

Inferencia Estadística I

Introducción a la Estadística

Cristina Figueroa Sisniega

DVM

June 3, 2020

Contenidos

- 1 Inferencia estadística
- 2 Conceptos básicos
- 3 Muestreo
- 4 Teorema Central del límite
- 5 Estimación puntual
- 6 Dos distribuciones importantes
- 7 Estimación por intervalos

¿Qué es la inferencia estadística?



Tipos de inferencia

Veremos tres maneras de inferir información d la población desde la muestra :

- Estimación Puntual
- Estimación por intervalos
- Test de Hipótesis

A continuación veremos una introducción a la naturaleza de la inferencia en cada una de ellas, sus diferencias y el tipo de conclusiones que de ellas extraemos sobre la población basandonos en los datos muestrales.

Tipos de inferencia

- Estimación Puntual: estimación de un parámetro desconocido de la población usando un estadístico muestral, e.g., a partir de los datos, estimamos la proporción de españoles que están a favor de alargar las medidas de confinamiento absoluto)
- Estimación por intervalos: estimamos un parámetro desconocido usando un intervalo de valores que probablemente contenga el valor verdadero del parámetro poblacional (y también establecemos cómo de seguros estamos de que este intervalo captura el valor verdadero del parámetro). e.g. a partir de los datos de la muestra, afirmamos con una confianza del 95% que la proporción de españoles que están a favor de alargar las medidas de confinamiento absoluto está entre 0,57 y 0,63.

Tipos de inferencia

- Test de Hipótesis: establecemos una afirmación sobre un parámetro de la población (H_0) y verificamos si hay evidencia en contra de esta afirmación (a favor de la alternativa H_1) en la muestra.

Contenidos

- 1 Inferencia estadística
- 2 Conceptos básicos
- 3 Muestreo
- 4 Teorema Central del límite
- 5 Estimación puntual
- 6 Dos distribuciones importantes
- 7 Estimación por intervalos

Conceptos básicos

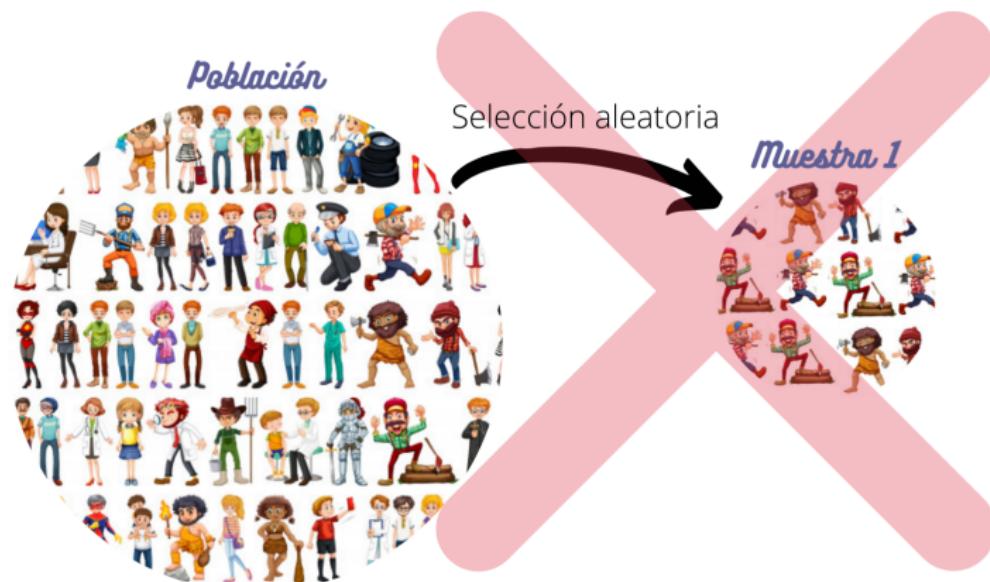
- Población: “Conjunto de elementos en los que se observa alguna característica común”
- Observaciones: “Valores que toma la característica observada en cada elemento de la población”
- Muestra: “Conjunto de unidades representativas de una población”. Para poder usarla objeto de un estudio de inferencia estadística, esta debe de ser aleatoria.
- Parámetro: “Característica numérica que describe una variable observada en la población”
- Estadístico: función de los valores de la muestra
- Estimación: Método mediante el cual tratamos de aproximarnos a los verdaderos valores de los parámetros de la distribución en población. Tenemos dos grandes grupos de estimación.
 - ▶ Métodos de estimación puntual
 - ▶ Métodos de estimación por intervalos.

