Análisis e Interpretación de Datos

MÁSTER UNIVERSITARIO EN ANÁLISIS Y VISUALIZACIÓN DE DATOS MASIVOS / VISUAL ANALYTICS AND BIG DATA

Miller Janny Ariza Garzón

Tema 6. Distribución en el muestreo

unir LA UNIVERSIDAD EN INTERNET

Universidad Internacional de La Rioja

Tabla de contenido

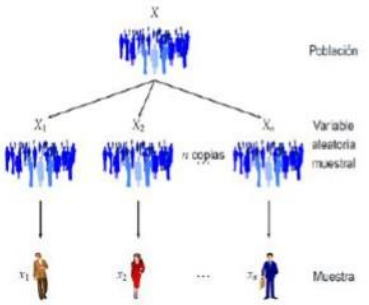
Contenido

Distribución en el muestreo

Distribución del conteo y la proporción muestral

$X \sim Bi(n,p)X$
$= \text{Número de éxitos}$

$$p = \frac{X}{np}$$
$= \text{Proporción de éxitos}$



Distribución del conteo y la proporción muestral

X v.a. con
$E(X) = \mu$ y $V(X) = \sigma$

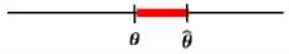
Muestra X_i independientes
$X_1, X_2, \ldots, Xn$

$$\bar{x} \sim N(\mu; \frac{\sigma}{\sqrt{n}})$$

Distribución de la media muestral

Tipos de estimación

Estimación puntual



Estimación por intervalos

Estimador ⇨ Parámetro

$\hat{p} \rightarrow p$
$\hat{\mu} \rightarrow \mu$
$\hat{\sigma} \rightarrow \sigma$

Propiedades de los estimadores

$Sesgo(\hat{\theta}) = E(\hat{\theta}) - \theta$

Insesgadez
$E(\hat{\theta}) = \theta$

Media
$E(\hat{\theta}) = \theta$

Proporción
$E(\bar{x}) = \mu$

Cuasivarianza
$E(s_c^2) = E\left(\frac{\sum (x_i - \bar{x})^2}{n-1}\right) = \sigma^2$

Una propiedad útil en muestreo

- Sean $X_1, X_2, \ldots, X_n$ un conjunto de variables aleatorias, entonces:

$$E\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} E(X_i)$$

$$Var\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} Var(X_i) + 2\sum_{i<j} Cov(X_i, X_j)$$

- Sean $X_1, X_2, \ldots, X_n$ un conjunto de variables aleatorias iid, independientes e igualmente distribuidas (muestra aleatoria), con media μ y varianza σ^2, entonces:

$$E\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} E(X_i) = n\mu$$

$$Var\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} Var(X_i) + 2\sum_{i<j} Cov(X_i, X_j)$$

$$= n\sigma^2$$

Estimador vs Parámetro

Parámetro:

Valor numérico que describe una característica poblacional o de la distribución de una variable aleatoria.
Ej.
$X \sim N(\mu, \sigma^2)$; μ y σ^2 son los parámetros
$X \sim B(n, p)$; p es un parámetro.

Estimador:

Función de variables aleatorias que describe características medidas en una muestra buscando estimar un parámetro.

