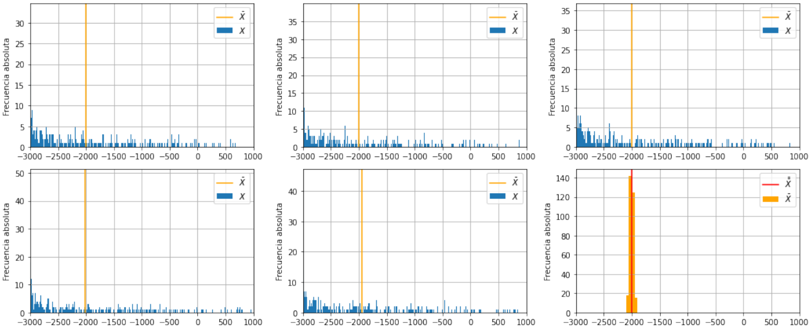
- Suponiendo $Y^{(i)} \sim \text{Beta}(0{,}5, 1{,}5)$ y $X^{(i)} = 4000\,Y^{(i)} - 3000$.
- 100 juegos.



Media muestral – simulación

Ejemplo: tragamonedas

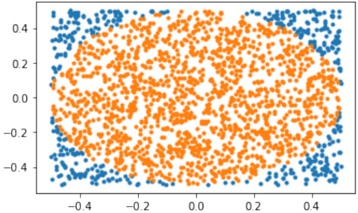
- Suponiendo $Y^{(i)} \sim \mathrm{Beta}(0{,}5, 1{,}5)$ y $X^{(i)} = 4000Y^{(i)} - 3000$.
- 1000 juegos.



Ejemplo: aproximación de Monte Carlo de π

Supuesto

- $X \sim \text{Bernoulli}\left(\frac{\pi}{4}\right)$.
- $\mathbb{E}[X] = \mu = \frac{\pi}{4}$.
- $\mathbb{V}[X] = \sigma^2 = \frac{\pi}{4}\left(1 - \frac{\pi}{4}\right)$.



Teorema del límite central

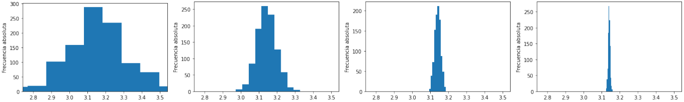
- $\bar{X}_n \sim \mathcal{N}\left(\frac{\pi}{4}, \frac{\frac{\pi}{4}\left(1-\frac{\pi}{4}\right)}{n}\right)$

Ejemplo: aproximación de Monte Carlo de π

Teorema del límite central

- $\bar{X}_n \sim \mathcal{N}\left(\frac{\pi}{4}, \frac{\frac{\pi}{4}\left(1-\frac{\pi}{4}\right)}{n}\right)$. Si $Y_n = 4\bar{X}_n$, entonces $Y_n \sim \mathcal{N}\left(\pi, \frac{\pi(4-\pi)}{n}\right)$.

n	100	1000	10000	100000
Secuencia 1	3,00	3,088	3,1616	3,13976
Secuencia 2	3,12	3,104	3,1652	3,14968
Secuencia 3	3,04	3,164	3,1496	3,14356
$\mathbb{E}[Y_n]$	3,14	3,142	3,1416	3,14159
$\sqrt{\mathbb{V}[Y_n]}$	0,16	0,052	0,0164	0,00519



Algunas aproximaciones

Media

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X^{(i)} \to \mathbb{E}[X].$$

Varianza (con media conocida)

$$S_n^2 = \frac{1}{n} \sum_{i=1}^{n} \left(X^{(i)} - \mathbb{E}[X] \right)^2 \to \mathbb{V}[X] = \mathbb{E}\left[(X - \mathbb{E}[X])^2 \right].$$

Función de distribución acumulada

$$\frac{1}{n} \left| \left\{ X^{(i)} \text{ tal que } X^{(i)} \leq c \right\} \right| \to F_X(c) = \mathbb{E}\left[1_{(-\infty, c)}(X) \right].$$

Ejemplo

Errores en un programa computacional

- Sea X variable aleatoria asociada a cantidad de errores por semana.
- Supongamos X sigue una distribución de Poisson Pois($\lambda = 5$).
- Hay 125 programas independientes corriendo.
- ¿Probabilidad de que cantidad de errores promedio sea menor a 5,5?

Ejemplo

Errores en un programa computacional

- Sea X variable aleatoria asociada a cantidad de errores por semana.
- Supongamos X sigue una distribución de Poisson $\text{Pois}(\lambda = 5)$.
- Hay 125 programas independientes corriendo.
- ¿Probabilidad de que cantidad de errores promedio sea menor a 5,5?

Desarrollo

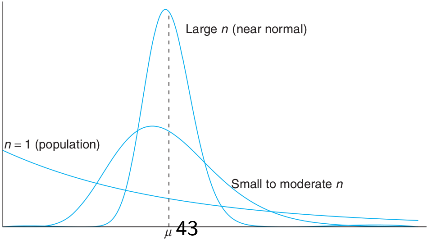
- $\mathbb{E}[X] = \mu = 5$.
- $\mathbb{V}[X] = \sigma^2 = 5$.
- Sabemos que $\bar{X}_{125}$ se parece a $\mathcal{N}\left(\mu, \frac{\sigma^2}{125}\right) = \mathcal{N}\left(5, \frac{1}{25}\right)$.

$$P\left(\bar{X}_{125} < 5{,}5\right) = F_{\bar{X}_{125}}(5{,}5) = \Phi\left(\frac{5{,}5 - 5}{\sqrt{\frac{1}{25}}}\right) = \Phi(2{,}5) \approx 0{,}9938.$$

Intervalos de confianza

Errores en un programa computacional

- Sea X variable aleatoria asociada a cantidad de errores por semana.
- Supongamos X sigue una distribución con media desconocida μ y varianza $\sigma^2 = 5$.
- Hay 125 programas independientes corriendo una semana.
- Medimos la cantidad de errores promedio, obteniendo $\bar{x}_{125} = 6$.
- $\bar{x}_{125} = 6$ es una estimación puntual de μ.
- ¿Entre qué valores está μ, con un 89 % de probabilidad?



Intervalos de confianza

Errores en un programa computacional

- $\bar{x}_{125} = 6$ es una estimación puntual de μ.
- ¿Entre qué valores está μ, con un 89 % de probabilidad?

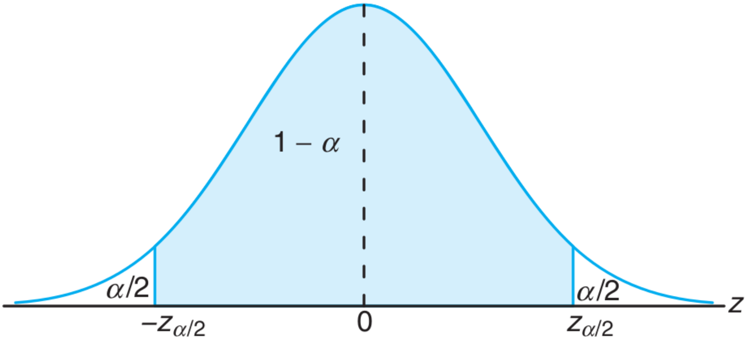
Desarrollo

- Sabemos que $\bar{X}_{125}$ se parece a $\mathcal{N}\left(\mu, \frac{\sigma^2}{125}\right) = \mathcal{N}\left(\mu, \frac{1}{25}\right)$.
- Podemos escribir $\bar{X}_{125} \approx \frac{1}{5}Z + \mu$, con $Z \sim \mathcal{N}(0,1)$.
- Buscamos el intervalo centrado en 0 para Z:
$$P(-1{,}598 \leq Z < 1{,}598) \approx 0{,}89.$$
- Resolvemos para μ:
$$P\left(-1{,}598 \leq 5\left(\bar{X}_{125} - \mu\right) < 1{,}598\right) \approx 0{,}89$$
$$\Rightarrow -6{,}32 \leq -\mu < -5{,}68 \Rightarrow \mu \in (5{,}68, 6{,}32]\,.$$

Intervalos de confianza

En general

- Dado un $\alpha \in [0, 1]$.
- Buscamos un intervalo con un nivel dado de *confianza* $1 - \alpha$.
- En este curso vamos a suponer intervalos centrados.