Contenidos

- 1 Inferencia estadística
- 2 Conceptos básicos
- 3 Muestreo
- 4 Teorema Central del límite
- 5 Estimación puntual
- 6 Dos distribuciones importantes
- 7 Estimación por intervalos

Muestreo

Las técnicas de muestreo son métodos para obtener miembros de una población. El principal requisito de una muestra es que sea aleatoria, si no, esta y las inferencias estarán sesgadas de alguna manera.



Muestreo

- **Muestreo por conveniencia:** Se elige una muestra que ya está disponible de manera no aleatoria. *e.g.*, encuestas en la calle.



- **Muestreo voluntario:** se pide a los miembros de la población la participación voluntaria y son estos los que deciden formar parte de la muestra. *e.g.*, un presentador de TV pide a la audiencia que respondan a unas preguntas.

- ▶ Sesgo de auto-selección: la decisión de participar está correlacionada con el objetivo de estudio.



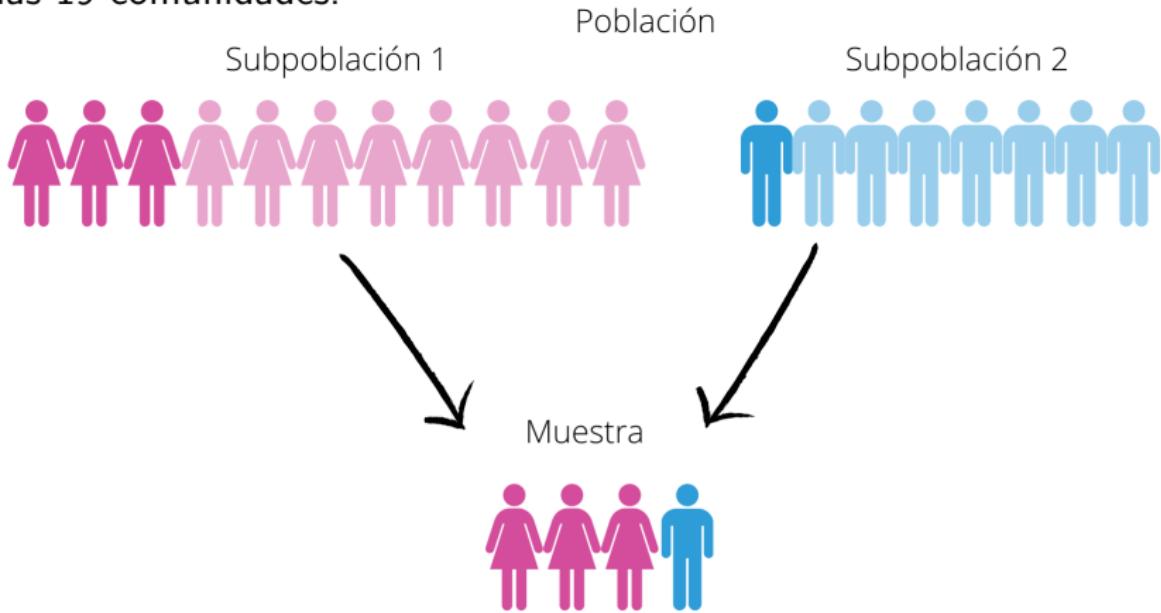
Muestreo

- **Muestreo aleatorio simple:** La presencia de los diferentes miembros de la población en la muestra es equiprobable . e.g. seleccionar estudiantes si su nombre es extraido del sombrero.

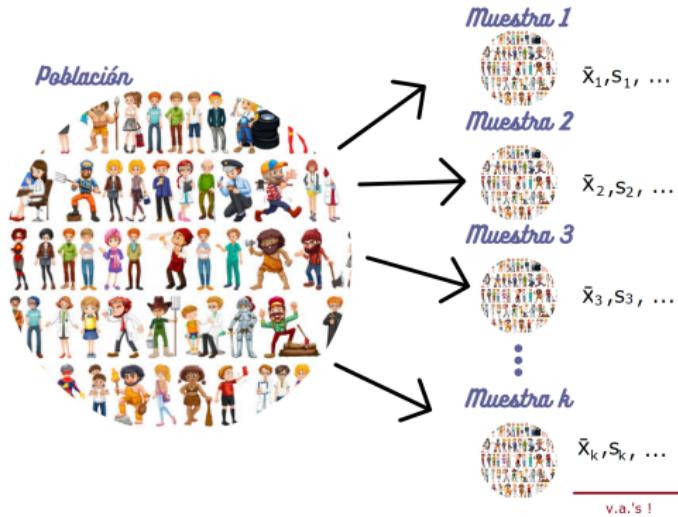


Muestreo

- **Muestreo Estratificado:** la población se divide en grupos. La muestra consiste en una selección aleatoria de miembros para cada uno de los grupos. e.g. el gobierno hace una encuesta a 38000 personas formada por 2000 personas elegidas al azar de cada una de las 19 comunidades.