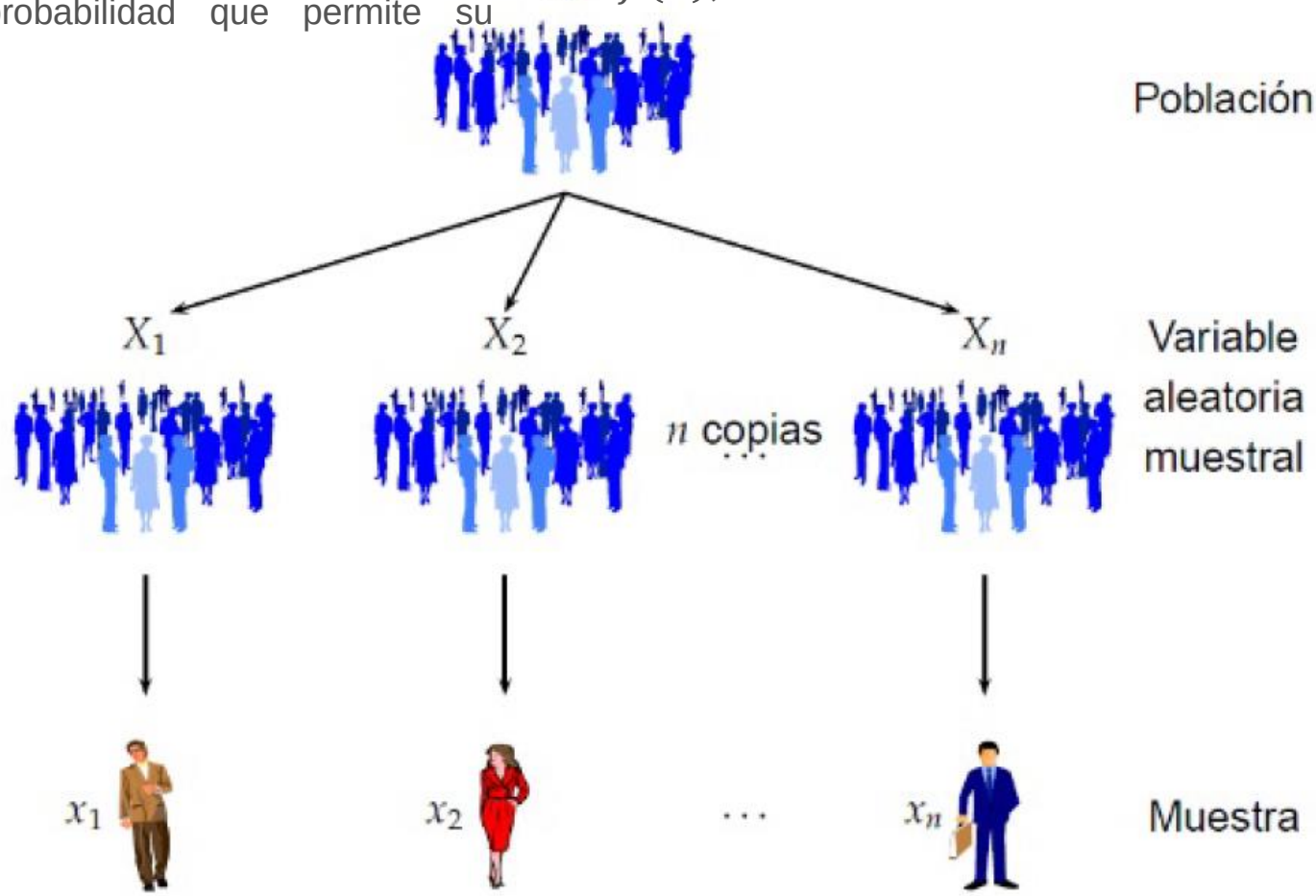
Para diferentes fenómenos pueden ser desconocidos.

Parámetro	Estimador
• μ	• $\bar{X}$
• P	• $\hat{P}$
• σ	• s

Muestra aleatoria y distribución muestral

Describe diferentes fenómenos y tiene una función de probabilidad que permite su modelación

$X \sim f(\theta);\ \theta\ desconocido$



¿Qué distribución tiene $\hat{\theta} = h(X_1, X_2, \dots, X_n)$ -el estimador- que permite estimar θ?

Ej. $\bar{X} = h(X_1, X_2, \dots, X_n) = \frac{\sum_{i=1}^{n} X_i}{n} \sim$? ← **Distribución muestral,** necesaria para estimar $\boldsymbol{\mu}$

Distribución muestral de $\hat{p}$

DISTRIBUCIÓN BINOMIAL

Función de Masa de Probabilidad:

- P(X) = Probabilidad de éxito dados los parámetros n y p.
- n = Tamaño de la muestra
- p = Probabilidad de Éxito
- 1 – p = Probabilidad de fracaso
- X = Número de éxitos en n ensayos independientes

$$P(X) = \binom{n}{X} p^X (1-p)^{n-X}$$

- Parámetros Poblacionales

$$E(X) = np$$
$$Var(X) = np(1-p)$$

Media y varianza depende de p, parámetro desconocido.