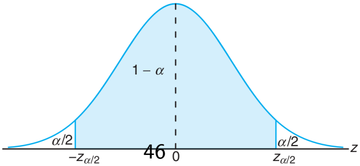
Intervalos de confianza

Si conocemos varianza σ^2 y queremos estimar μ

- Sabemos que $\bar{X}_n$ se parece a $\mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$.
- Podemos escribir $\bar{X}_n \approx \frac{\sigma}{\sqrt{n}} Z + \mu$, con $Z \sim \mathcal{N}(0,1)$.
- Buscamos intervalo centrado para Z:
  $P\left(-z_{\frac{\alpha}{2}} \leq Z < z_{\frac{\alpha}{2}}\right) = 1 - \alpha$.
- Resolvemos para μ:

$$P\left(-z_{\frac{\alpha}{2}} \leq \sqrt{n}\frac{\bar{X}_n - \mu}{\sigma} < z_{\frac{\alpha}{2}}\right) = 1 - \alpha$$

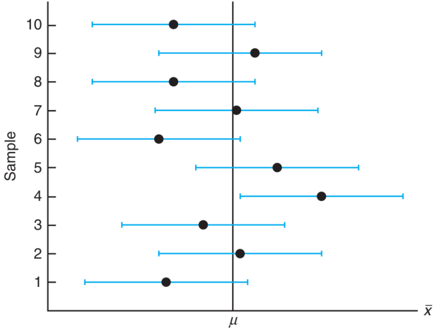
$$\Rightarrow \bar{X}_n - \frac{\sigma}{\sqrt{n}} z_{\frac{\alpha}{2}} < \mu \leq \bar{X}_n + \frac{\sigma}{\sqrt{n}} z_{\frac{\alpha}{2}}.$$



Intervalos de confianza

Límites del intervalo

- Notar que $\bar{X}_n - \frac{\sigma}{\sqrt{n}} z_{\frac{\alpha}{2}}$ y $\bar{X}_n + \frac{\sigma}{\sqrt{n}} z_{\frac{\alpha}{2}}$ son variables aleatorias...



Nota: *Bootstrap*

- Obtener intervalos de confianza usando distintas muestras generadas basadas la muestra original, con repetición.

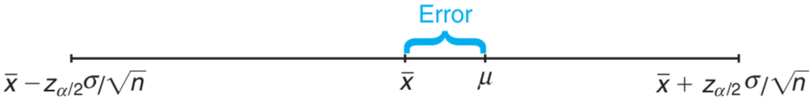
Intervalos de confianza

Frecuentista

- En este caso μ tiene un valor fijo.
- La distribución es de $\bar{X}_n$ y los límites del intervalo.

Interpretación

- Si se usa $\bar{X}_n$ para estimar μ, se tiene confianza de $1-\alpha$ de que el error no excede $\frac{\sigma}{\sqrt{n}} z_{\frac{\alpha}{2}}$.



Interpretación

- Si se usa $\bar{X}_n$ para estimar μ, se tiene confianza de $1-\alpha$ de que el error no excede e cuando la muestra es de tamaño $n = \left(\frac{\sigma}{e} z_{\frac{\alpha}{2}}\right)^2$.

Intervalos de confianza

Si no conocemos varianza σ^2 y queremos estimar μ

- Supongamos que $X^{(i)}$ tiene distribución normal $\mathcal{N}(\mu, \sigma^2)$.
- Podemos estimar σ^2 con el estadístico varianza muestral
$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(X^{(i)} - \bar{X}_n \right)^2 .$$

Distribución de la media y varianza muestrales

- Sabemos que $\bar{X}_n \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$.
- Sabemos que $Z = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}$ es normal estándar, $Z \sim \mathcal{N}(0,1)$.
- $V = \frac{(n-1)S_n^2}{\sigma^2}$ tiene distribución ji al cuadrado con $n-1$ grados de libertad, $V \sim \chi^2(n-1)$:
$$p_V(v) = \frac{1}{2^{\frac{n-1}{2}} \Gamma\left(\frac{n-1}{2}\right)} v^{\frac{n-1}{2}-1} e^{-\frac{v}{2}}, \text{ con } v \in [0, +\infty) .$$

Intervalos de confianza

Distribución de la media y varianza muestrales

- Sabemos que $\bar{X}_n \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$.
- Sabemos que $Z = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}$ es normal estándar, $Z \sim \mathcal{N}(0,1)$.
- $V = \frac{(n-1)S_n^2}{\sigma^2}$ tiene distribución ji al cuadrado con $n-1$ grados de libertad, $V \sim \chi^2(n-1)$:
$$p_V(v) = \frac{1}{2^{\frac{n-1}{2}}\Gamma\left(\frac{n-1}{2}\right)} v^{\frac{n-1}{2}-1} e^{-\frac{v}{2}}, \text{ con } v \in [0, +\infty)\,.$$

Teorema

- La variable aleatoria $T = \frac{Z}{\sqrt{\frac{V}{n-1}}}$ es distribución t de Student con $n-1$ grados de libertad:
$$p_T(t) = \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)\sqrt{\pi(n-1)}}\left(1 + \frac{t^2}{n-1}\right)^{-\frac{n}{2}}.$$

Intervalos de confianza

Distribución de la media y varianza muestrales

- Sabemos que $Z = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}$ es normal estándar, $Z \sim \mathcal{N}(0,1)$.
- $V = \frac{(n-1)S_n^2}{\sigma^2}$ tiene distribución ji al cuadrado con $n-1$ grados de libertad, $V \sim \chi^2(n-1)$.

Reescribimos T

- La variable aleatoria $T = \frac{Z}{\sqrt{\frac{V}{n-1}}}$ es distribución t de Student con $n-1$ grados de libertad.

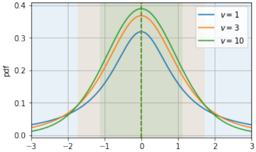
$$T = \frac{Z}{\sqrt{\frac{V}{n-1}}} = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma\sqrt{\frac{(n-1)S_n^2}{\sigma^2(n-1)}}} = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sqrt{S_n^2}}.$$

Distribución t de Student con ν grados de libertad $t(\nu)$

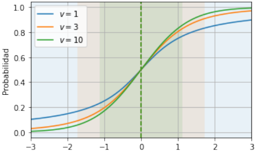
Definición

$$p(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\pi\nu}}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}.$$

$pdf\ (p(x))$



$cdf\ (F(x))$



Propiedades

- $\mathbb{E}_{X\sim t(\nu)}[X] = \mu_X = 0.$
- $\mathbb{V}_{X\sim t(\nu)}[X] = \sigma_X^2 = \begin{cases} +\infty \text{ si } \nu \leq 2, \\ \frac{\nu}{\nu-2} \text{ si } \nu > 2 \end{cases}.$