Distribucion muestral

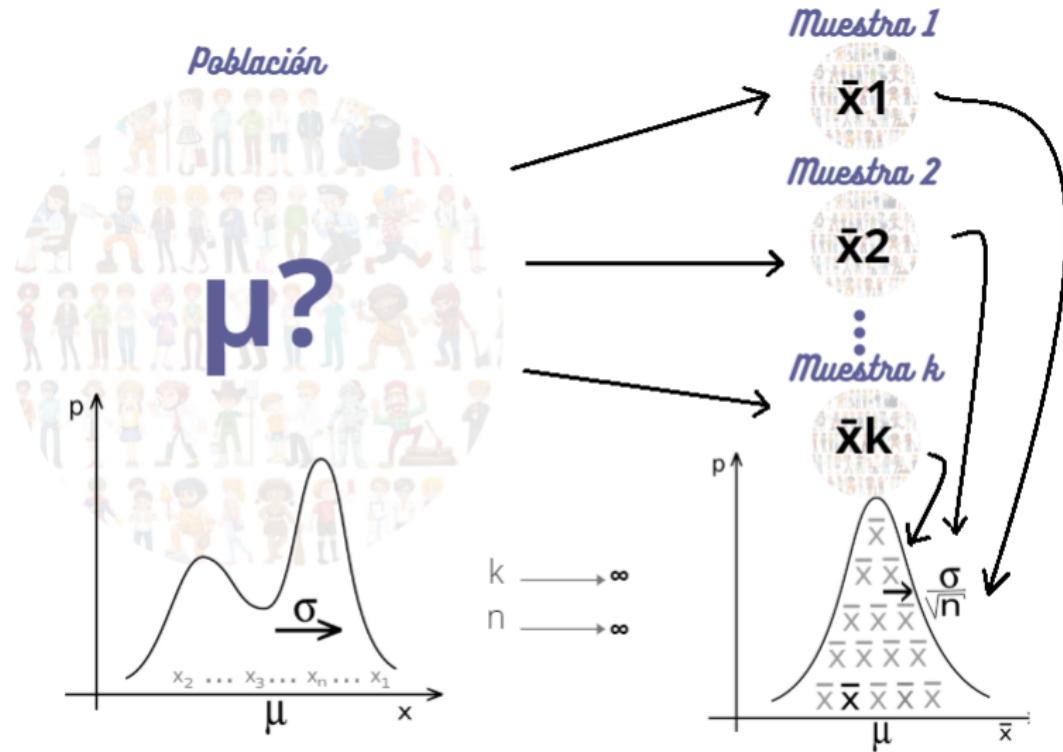


- Media muestral:
$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$$
- Varianza muestral:
$$S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}$$
- Desviación Típica muestral:
$$S = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}}$$

Contenidos

- 1 Inferencia estadística
- 2 Conceptos básicos
- 3 Muestreo
- 4 Teorema Central del límite
- 5 Estimación puntual
- 6 Dos distribuciones importantes
- 7 Estimación por intervalos

Teorema Central del límite



Teorema Central del límite

| POBLACIÓN | Condiciones para la aproximación a la Normal | Aproximación a la normal | Estimador \bar{X} o p Con $n > 30$ Aplicamos T.C.L Distribución muestral |
|--|--|--|---|
| Binomial $X \rightarrow B(n, p)$ | $n > 30$ si $n \cdot p \cdot q > 5$ | $X \rightarrow N((n \cdot p); \sqrt{n \cdot p \cdot q})$ | $\hat{p} \rightarrow N\left(p; \sqrt{\frac{p \cdot q}{n}}\right)$ |
| Poisson $X \rightarrow P(\lambda)$ | $\lambda > 5$ | $X \rightarrow N(\lambda; \sqrt{\lambda})$ | $\bar{X} \rightarrow N\left(\lambda; \sqrt{\frac{\lambda}{n}}\right)$ |
| Normal $X \rightarrow N(\mu; \sigma)$ | | $X \rightarrow N(\mu; \sigma)$ | $\bar{X} \rightarrow N\left(\mu; \sqrt{\frac{\sigma^2}{n}}\right)$ |

Contenidos

- 1 Inferencia estadística
- 2 Conceptos básicos
- 3 Muestreo
- 4 Teorema Central del límite
- 5 Estimación puntual
- 6 Dos distribuciones importantes
- 7 Estimación por intervalos

Estimación Puntual

Un estimador puntual $\hat{\theta}$, usa un estadístico (valor función de los datos muestrales) para estimar un parámetro desconocido de la población θ .