Distribución muestral de $\hat{p}$

Estimador de p:

$$\hat{p} = \frac{\sum_{i=1}^{n} X_i}{n} \qquad X_i = \begin{cases} 1 \; si \; se \; da \; un \; éxito \\ 0 \; en \; otro \; caso \end{cases}$$

Para una muestra aleatoria $X_1, X_2, \ldots, X_n$ donde cada $X_i \sim B(1, p)$

$$E(X_i) = 1 * p = p$$
$$V(X_i) = 1 * p = p(1-p)$$

Luego:

$$E(\hat{p}) = E\left(\frac{\sum_{i=1}^{n} X_i}{n}\right) = p$$
$$V(\hat{p}) = \left(\frac{\sum_{i=1}^{n} X_i}{n}\right) = \frac{p(1-p)}{n}$$

$\sum_{i=1}^{n} X_i \sim B(n, p)$ y $\hat{p} = \frac{\sum_{i=1}^{n} X_i}{n} \sim ?$ ¿Podrá ser normal?

Distribución muestral de $\hat{\mu} = \bar{X}$

$X \sim N(\mu, \sigma^2)$

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Para una muestra aleatoria $X_1, X_2, \ldots, X_n$ donde cada $X_i \sim N(\mu, \sigma^2)$

$$E(X_i) = \mu$$
$$V(X_i) = \sigma^2$$

Luego:

$$E(\bar{X}) = E\left(\frac{\sum_{i=1}^{n} X_i}{n}\right) = \mu$$
$$V(\bar{X}) = \left(\frac{\sum_{i=1}^{n} X_i}{n}\right) = \frac{\sigma^2}{n}$$

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

$Si\ X \sim ?$

$\bar{X} \sim ?$

¿Podrá ser normal?

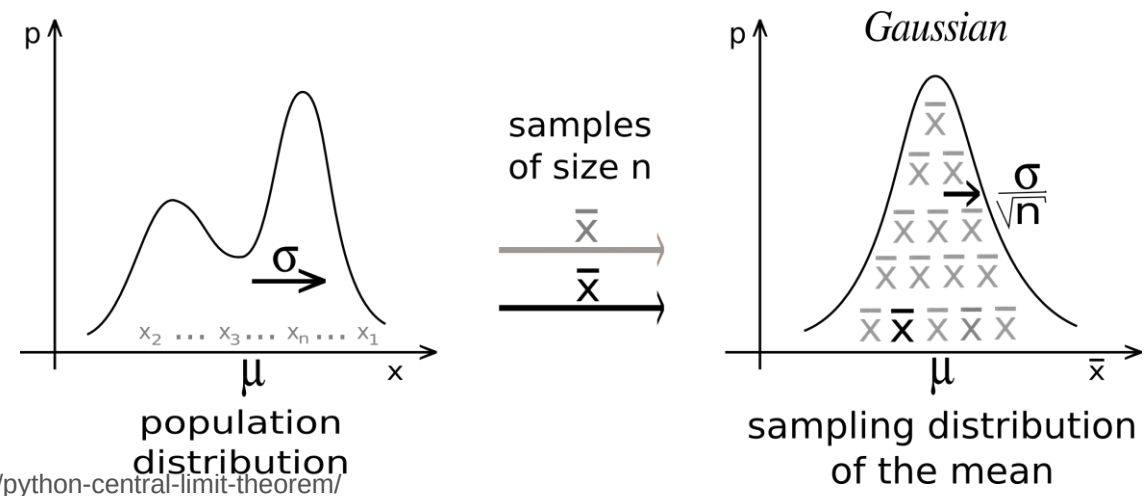
Teorema del límite central (TLC)

Afirma que cuando tenemos n variables $X_1, X_2, \ldots, X_n$ iid (con n suficientemente grande), con media μ y varianza σ^2, su suma $X_1 + X_2 + \cdots + X_n$ es una variable aleatoria que se distribuye aproximadamente como una normal. Esta aproximación será mejor cuanto mayor sea n.