Intervalos de confianza

Si no conocemos varianza σ^2 y queremos estimar μ

- Supongamos que $X^{(i)}$ tiene distribución normal $\mathcal{N}(\mu, \sigma^2)$.
- Calculamos media y varianza muestrales $\bar{X}_n$ y S_n^2.
- Sabemos que $T = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sqrt{S_n^2}}$ sigue distribución $t(n-1)$.
- Buscamos intervalo centrado para T:
  $P\left(-t_{\frac{\alpha}{2}} \leq T < t_{\frac{\alpha}{2}}\right) = 1 - \alpha$.
- Resolvemos para μ:

$$P\left(-t_{\frac{\alpha}{2}} \leq \sqrt{n}\frac{\bar{X}_n - \mu}{\sqrt{S_n^2}} < t_{\frac{\alpha}{2}}\right) = 1 - \alpha$$

$$\Rightarrow \bar{X}_n - \frac{\sqrt{S_n^2}}{\sqrt{n}} t_{\frac{\alpha}{2}} < \mu \leq \bar{X}_n + \frac{\sqrt{S_n^2}}{\sqrt{n}} t_{\frac{\alpha}{2}}.$$

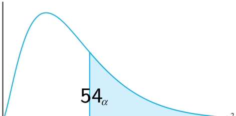
Intervalos de confianza

Si conocemos μ y queremos estimar σ^2

- Supongamos que $X^{(i)}$ tiene distribución normal $\mathcal{N}(\mu, \sigma^2)$.
- Sabemos que $V = \frac{(n-1)S_n^2}{\sigma^2}$ tiene distribución ji al cuadrado con $n-1$ grados de libertad, $V \sim \chi^2(n-1)$.
- Buscamos intervalo centrado para V:
  $P\left(v_{\frac{\alpha}{2}} \leq V < v_{1-\frac{\alpha}{2}}\right) = 1 - \alpha.$
- Resolvemos para σ^2:

$$P\left(v_{\frac{\alpha}{2}} \leq \frac{(n-1)S_n^2}{\sigma^2} < v_{1-\frac{\alpha}{2}}\right) = 1 - \alpha$$

$$\Rightarrow \frac{(n-1)S_n^2}{v_{\frac{1-\alpha}{2}}} < \sigma^2 \leq \frac{(n-1)S_n^2}{v_{\frac{\alpha}{2}}}.$$

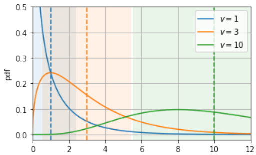


Distribución ji al cuadrado v grados de libertad $\chi^2(v)$

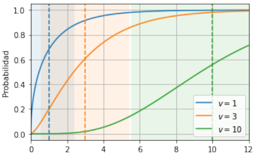
Definición

$$p(x) = \frac{x^{\frac{v}{2}-1} e^{-\frac{v}{2}}}{2^{\frac{v}{2}} \Gamma\left(\frac{v}{2}\right)}.$$

$pdf \; (p(x))$



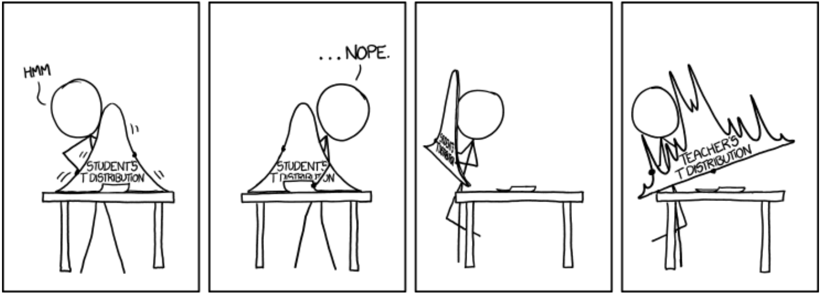
$cdf \; (F(x))$



Propiedades

- $\mathbb{E}_{X\sim\chi^2(v)}[X] = \mu_X = v.$
- $\mathbb{V}_{X\sim\chi^2(v)}[X] = \sigma_X^2 = 2v.$

XKCD


HMM
STUDENT'S
T DISTRIBUTION
. . . NOPE.
STUDENT'S
TEACHER'S
T DISTRIBUTION

Sesgo

Estimador insesgado

- Un estimador $\hat{\theta}$ es *insesgado* si $\mathbb{E}\left[\hat{\theta}\right] = \theta$ para cualquier valor de θ.
- En caso contrario, el estimador es sesgado y $\mathbb{E}\left[\hat{\theta}\right] - \theta$ se llama *sesgo*.

Ejemplo: varianza muestral

- Consideremos el siguiente estimador de la varianza:

$$\hat{\sigma}_n^2 = \frac{1}{n} \sum_{i=1}^{n} \left(X^{(i)} - \bar{X}_n\right)^2.$$

- ¿Es sesgado?

Ejemplo: varianza muestral

$$\mathbb{E}\left[\hat{\sigma}_n^2\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left(X^{(i)} - \bar{X}_n\right)^2\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[X^{(i)2} + \bar{X}_n^2 - 2X^{(i)}\bar{X}_n\right]$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left(\mathbb{E}\left[X^{(i)2}\right] + \mathbb{E}\left[\bar{X}_n^2\right] - 2\mathbb{E}\left[X^{(i)}\bar{X}_n\right]\right)$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left(\mathbb{E}\left[X^{(i)2}\right] + \mathbb{E}\left[\frac{1}{n^2}\sum_{j=1}^{n}\sum_{k=1}^{n}X^{(j)}X^{(k)}\right] - 2\mathbb{E}\left[\frac{1}{n}X^{(i)}\sum_{j=1}^{n}X^{(j)}\right]\right)$$

Sesgo

Ejemplo: varianza muestral

$$= \sum_{i=1}^{n} \left(\frac{\sigma^2 + \mu^2}{n} + \frac{(n^2 - n)\mu^2 + n(\sigma^2 + \mu^2)}{n^3} - \frac{2((n-1)\mu^2 + \sigma^2 + \mu^2}{n^2} \right.$$

$$= \frac{1}{n^2}\left(n^2\sigma^2 + n^2\mu^2 + n^2\mu^2 + n\sigma^2 - 2n^2\mu^2 - 2n\sigma^2\right)$$

$$= \frac{(n^2 - n)}{n^2}\sigma^2 = \frac{n-1}{n}\sigma^2.$$

Sesgo

Ejemplo: varianza muestral

- ¡El estimador $\hat{\sigma}_n^2$ es sesgado!
- $\mathbb{E}\left[\hat{\sigma}_n^2\right] = \frac{n-1}{n}\sigma^2$.

Corrección: varianza muestral con $n-1$

- Consideremos el siguiente estimador de la varianza:

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X^{(i)} - \bar{X}_n\right)^2 = \frac{n}{n-1}\hat{\sigma}_n^2.$$

- Ahora tenemos que:

$$\mathbb{E}\left[S_n^2\right] = \frac{1}{n-1}\mathbb{E}\left[\hat{\sigma}_n^2\right] = \frac{n}{n-1}\frac{n-1}{n}\sigma^2 = \sigma^2.$$

Otros estimadores

¿Cómo podemos obtener estimadores?

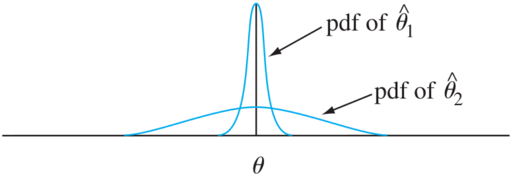
- Minimizar una función de error.
- Máxima verosimilitud.
- Máximo a posteriori.

Estimador insesgado de mínima varianza (*MVUE*)

Definición

- Estimador insesgado de θ:
$$\mathbb{E}\left[\hat{\theta}\right] = \theta.$$

- Menor varianza:
$$\hat{\theta} = \underset{e \text{ estimador insesgado}}{\arg\min} \mathbb{E}\left[(e-\theta)^2\right].$$



Nota

- Si es que existe, es único. - No siempre existe.