- La media de la población se puede estimar puntualmente mediante la media de la muestra: $\bar{x} = \mu$
- La proporción de la población se puede estimar puntualmente mediante la proporción de la muestra: $\hat{p} = p$
- La desviación típica de la población se puede estimar puntualmente mediante la desviación típica de la muestra: $s = \sigma$

Los estimadores son v.a.'s y tienen una distribución de probabilidad que corresponde a las distribuciones muestrales.

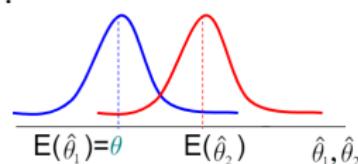
Estimación Puntual: Propiedades de un estimador

No todos los estimadores son apropiados. Para evaluar su calidad, podemos comparar su Error Cuadrático Medio

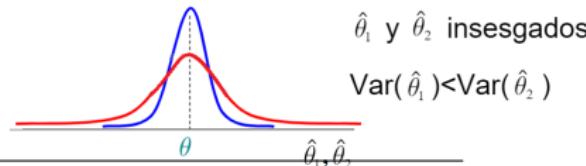
$$ECM(\hat{\theta}) = \text{Var}(\hat{\theta}) + (E[\hat{\theta}] - \theta)^2.$$

Esto supone tener un estimador:

- a) Insesgado (Sesgo): El estimador es "insesgado" si la media de su distribución es igual al parámetro, es decir, $E[\hat{\theta}] = \theta$ ¹



- b) Eficiente (Minima varianza): Entre dos estimadores insesgados preferimos el que tenga menos varianza.



¹Estimadores insesgados son: \bar{x} , s^2

Estimación Puntual: Construcción de un estimador

Para construir los estimadores hay dos métodos principales:

- a) Método de los momentos
- b) Método de máxima verosimilitud

Estimación Puntual: Método de los momentos

Consideremos: una v.a. X y $k \geq 1$, el k -ésimo momento de X es el número $E(X^k)$. Se denominan **momentos poblacionales** y el primer y el segundo momento son:

$$E(X)$$

$$E(X^2)$$

Por otro lado, consideremos una m.a.s. de la distribución $f(X|\theta)$ con θ un parámetro (o vector) desconocido. El k -ésimo **momento muestral** es la variable aleatoria:

$$\frac{1}{n} \sum_i^n x_i^k$$

El método de los momentos consiste en igualar los momentos poblacionales a los muestrales para estimar los primeros:

$$E(X) = \frac{1}{n} \sum_i^n x_i$$

$$E(X^2) = \frac{1}{n} \sum_i^n x_i^2$$

$$E(X^k) = \frac{1}{n} \sum_i^n x_i^k$$

Estimación Puntual: Método de los momentos

Ejemplos

- 1 Tenemos una muestra x_1, \dots, x_n de una v.a. X con distribución Poisson $P(\lambda)$ donde $\lambda > 0$ y desconocido. Se nos pide estimarla usando el método de momentos.

Sabemos que el primer momento poblacional: $E(X) = \lambda$ y el primer momento muestral: $\frac{1}{n} \sum_i^n x_i$

Así, $E(X) = \frac{1}{n} \sum_i^n x_i$ y entonces $\hat{\lambda} = \bar{X}$

- 2 Tenemos una muestra x_1, \dots, x_n de una v.a. X con distribución normal $N(\mu, \sigma^2)$ Se pide estimar μ y σ^2 con el método de los momentos.

Sabemos que el 1er y el 2ndo momento poblacional: $E(X) = \mu$ y $E(X^2) = \sigma^2 + \mu^2$

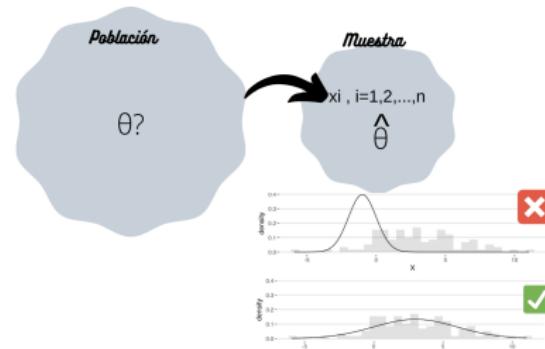
Sabemos que el 1er y el 2ndo momento muestral: $\frac{1}{n} \sum_i^n x_i$ y $\frac{1}{n} \sum_i^n x_i^2$, respectivamente. Así que:

$$\hat{\mu} = \bar{X}$$

$$\hat{\sigma}^2 + \mu^2 = \frac{1}{n} \sum_i^n x_i^2$$

Estimación Puntual: Método de máxima verosimilitud(MLE)

Teniendo θ (el parámetro poblacional desconocido) y una m.a.s con observaciones (independientes e identicamente distribuidas) x_1, x_2, \dots, x_n extraídas de una función de distribución $f(X|\theta)$. Nos preguntamos: cuál es el valor más verosímil para θ dada esa muestra($\hat{\theta}$)?