Si n es grande:

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

$$X \sim ?, n\ grande\ (n \geq 30), \bar{X} \sim N$$



https://www.geeksforgeeks.org/python-central-limit-theorem/

Teorema del límite central (TLC)

Consecuencias (n grande):

- Para una muestra aleatoria $X_1, X_2, \dots, X_n$ donde cada $X_i \sim B(1,p)$

$$\hat{p} = \frac{\sum_{i=1}^{n} X_i}{n} \sim \cancel{?}$$

$$X_i = \begin{cases} 1 \; si \; se \; da \; un \; éxito \\ 0 \; en \; otro \; caso \end{cases}$$

$$\hat{p} = \frac{\sum_{i=1}^{n} X_i}{n} \sim N\left(p, \frac{p(1-p)}{n}\right)$$

- Para una muestra aleatoria $X_1, X_2, \dots, X_n$ donde cada X_i

$$E(X_i) = \mu$$
$$V(X_i) = \sigma^2$$

$$Si \; X \sim ? \longrightarrow \bar{X} \sim \cancel{?}$$

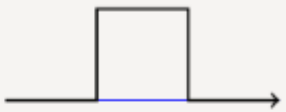
$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

Teorema del límite central (TLC)

$X_1, X_2 \ldots$ are iid Uniform(0,1)

$$Z_1 = \frac{X_1 - \frac{1}{2}}{\sqrt{\frac{1}{12}}}$$

PDF of Z_1



$$Z_2 = \frac{X_1 + X_2 - 1}{\sqrt{\frac{2}{12}}}$$

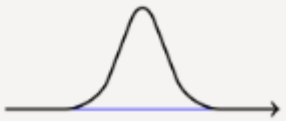
PDF of Z_2

$$Z_3 = \frac{X_1 + X_2 + X_3 - \frac{3}{2}}{\sqrt{\frac{3}{12}}}$$

PDF of Z_3

$$Z_{30} = \frac{\sum_{i=1}^{30} X_i - \frac{30}{2}}{\sqrt{\frac{30}{12}}}$$

PDF of Z_{30}



$X_1, X_2 \ldots$ are iid Bernoulli(p)