Minimizar una función de error

Error cuadático medio

- Definición:
$$\mathsf{MSE}\left(\hat{\theta}\right) = \mathbb{E}\left[\left(\hat{\theta} - \theta\right)^2\right].$$
- Al igual que antes, se minimiza MSE:
$$\hat{\theta} = \underset{e \text{ estimador}}{\arg\min}\ \mathsf{MSE}\left(\hat{\theta}\right) = \underset{e \text{ estimador}}{\arg\min}\ \mathbb{E}\left[(e - \theta)^2\right].$$

Descomposición

$$\begin{aligned}
\mathsf{MSE}\left(\hat{\theta}\right) &= \mathbb{E}\left[\left(\hat{\theta} - \mathbb{E}\left[\hat{\theta}\right] + \mathbb{E}\left[\hat{\theta}\right] - \theta\right)^2\right] \\
&= \underbrace{\mathbb{E}\left[\left(\hat{\theta} - \mathbb{E}\left[\hat{\theta}\right]\right)^2\right]}_{\text{varianza}} + \underbrace{\left(\mathbb{E}\left[\hat{\theta}\right] - \theta\right)^2}_{\text{sesgo}^2}
\end{aligned}$$

Minimizar una función de error

Ejemplo

- Sea $\hat{\theta}_{\mathsf{MVUE}}$ el estimador insesgado de mínima varianza.
- Sea $\hat{\theta}_\alpha = (1+\alpha)\hat{\theta}_{\mathsf{MVUE}}$ otro estimador.
- Notar que $\mathbb{E}\left[\hat{\theta}_\alpha\right] = (1+\alpha)\theta$. Si $\alpha \neq 0$, es sesgado.

Se tiene que

$$\mathsf{MSE}\left(\hat{\theta}_\alpha\right) = (1+\alpha)^2 \mathsf{MSE}\left(\hat{\theta}_{\mathsf{MVUE}}\right) + \alpha^2\theta^2.$$

El error es menor si

$$-\frac{2\mathsf{MSE}\left(\hat{\theta}_{\mathsf{MVUE}}\right)}{\mathsf{MSE}\left(\hat{\theta}_{\mathsf{MVUE}}\right) + \theta^2} < \alpha < 0.$$

Consistencia

Definición

- Un estimador $\hat{\theta}_n$ es consistente si converge a θ cuando $n \to \infty$.

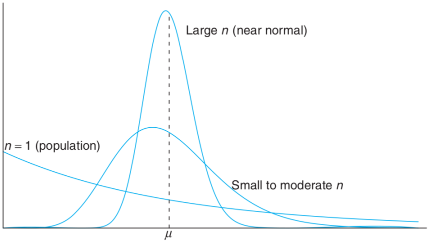
¿Por qué?

- Todo estimador cuyo sesgo y varianza convergen a 0 si $n \to \infty$, es consistente.
- Un estimador insesgado lo es para todo n.
- Aquí nos interesa que sea bueno cuando n es grande.
- La media muestral $\bar{X}_n$ es consistente.

Recordar – estimadores

Funciones de la muestra aleatoria

- $\hat{\theta}_n = g\left(X^{(1)}, X^{(2)}, \ldots, X^{(n)}\right)$.
- Son variables aleatorias: tienen una distribución de probabilidad.



En realidad no conocemos θ

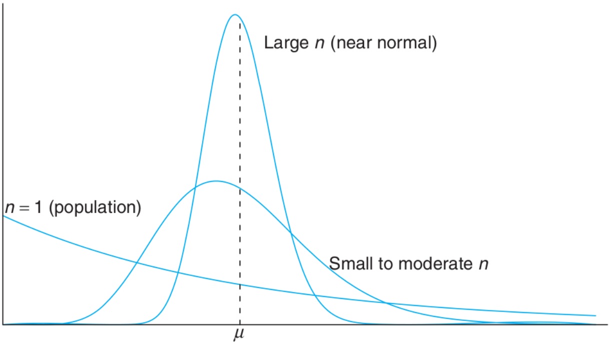
- Bayes:

$$p(\theta \mid \mathrm{X}_n) = \frac{p(\mathrm{X}_n \mid \theta) p(\theta)}{p(\mathrm{X}_n)}.$$

Enfoque bayesiano

En realidad no conocemos θ

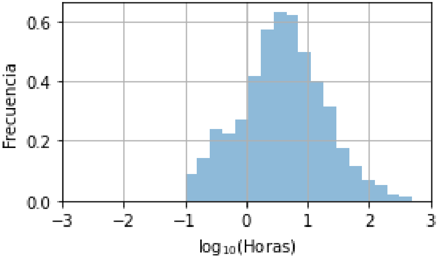
$$p(\theta \mid \mathrm{X}_n) = \frac{p(\mathrm{X}_n \mid \theta)p(\theta)}{p(\mathrm{X}_n)}.$$



Ejemplo: modelo de distribución normal

Datos

- Logaritmo de cantidad de horas jugadas en promedio de 3000 juegos de plataforma Steam.



Modelo

- Supongamos un modelo como $\mathcal{N}\left(\mu, \sigma^2\right)$.
- Queremos estimar μ y σ^2.

Ejemplo: modelo de distribución normal $\mathcal{N}(\mu, \sigma^2)$

Distribución conjunta de μ y σ^2

$$p\left(\mu, \sigma^2 \mid \mathrm{X}_n\right) = \frac{\overbrace{p\left(\mathrm{X}_n \mid \mu, \sigma^2\right)}^{\text{verosimilitud}} \overbrace{p\left(\mu, \sigma^2\right)}^{\text{a priori}}}{\underbrace{p\left(\mathrm{X}_n\right)}_{\text{evidencia}}}.$$

Componentes

- Verosimilitud:

$$p\left(\mathrm{X}_n \mid \mu, \sigma^2\right) = \prod_{i=1}^{n} p\left(x^{(i)} \mid \mu, \sigma^2\right) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\left(x^{(i)}-\mu\right)^2}{2\sigma^2}}.$$

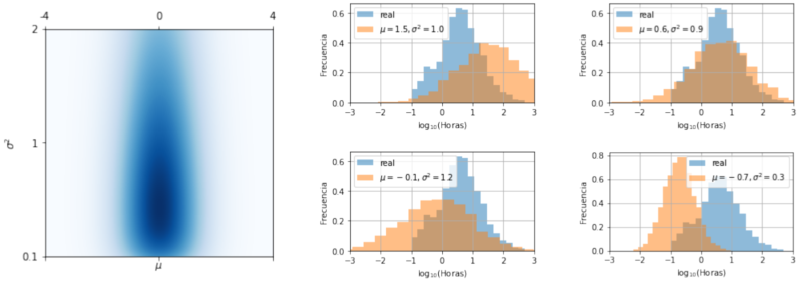
Ejemplo: modelo de distribución normal $\mathcal{N}(\mu, \sigma^2)$

Componentes

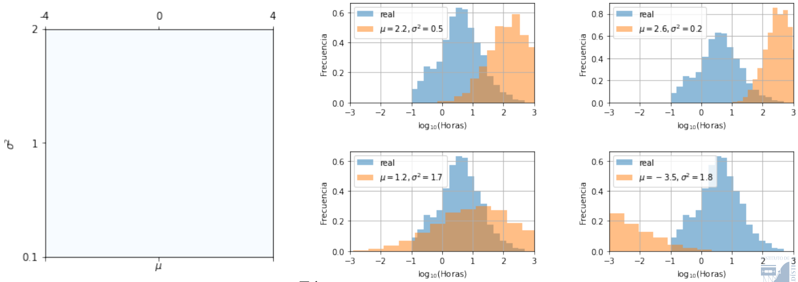
- A priori: veremos dos casos:
  1. $\mu \sim \mathcal{N}(0, 1)$ y $\sigma^2 \sim \Gamma(1{,}5, 0{,}01)$.
  2. $\mu \sim U(-4, 4)$ y $\sigma^2 \sim U(0{,}1, 2)$.
- Evidencia: cte. de normalización para $\iint p(\mu, \sigma^2 \mid X_n) d\mu d\sigma^2 = 1$.