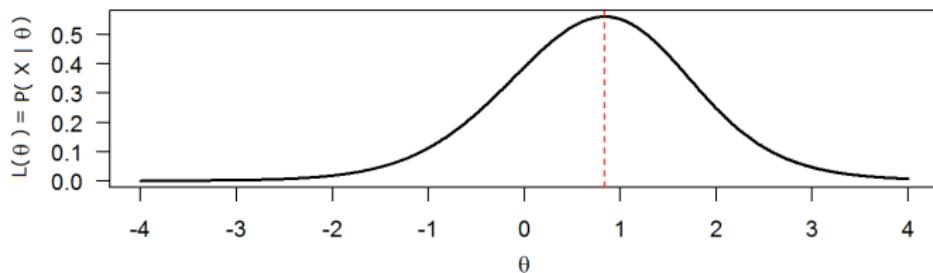
La función de densidad conjunta(pmf, pdf) de todas las observaciones es:
 $f(x_1, x_2, \dots, x_n | \theta) = f(x_1 | \theta) * f(x_2 | \theta) * \dots * f(x_n | \theta)$ (independencia)

Estimación Puntual: Método de máxima verosimilitud(MLE)

No sabemos θ pero evaluamos una función de verosimilitud donde tendremos diferentes valores θ para los datos muestrales dados:

$$L(\theta|x_1, \dots, x_n) = \prod_{i=1}^n f(x_i|\theta)^2$$

Maximizamos la función para encontrar $\hat{\theta}$. Este es el llamado estimador de máxima verosimilitud (MLE).



Intuición: se selecciona el valor de θ que hace que los datos observados sean lo más probable.

²En la práctica, se suele utilizar el logaritmo de esta función:

$$\hat{\ell}(\theta|x_1, \dots, x_n) = \ln L = \sum_{i=1}^n \ln f(x_i|\theta)$$

Estimación Puntual: Método de máxima verosimilitud

Ejemplo

Veamos un ejemplo en R.³ Tenemos los siguientes datos sobre el número de giros a la derecha durante 300 intervalos de 3 minutos cada uno en una intersección :

| x | freq |
|-----|------|
| 0 | 14 |
| 1 | 30 |
| 2 | 36 |
| 3 | 68 |
| 4 | 43 |
| 5 | 43 |
| 6 | 30 |
| 7 | 14 |
| 8 | 10 |
| 9 | 6 |
| 10 | 4 |
| 11 | 1 |
| 12 | 1 |
| +13 | 0 |

Si suponemos que esta muestra procede de una población con distribución Poisson, lo que intentaríamos estimar es el parámetro λ . Recordemos que : $P(X = x) = \frac{e^{-\lambda} * \lambda^x}{x!}$

³<http://people.missouristate.edu/songfengzheng/Teaching/MTH541/MLE-R.pdf>

Estimación Puntual: Método de máxima verosimilitud

Ejemplo

```
1 X <- c(rep(0,14), rep(1,30), rep(2,36), rep(3,68), rep(4, 43),
      , rep(5,43), rep(6, 30), rep(7,14), rep(8,10), rep(9, 6),
      rep(10,4), rep(11,1), rep(12,1))
2 hist(X, main="histograma del numero de giros a la derecha",
      right=FALSE, prob=TRUE)
3 n <- length(X)
4 negloglike <-function(lam) { n* lam -sum(X) *log(lam) + sum(
      log(factorial(X))) }
5 negloglike(0.3)
6 out <- nlm(negloglike, p=c(0.5))
7 out
8 mean(X)
```

Contenidos

- 1 Inferencia estadística
- 2 Conceptos básicos
- 3 Muestreo
- 4 Teorema Central del límite
- 5 Estimación puntual
- 6 Dos distribuciones importantes
- 7 Estimación por intervalos

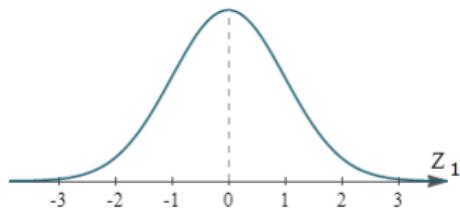
Dos distribuciones importantes

- ① Distribución Chi-cuadrado
- ② Distribución t-student

Distribución chi cuadrado

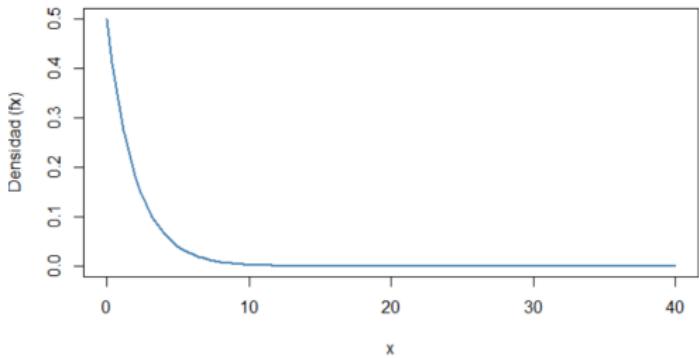
Puede considerarse como el cuadrado de una selección de una normal estandar.