$$Z_1 = \frac{X_1 - p}{\sqrt{p(1-p)}}$$

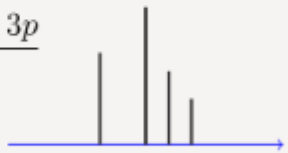
PMF of Z_1

$$Z_2 = \frac{X_1 + X_2 - 2p}{\sqrt{2p(1-p)}}$$

PMF of Z_2

$$Z_3 = \frac{X_1 + X_2 + X_3 - 3p}{\sqrt{3p(1-p)}}$$

PMF of Z_3

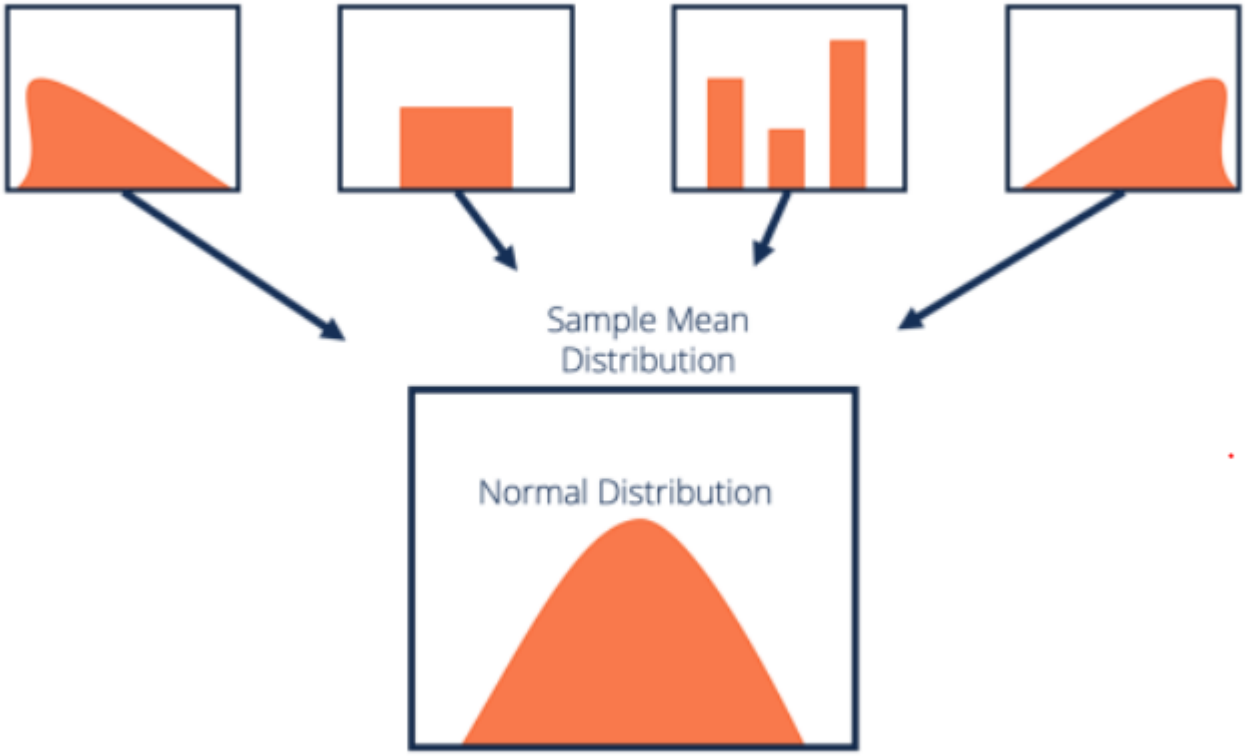


$$Z_{30} = \frac{\sum_{i=1}^{30} X_i - 30p}{\sqrt{30p(1-p)}}$$

PMF of Z_{30}

https://www.probabilitycourse.com/chapter7/7_1_2_central_limit_theorem.php

Teorema del límite central (TLC)



https://www.analyticsvidhya.com/blog/2019/05/statistics-101-introduction-central-limit-theorem/

Teorema del límite central (TLC)

Ejemplo 3: Una votación independentista Imaginemos (aunque se trata de un tema perfectamente posible) que un conocido periódico catalán desea interrogar a sus lectores por medio de una encuesta sobre si están a favor o no de ser independientes de España. Supongamos ahora que en este periódico se asume que el 70% de sus lectores tiene tendencia independentista y por tanto votarían a favor de esta. ¿Cuál sería entonces la probabilidad qué con una muestra aleatoria de 500 lectores se alcance al menos una cifra del 65% a favor de la independencia o mayor?

$$X_i \sim B(1, p = 0.7)$$

$$n = 500$$

$$\hat{p} = \frac{\sum_{i=1}^{n} X_i}{n} \sim N\left(0.7, \frac{0.7(1-0.7)}{500}\right)$$

$$P(\hat{p} \geq 0.65) = 1 - F\left(\frac{0.65 - 0.7}{\sqrt{\frac{0.7(1-0.7)}{500}}}\right) = 1 - \Phi\left(\frac{0.65 - 0.7}{\sqrt{\frac{0.7(1-0.7)}{500}}}\right) \cong 0.99$$

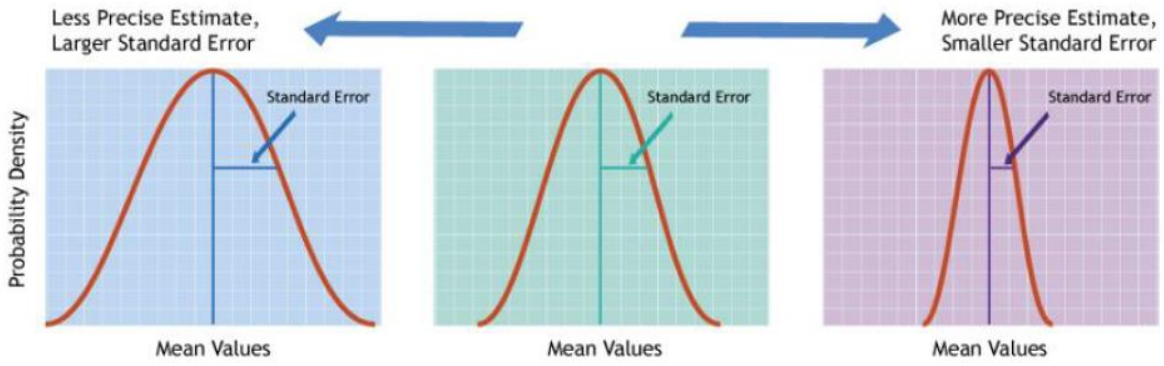
Teorema del límite central (TLC)