Ejemplo: modelo de distribución normal $\mathcal{N}(\mu, \sigma^2)$

Caso 1

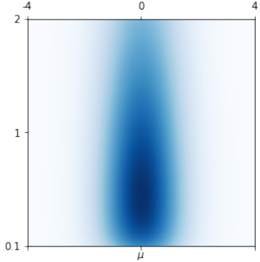


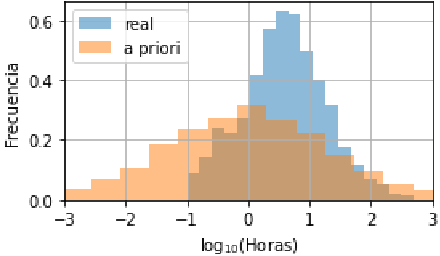
Caso 2



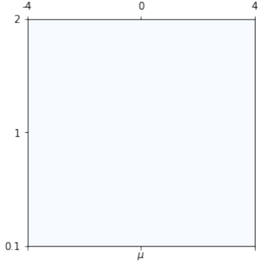
Ejemplo: modelo de distribución normal $\mathcal{N}(\mu, \sigma^2)$

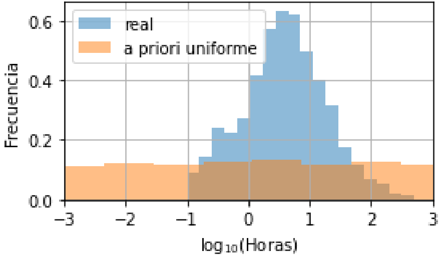
Caso 1



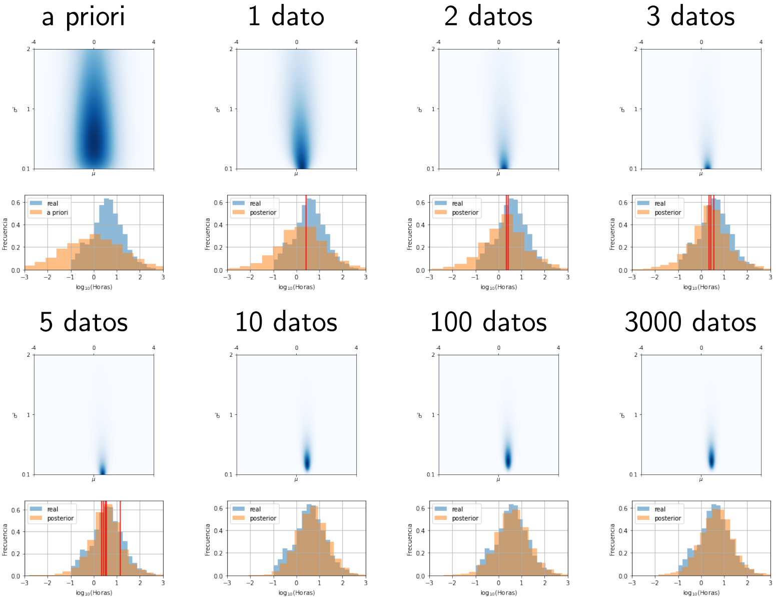


Caso 2

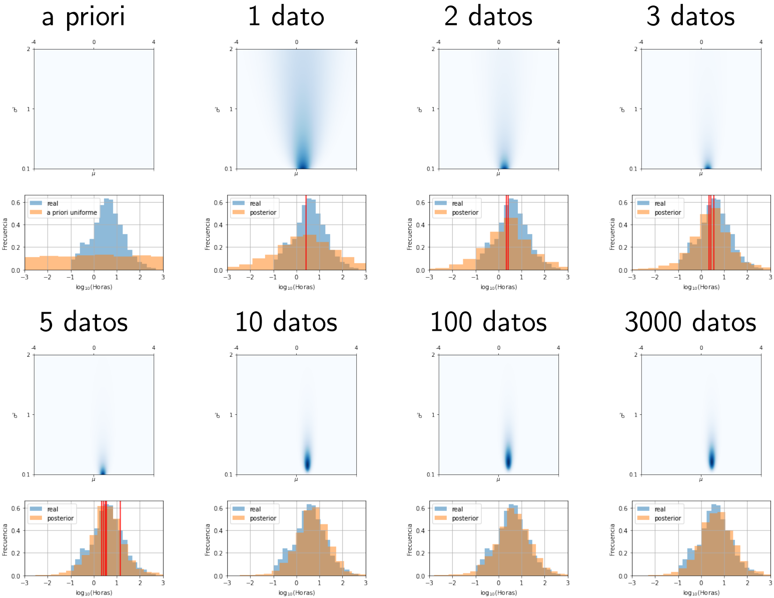




Ejemplo: modelo de distribución normal – caso 1



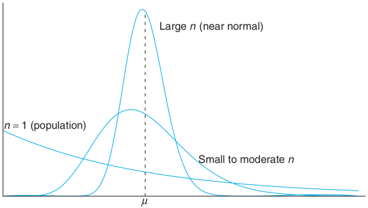
Ejemplo: modelo de distribución normal – caso 2

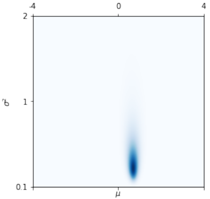


Enfoque bayesiano

Distribución de θ

$$p(\theta \mid \mathrm{X}_n) = \frac{p(\mathrm{X}_n \mid \theta)p(\theta)}{p(\mathrm{X}_n)}.$$



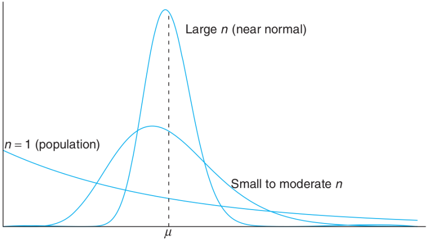


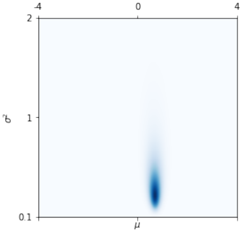
Una posibilidad: buscar el máximo de $p(\theta \mid \mathrm{X}_n)$

$$\hat{\theta}_n = \arg\max_{\theta} p(\theta \mid \mathrm{X}_n) = \arg\max_{\theta} \frac{p(\mathrm{X}_n \mid \theta)p(\theta)}{p(\mathrm{X}_n)}$$

$$= \arg\max_{\theta} p(\mathrm{X}_n \mid \theta)p(\theta).$$

Máxima verosimilitud (*Maximum likelihood estimate MLE*)





Máximo de $p(\theta \mid X_n)$

- En este enfoque, se ignora la distribución a priori.
- En caso de que θ es acotado, sería equivalente a una distribución uniforme a priori.

$$\hat{\theta}_n = \arg\max_{\theta} p(X_n \mid \theta).$$

Máxima verosimilitud (*Maximum likelihood estimate MLE*)

Máximo de $p(\theta \mid \mathrm{X}_n)$

$$\hat{\theta}_n = \arg\max_{\theta} p(\mathrm{X}_n \mid \theta).$$

Ejemplo: modelo de distribución normal $\mathcal{N}(\mu, \sigma^2)$

- Verosimilitud:
$$p\left(\mathrm{X}_n \mid \mu, \sigma^2\right) = \prod_{i=1}^{n} p\left(x^{(i)} \mid \mu, \sigma^2\right) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\left(x^{(i)}-\mu\right)^2}{2\sigma^2}}.$$