Consideremos Z_1



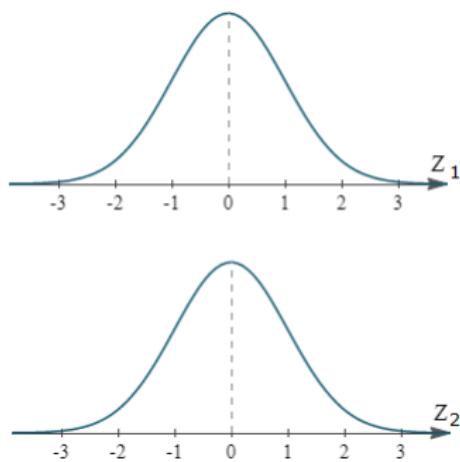
Consideremos $Q = Z_1^2 \rightarrow \chi_1^2$

Distribución Chi-cuadrado (df = 1)

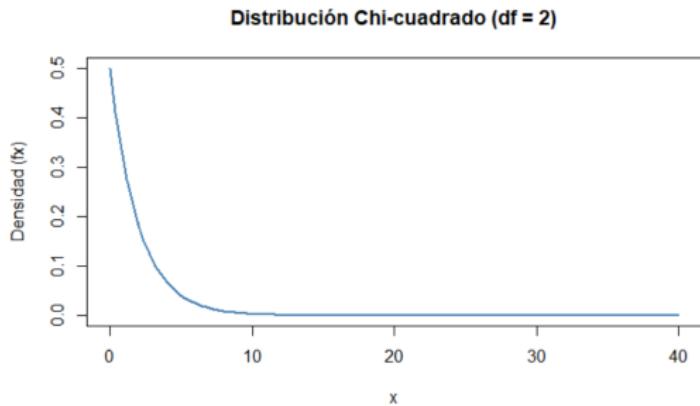


Distribución chi cuadrado

Consideremos Z_1 y Z_2

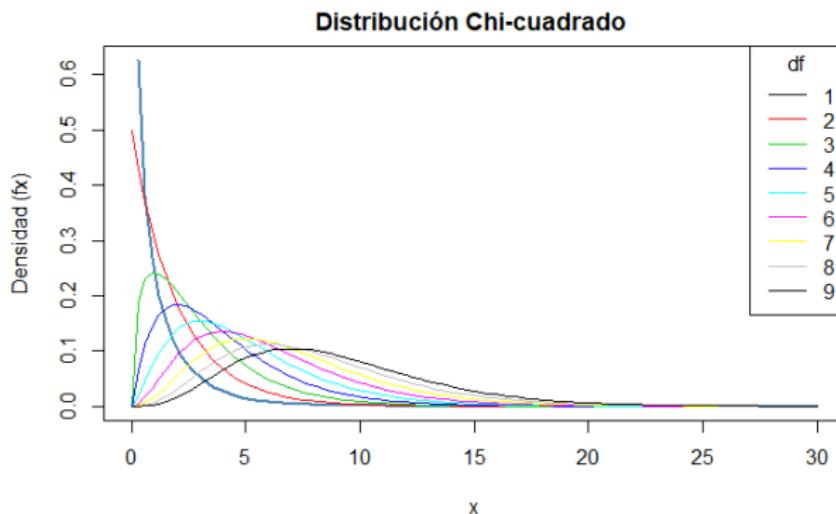


Consideremos $Q = Z_1^2 + Z_2^2 \rightarrow \chi^2$



Distribución chi cuadrado

En general: $\sum_{i=1}^k Z_i^2 \rightarrow \chi_k^2$

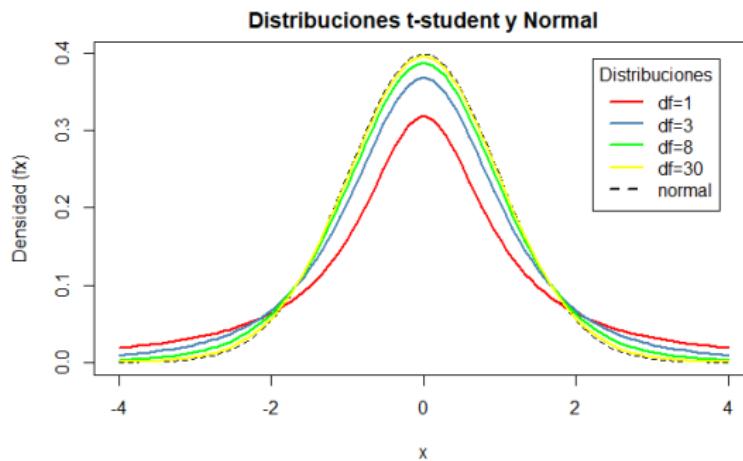


- Como vemos, su función de distribución depende de los grados de libertad (df)
- Su media es k y su varianza es $2k$.
- El estadístico chi: $X_k^2 = \frac{(n-1)S^2}{\sigma^2}$. Lo usaremos más adelante.

Distribución t-student

Dadas dos variables aleatorias independientes; una $Z \rightarrow N(0, 1)$ y otra χ_k^2 con k grados de libertad, generamos una nueva variable aleatoria llamada t donde:

$$\frac{Z}{\sqrt{\chi_k^2/k}} \rightarrow t_k$$



- Su media es cero y su varianza $\frac{n}{n-2}$
- El estadístico $t = \frac{\bar{X} - \mu}{s/\sqrt{n}}$

Contenidos

- 1 Inferencia estadística
- 2 Conceptos básicos
- 3 Muestreo
- 4 Teorema Central del límite
- 5 Estimación puntual
- 6 Dos distribuciones importantes
- 7 Estimación por intervalos

Estimación por intervalos de confianza

Hasta ahora hemos usado datos muestrales y :

- 1 Hemos asumido que provienen de una distribución en específico con ciertos parámetros. e.g, que proceden de una $N(\mu, \sigma^2)$
- 2 Hemos estimado el valor de los parámetros desconocidos .

Pero, son nuestros estimaciones validas? Tenemos que poder evaluar nuestro margen de error: Intervalos de confianza.

Estimación por intervalos de confianza