- Para una muestra aleatoria $X_1, X_2, \ldots, X_n$ donde cada X_i

$$E(X_i) = \mu$$
$$V(X_i) = \sigma^2$$

$$Si\ X \sim ? \longrightarrow \bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

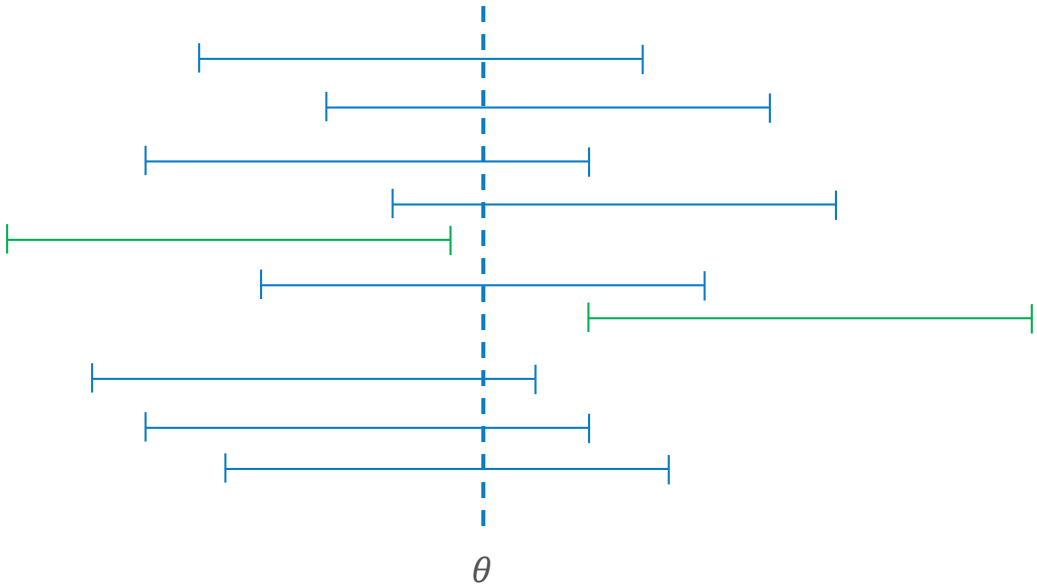
Su raíz, equivale al error típico o error estándar. Se disminuye cuando el tamaño de muestra crece.



Estimación por intervalo

$\widehat{\theta} \longrightarrow \theta$ Estimador Puntual

$\hat{\theta} \pm \mathrm{Q} * \text{Error típico de } \hat{\theta}$



Estimador por intervalo

Si se extraen repetida e independientemente muestras aleatorias de n observaciones de la población y se calculan intervalos de confianza al (1-α)%, entonces en el (1-α)% de estos intervalos contendrá el verdadero valor del parámetro.

confianza representa el porcentaje de intervalos que incluirían el parámetro de población si se tomara muestras de la misma población una y otra vez. **Confianza valida la formula**

Estimación por intervalo

Intervalo para μ:

- Para una muestra aleatoria $X_1, X_2, \dots, X_n$ donde cada X_i

$$E(X_i) = \mu$$
$$V(X_i) = \sigma^2$$

$$\downarrow$$

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

$$\downarrow$$

$$\bar{x} \pm z_{1-\alpha/2} * \sigma/\sqrt{n}$$

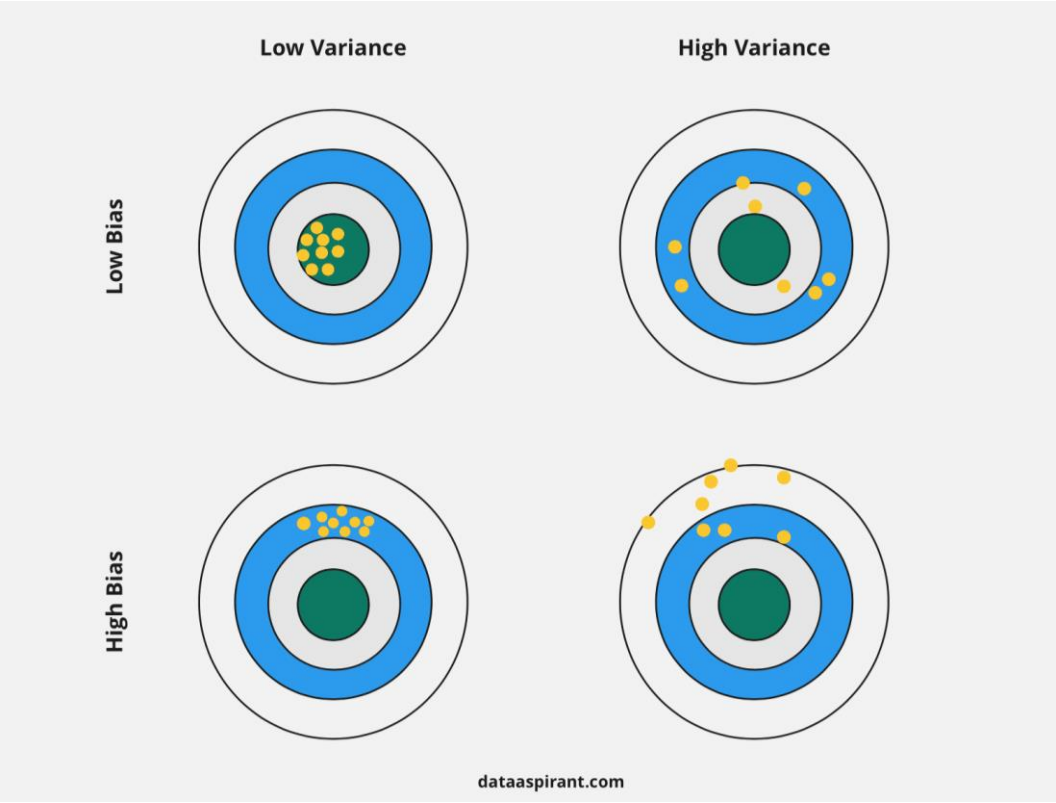
Intervalo de confianza para la verdadera media μ

Ej. Queremos estudiar cuánto tarda (tiempo) el COVID-19 en ser sintomático en la población española. Elegir muestra representativa, determinar qué parámetro podemos estudiar (media casos confirmados). Lo podemos calcular en una muestra y extender a la población.

Propiedades de un buen estimador $\widehat{\theta}$

Insesgado: $sesgo(\hat{\theta}) = E(\hat{\theta}) - \hat{\theta} = 0, \quad si\ E(\hat{\theta}) = \theta$

Eficiente: $Var(\hat{\theta}) < Var(\hat{\theta}_i)$ (Varianza mínima)



Propiedades de un buen estimador $\widehat{\theta}$

- $\bar{X}$ es un estimador insesgado y eficente de μ

(UMVUE)

- $\hat{p}$ es un estimador insesgado y eficente de p

(UMVUE)

¿Cuál será un buen estimador de σ^2?

Estimador de la varianza σ^2

Varianza muestral

$$s^2 = \frac{\sum(x_i - \bar{x})^2}{n}$$

Es un estimador sesgado

$$E(s^2) = E\left(\frac{\sum (x_i - \bar{x})^2}{n}\right) \neq \sigma^2$$

Cuasivarianza

$$s_c^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$$

Es un estimador insesgado

$$E(s_c^2) = E\left(\frac{\sum (x_i - \bar{x})^2}{n-1}\right) = \sigma^2$$

Hemos de usar la cuasivarianza de cara a los análisis de Estadística Inferencial

Próxima sesión

Tema 7: Intervalos de confianza

- Distribución en el muestreo del conteo y la proporción muestral
- Introducción a los intervalos de confianza
- Intervalo de confianza para la media de una población normal: varianza conocida y desconocida
- Calculando el tamaño de la muestra
- Intervalo de confianza para la proporción
- Intervalo de confianza para la varianza de una población normal
- Intervalo de confianza para la diferencia de medias y proporciones
- Intervalos de confianza robustos

www.unir.net