- Maximizamos logaritmo con respecto a μ, derivando e igualando a 0:
$$\frac{d\left[\ln\left(p\left(\mathrm{X}_n \mid \mu, \sigma^2\right)\right)\right]}{d\mu} = \sum_{i=1}^{n} \frac{1}{\sigma^2}\left(x^{(i)} - \mu\right) = 0 \Rightarrow \hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} x^{(i)}.$$

Máxima verosimilitud (*Maximum likelihood estimate MLE*)

Máximo de $p(\theta \mid \mathrm{X}_n)$

$$\hat{\theta}_n = \arg\max_{\theta} p(\mathrm{X}_n \mid \theta).$$

Ejemplo: modelo de distribución normal $\mathcal{N}(\mu, \sigma^2)$

$$p\left(\mathrm{X}_n \mid \mu, \sigma^2\right) = \prod_{i=1}^{n} p\left(x^{(i)} \mid \mu, \sigma^2\right) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\left(x^{(i)}-\mu\right)^2}{2\sigma^2}}.$$

- Maximizamos logaritmo con respecto a σ^2, derivando e igualando a 0:

$$\frac{d\left[\ln\left(p\left(\mathrm{X}_n \mid \mu, \sigma^2\right)\right)\right]}{d\sigma^2} = \sum_{i=1}^{n} \frac{d}{d\sigma^2}\left[-\frac{1}{2}\left(\ln\left(\sigma^2\right) + \frac{\left(x^{(i)} - \mu\right)^2}{\sigma^2}\right)\right]$$

$$= \sum_{i=1}^{n} -\frac{1}{2}\left(\frac{1}{\sigma^2} - \frac{\left(x^{(i)} - \mu\right)^2}{\left(\sigma^2\right)^2}\right) = 0 \Rightarrow \hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x^{(i)} - \mu\right)^2.$$

Máxima verosimilitud (*Maximum likelihood estimate MLE*)

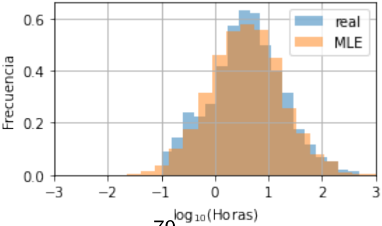
Máximo de $p(\theta \mid \mathrm{X}_n)$

$$\hat{\theta}_n = \arg\max_{\theta} p(\mathrm{X}_n \mid \theta).$$

Ejemplo: modelo de distribución normal $\mathcal{N}(\mu, \sigma^2)$

- Estimadores de máxima verosimilitud (*MLE*):

$$\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} x^{(i)} \text{ y } \hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x^{(i)} - \mu\right)^2.$$



Máxima verosimilitud (*Maximum likelihood estimate MLE*)

Máximo de $p(\theta \mid X_n)$

- En este enfoque, se ignora la distribución a priori.
- En caso de que θ es acotado, sería equivalente a una distribución uniforme a priori.

$$\hat{\theta}_n = \arg\max_{\theta} p(X_n \mid \theta).$$

Propiedades

- Es asintóticamente insesgado: $\lim_{n\to\infty} \mathbb{E}\left[\hat{\theta}_n\right] = \theta$.
- Es asintóticamente consistente: $\lim_{n\to\infty} P\left(\left|\hat{\theta}_n - \theta\right| > \epsilon\right) = 0$.
- Es *eficiente* (su varianza es la menor posible para un estimador insesgado. Buscar teorema de Cramér-Rao).
- Si existe un estadístico suficiente para θ, el estimador de máxima verosimilitud se puede expresar en base a ese estadístico.

Máxima verosimilitud (*Maximum likelihood estimate MLE*)

Ejemplo: distribución de Poisson

- Muestra $\left\{x^{(1)}, \ldots, x^{(n)}\right\}$ de tamaño n.
- Recordar: $f\left(x^{(i)}\right) = e^{-\lambda} \frac{\lambda^{x^{(i)}}}{x^{(i)}!}$.
- ¿Estimador $\hat{\lambda}_n^{\mathrm{MLE}}$ de λ?

Desarrollo

- Verosimilitud:

$$P(\mathrm{X}_n \mid \lambda) = \prod_{i=1}^{n} e^{-\lambda} \frac{\lambda^{x^{(i)}}}{x^{(i)}!} = e^{-\lambda n} \frac{\lambda^{\sum_{i=1}^{n} x^{(i)}}}{\prod_{i=1}^{n} x^{(i)}!}.$$

- Maximizamos derivando con respecto a λ e igualando a 0:

$$\left(-n + \frac{1}{\lambda} \sum_{i=1}^{n} x^{(i)}\right) e^{-\lambda n + \ln(\lambda) \sum_{i=1}^{n} x^{(i)}} = 0 \Rightarrow \hat{\lambda}_n^{\mathrm{MLE}} = \frac{1}{n} \sum_{i=1}^{n} x^{(i)}.$$

Máxima verosimilitud (*Maximum likelihood estimate MLE*)

Ejemplo: distribución Gamma

- Muestra $\left\{x^{(1)}, \dots, x^{(n)}\right\}$ de tamaño n.
- Recordar: $p\left(x^{(i)}\right) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{(i)\,\alpha-1} e^{-\beta x^{(i)}}$.
- ¿Estimador $\hat{\beta}_n^{\text{MLE}}$ de β, si α es conocido?

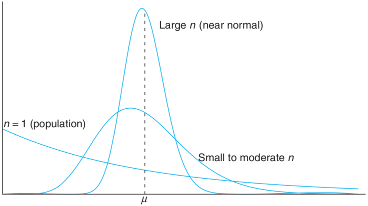
Desarrollo

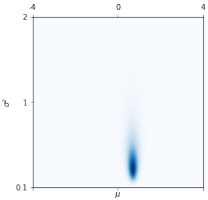
- Verosimilitud:
$$p(\mathrm{X}_n \mid \alpha, \beta) = \prod_{i=1}^{n} \frac{\beta^\alpha}{\Gamma(\alpha)} x^{(i)\,\alpha-1} e^{-\beta x^{(i)}} = \frac{\beta^{\alpha n} e^{-\beta \sum_{i=1}^{n} x^{(i)}}}{\Gamma(\alpha)^n} \prod_{i=1}^{n} x^{(i)\,\alpha-1}.$$
- Maximizamos derivando con respecto a β e igualando a 0:
$$\left(\frac{\alpha n}{\beta} - \sum_{i=1}^{n} x^{(i)}\right) e^{-\beta \sum_{i=1}^{n} x^{(i)} + \ln(\beta)\alpha n} = 0 \Rightarrow \hat{\beta}_n^{\text{MLE}} = \frac{\alpha n}{\sum_{i=1}^{n} x^{(i)}}.$$

Enfoque bayesiano

Distribución de θ

$$p(\theta \mid X_n) = \frac{p(X_n \mid \theta)p(\theta)}{p(X_n)}.$$



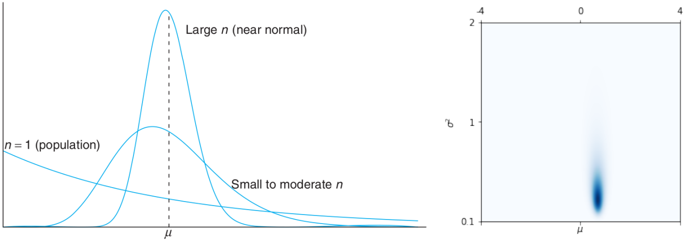


Una posibilidad: buscar el máximo de $p(\theta \mid X_n)$

$$\hat{\theta}_n = \arg\max_{\theta} p(\theta \mid X_n) = \arg\max_{\theta} \frac{p(X_n \mid \theta)p(\theta)}{p(X_n)}$$

$$= \arg\max_{\theta} p(X_n \mid \theta)p(\theta).$$

Máximo a posteriori (*Maximum a posteriori MAP*)