- Un intervalo de confianza consiste en un rango donde con una determinada probabilidad se encontrará el valor del parámetro poblacional. e.g., podemos afirmar con un 90% que la media de altura de los españoles se encuentra entre 1,6 y 1,8 metros.
- Esa probabilidad de que el intervalo calculado contenga el verdadero valor del parámetro es lo que denominamos nivel de confianza y lo denotamos por $1 - \alpha$ donde α es el error (o el nivel de significacion) que estamos dispuestos a asumir.

Estimación por intervalos de confianza de la media de población normal con varianza conocida

Tenemos una v.a. que se distribuye normalmente : $X \rightarrow N(\mu, \sigma)$ y una muestra x_1, x_2, \dots, x_n . Se quiere estimar μ mediante un intervalo de confianza $(1 - \alpha)\%$.

Del estimador media muestral, gracias al TCL, tenemos: $\bar{X} \rightarrow N(\mu, \frac{\sigma}{\sqrt{n}})$

Elegimos un nivel de significancia α y usamos el estadístico Z : $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \rightarrow N(0, 1)$

$$P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1 - \alpha$$

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

ó

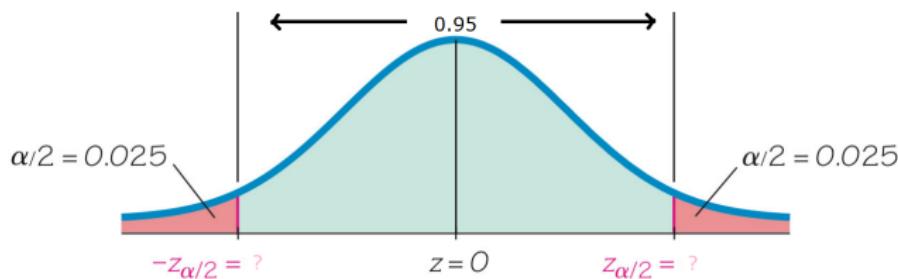
$$P(\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}) = 1 - \alpha$$

I.C. para μ con confianza de $(1 - \alpha)\%$ y σ conocida: $(\bar{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}})$

Estimación por intervalos de confianza de la media de población normal con varianza conocida

Para un nivel de significación de $\alpha = 0,05$ (un nivel de confianza de 0,95), tenemos:

$$P(-z_{0,025} \leq Z \leq z_{0,025}) = 0,95$$



Buscamos en la tabla el valor correspondiente a $\pm z_{0,05/2}$.