Máximo de $p(\theta \mid X_n)$

- En este enfoque, se debe definir un a priori.
- Como vimos, es más importante cuando n es pequeño.

$$\hat{\theta}_n = \arg\max_{\theta} p(X_n \mid \theta) p(\theta).$$

Máximo a posteriori (*Maximum a posteriori MAP*)

Máximo de $p(\theta \mid X_n)$

$$\hat{\theta}_n = \arg\max_{\theta} p(X_n \mid \theta) p(\theta).$$

Ejemplo: modelo de distribución normal $\mathcal{N}(\mu, \sigma^2)$ y prior $\mathcal{N}(0, 1)$

$$p\left(X_n \mid \mu, \sigma^2\right) p\left(\mu, \sigma^2\right) = p\left(\sigma^2\right) \frac{1}{\sqrt{2\pi}} e^{-\frac{\mu^2}{2}} \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\left(x^{(i)} - \mu\right)^2}{2\sigma^2}}.$$

- Supongamos σ^2 conocido, y estimemos μ:
$$p(X_n \mid \mu) p(\mu) = \frac{1}{\sqrt{2\pi}} e^{-\frac{\mu^2}{2}} \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\left(x^{(i)} - \mu\right)^2}{2\sigma^2}}.$$

Máximo a posteriori (*Maximum a posteriori MAP*)

Máximo de $p(\theta \mid \mathrm{X}_n)$

$$\hat{\theta}_n = \arg\max_{\theta} p(\mathrm{X}_n \mid \theta) p(\theta).$$

Ejemplo: modelo de distribución normal $\mathcal{N}(\mu, \sigma^2)$ y prior $\mathcal{N}(0,1)$

$$p(\mathrm{X}_n \mid \mu) p(\mu) = \frac{1}{\sqrt{2\pi}} e^{-\frac{\mu^2}{2}} \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\left(x^{(i)}-\mu\right)^2}{2\sigma^2}}.$$

- Maximizamos logaritmo con respecto a μ, derivando e igualando a 0:

$$\frac{d\left[\ln\left(p\left(\mathrm{X}_n \mid \mu, \sigma^2\right) p\left(\mu, \sigma^2\right)\right)\right]}{d\mu} = -\mu + \sum_{i=1}^{n} \frac{1}{\sigma^2}\left(x^{(i)} - \mu\right) = 0$$

$$\Rightarrow \hat{\mu}_n = \frac{1}{n+\sigma^2} \sum_{i=1}^{n} x^{(i)}.$$

Máximo a posteriori (*Maximum a posteriori MAP*)

Máximo de $p(\theta \mid \mathrm{X}_n)$

$$\hat{\theta}_n = \arg\max_{\theta} p(\mathrm{X}_n \mid \theta) p(\theta).$$

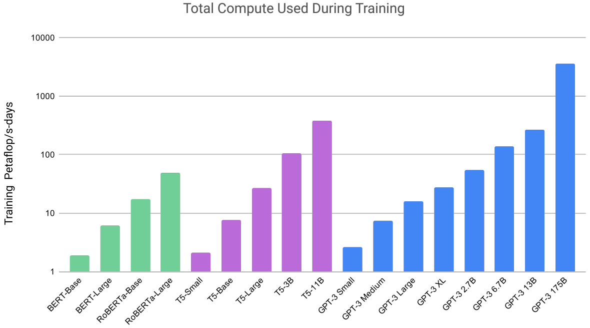
Ejemplo: modelo de distribución normal $\mathcal{N}(\mu, \sigma^2)$ y prior $\mathcal{N}(0, 1)$

- Suponiendo:
  - σ^2 conocido.
  - A priori $\mu \sim \mathcal{N}(0, 1)$.
- Entonces el estimador de máximo a posteriori (*MAP*) de μ es:

$$\hat{\mu}_n = \frac{1}{n + \sigma^2} \sum_{i=1}^{n} x^{(i)}.$$

Ejemplo: Modelo de lenguaje GPT-3

Model Name	n_{params}	n_{layers}	d_{model}	n_{heads}	d_{head}	Batch Size	Learning Rate
GPT-3 Small	125 M	12	768	12	64	0,5 M	$6{,}0 \times 10^{-4}$
GPT-3 Med.	350 M	24	1024	16	64	0,5 M	$3{,}0 \times 10^{-4}$
GPT-3 Large	760 M	24	1536	16	96	0,5 M	$2{,}5 \times 10^{-4}$
GPT-3 XL	1,3 B	24	2048	24	128	1 M	$2{,}0 \times 10^{-4}$
GPT-3 2.7B	2,7 B	32	2560	32	80	1 M	$1{,}6 \times 10^{-4}$
GPT-3 6.7B	6,7 B	32	4096	32	128	2 M	$1{,}2 \times 10^{-4}$
GPT-3 13B	13,0 B	40	5140	40	128	2 M	$1{,}0 \times 10^{-4}$
GPT-3 175B o "GPT-3"	175,0 B	96	12288	96	128	3,2 M	$0{,}6 \times 10^{-4}$



Aplicación: mezcla de Gaussianas

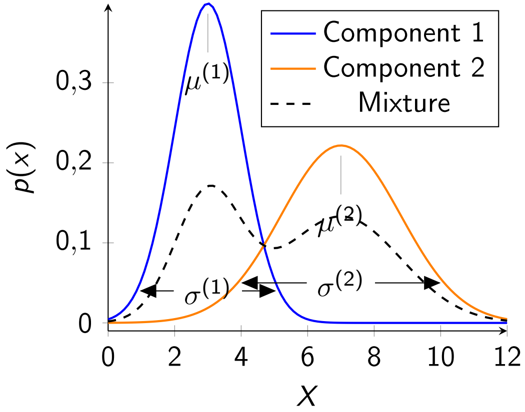
Ejemplo

- Se desea modelar una varibale aleatoria con un modelo de *mezcla de dos Gaussianas*:
$$p\left(x \mid \mu_1, \mu_2, \sigma^2\right) = \sum_{k=1}^{2} \pi_k \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu_k)^2}{2\sigma^2}}.$$
- $k = 1, 2$ indica la etiqueta de la Guassiana.
- Ambas tienen la misma varianza σ^2, pero distintas medias μ_1 y μ_2.
- Se supone una probabilidad a priori para cada clase $\pi_1 = \frac{1}{2}$ y $\pi_2 = \frac{1}{2}$.
- Se tiene una muestra iid de N datos, donde cada uno proviene de una Gaussiana.

Aplicación: mezcla de Gaussianas

Modelo

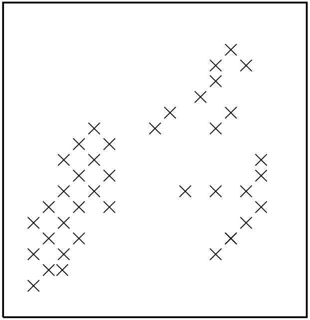
$$p\left(x \mid \mu_1, \mu_2, \sigma^2\right) = \sum_{k=1}^{2} \pi_k \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu_k)^2}{2\sigma^2}}.$$



Aplicación: mezcla de Gaussianas

Modelo

$$p(x \mid \mu_1, \mu_2, \sigma^2) = \sum_{k=1}^{2} \frac{1}{2} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu_k)^2}{2\sigma^2}}.$$



Aplicación: mezcla de Gaussianas

Modelo

$$p\left(x \mid \mu_1, \mu_2, \sigma^2\right) = \sum_{k=1}^{2} \frac{1}{2} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\left(x-\mu_k\right)^2}{2\sigma^2}}.$$

Posterior de clases

- Sea k_i la clase asociada al dato i.

$$p_1^{(i)} = p\left(k_i = 1 \mid x^{(i)}, \mu_1, \mu_2, \sigma^2\right)$$

$$= \frac{p\left(x^{(i)} \mid k_i = 1, \mu_1, \mu_2, \sigma^2\right) p(k_i = 1)}{p\left(x^{(i)} \mid \mu_1, \mu_2, \sigma^2\right)}$$

$$= \frac{\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\left(x-\mu_1\right)^2}{2\sigma^2}} \frac{1}{2}}{\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\left(x-\mu_1\right)^2}{2\sigma^2}} \frac{1}{2} + \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\left(x-\mu_2\right)^2}{2\sigma^2}} \frac{1}{2}} = \frac{e^{-\frac{\left(x-\mu_1\right)^2}{2\sigma^2}}}{e^{-\frac{\left(x-\mu_1\right)^2}{2\sigma^2}} + e^{-\frac{\left(x-\mu_2\right)^2}{2\sigma^2}}}$$

Aplicación: mezcla de Gaussianas

Modelo

$$p\left(x \mid \mu_1, \mu_2, \sigma^2\right) = \sum_{k=1}^{2} \frac{1}{2} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu_k)^2}{2\sigma^2}}.$$

Posterior de clases

$$p_1^{(i)} = \frac{e^{-\frac{(x-\mu_1)^2}{2\sigma^2}}}{e^{-\frac{(x-\mu_1)^2}{2\sigma^2}} + e^{-\frac{(x-\mu_2)^2}{2\sigma^2}}} = \frac{1}{1 + e^{\frac{2x^{(i)}(\mu_2-\mu_1)-\left(\mu_2^2-\mu_1^2\right)}{2\sigma^2}}},$$

$$p_2^{(i)} = \frac{1}{1 + e^{\frac{-2x^{(i)}(\mu_2-\mu_1)+\left(\mu_2^2-\mu_1^2\right)}{2\sigma^2}}}.$$

Aplicación: mezcla de Gaussianas

Modelo

$$p\left(x \mid \mu_1, \mu_2, \sigma^2\right) = \sum_{k=1}^{2} \frac{1}{2} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu_k)^2}{2\sigma^2}}.$$

Aplicación: mezcla de Gaussianas

Si no conocemos μ_1 y μ_2

- Los buscamos por MLE (máxima verosimilitud):

$$\frac{d}{d\mu_k} \ln \prod_{i=1}^{n} p\left(x^{(i)} \mid \mu_1, \mu_2, \sigma^2\right) = \sum_{i=1}^{n} \frac{d}{d\mu_k} \ln \left[\sum_{k'=1}^{2} \frac{1}{2} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\left(x^{(i)}-\mu_{k'}\right)^2}{2\sigma^2}} \right.$$

$$= \sum_{i=1}^{n} \frac{\frac{1}{2}\left(x^{(i)} - \mu_k\right) e^{-\frac{\left(x^{(i)}-\mu_k\right)^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}\sigma^2 \left[\frac{1}{2\sqrt{2\pi\sigma^2}} e^{-\frac{\left(x^{(i)}-\mu_1\right)^2}{2\sigma^2}} + \frac{1}{2\sqrt{2\pi\sigma^2}} e^{-\frac{\left(x^{(i)}-\mu_2\right)^2}{2\sigma^2}} \right]}$$

$$= \sum_{i=1}^{n} p_k^{(i)} \frac{\left(x^{(i)} - \mu_k\right)}{\sigma^2} = 0.$$

Aplicación: mezcla de Gaussianas

Modelo

$$p\left(x \mid \mu_1, \mu_2, \sigma^2\right) = \sum_{k=1}^{2} \frac{1}{2} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu_k)^2}{2\sigma^2}}.$$

Si no conocemos μ_1 y μ_2

- Tenemos que $\hat{\mu}_k = \frac{\sum_{i=1}^n p_k^{(i)} x^{(i)}}{\sum_{i=1}^n p_k^{(i)}}$, una media muestral "con pesos".
- Pero $p_k^{(i)}$ depende de μ_k.
- Se resuelve con algoritmo que actualiza $p_k^{(i)}$ y $\hat{\mu}_k$ iterativamente.

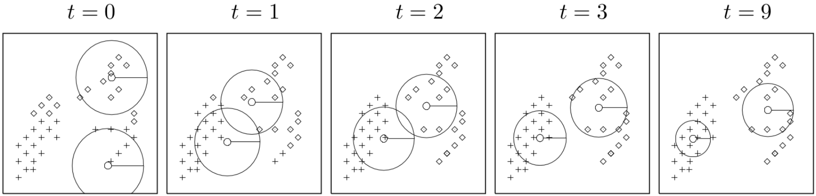
Ejercicio

- Verifique que la segunda derivada de la verosimilitud es negativa (es un máximo local).

Aplicación: mezcla de Gaussianas

Modelo

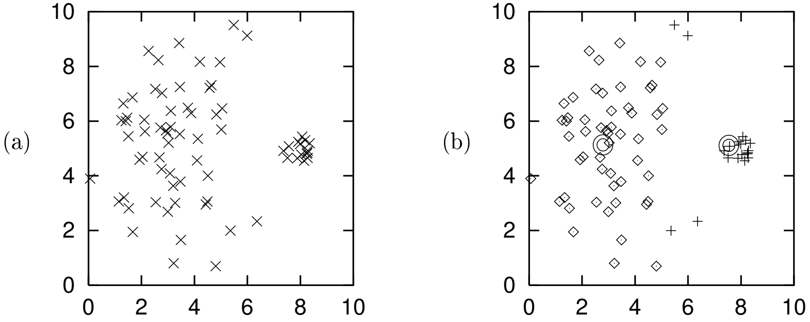
$$p\left(x \mid \mu_1, \mu_2, \sigma^2\right) = \sum_{k=1}^{2} \frac{1}{2} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu_k)^2}{2\sigma^2}}.$$



Aplicación: mezcla de Gaussianas

Modelo

$$p(x \mid \mu_1, \mu_2, \sigma^2) = \sum_{k=1}^{2} \frac{1}{2} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu_k)^2}{2\sigma^2}}.$$

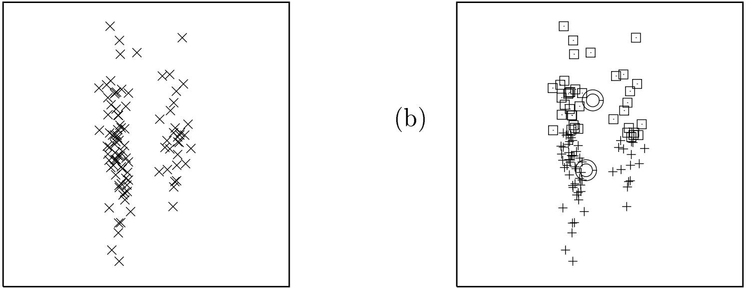


Aplicación: mezcla de Gaussianas

Modelo

$$p\left(x \mid \mu_1, \mu_2, \sigma^2\right) = \sum_{k=1}^{2} \frac{1}{2} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu_k)^2}{2\sigma^2}}.$$

(a) (b)



Aplicación: mezcla de Gaussianas

Modelo

$$p\left(x \mid \mu_1, \mu_2, \sigma_1^2, \sigma_2^2\right) = \sum_{k=1}^{2} \frac{1}{2} \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{-\frac{(x-\mu_k)^2}{2\sigma_k^2}}.$$