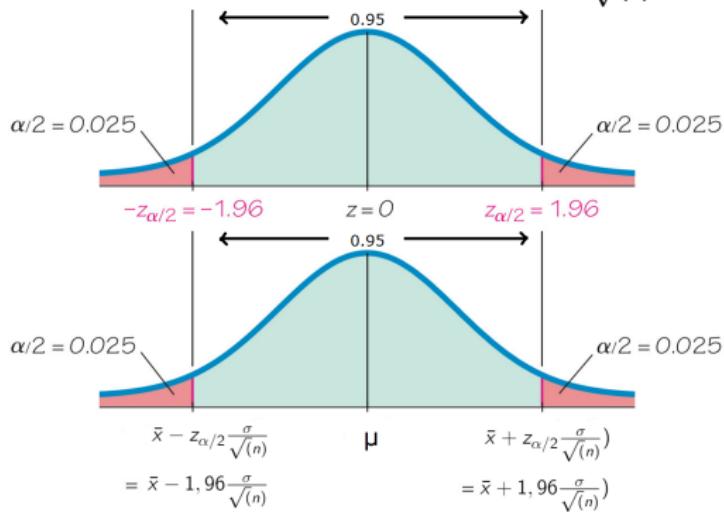
Estimación por intervalos de confianza de la media de población normal con varianza conocida

$$P(-1,96 \leq Z \leq 1,96) = 0,95$$

$$P\left(-1,96 \leq \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \leq 1,96\right) = 0,95$$

$$P\left(\bar{x} - 1,96 * \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + 1,96 * \frac{\sigma}{\sqrt{n}}\right) = 0,95$$

Podemos decir que hay una probabilidad de 0.95 de que este intervalo de confianza contenga al verdadero valor del parámetro : $(\bar{x} - 1,96 * \frac{\sigma}{\sqrt{n}}, \bar{x} + 1,96 * \frac{\sigma}{\sqrt{n}})$

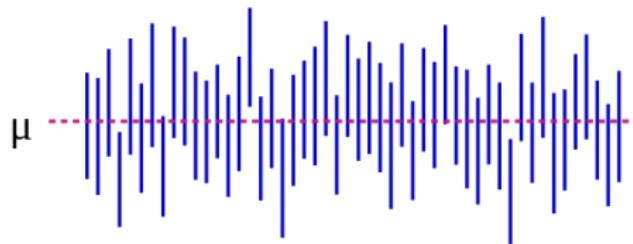


Estimación por intervalos de confianza de la media de población normal con varianza conocida

Ejemplo

Estimación por intervalos de confianza

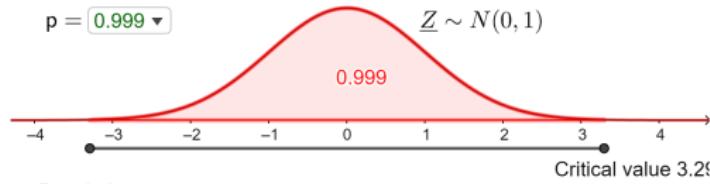
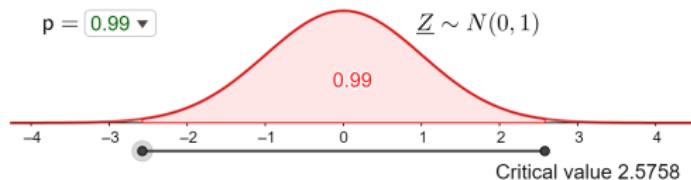
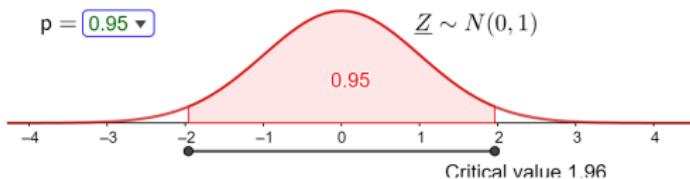
- Sabemos que si extrajeramos muchas muestras de la población y construyeramos intervalos , el $(1 - \alpha)\%$ contendría al verdadero valor del parámetro.



- Nótese que:
 - ▶ Cuanto mayor es la σ , mayor es la longitud del intervalo.
 - ▶ Cuanto mayor es n , menor es la longitud del intervalo →
mayor precisión. Cuanto menor es el nivel de significación, mayores es el nivel de confianza y menor precisión

Estimación por intervalos de confianza

Confidence level



Estimación por intervalos de confianza de la media de población normal con varianza desconocida

Si desconocemos la desviación típica de la población, procedemos como en el caso anterior pero usando el estimador muestral:⁴

$$S = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}}$$

Y usamos el estadístico-t : $t = \frac{\bar{X} - \mu}{S/\sqrt{n}} \rightarrow t_{n-1}$

$$P\left(-t_{n-1, \alpha/2} < \frac{\bar{X} - \mu}{s/\sqrt{n}} < t_{n-1, \alpha/2}\right) = 1 - \alpha$$

$$P\left(\bar{X} - t_{n-1, \alpha/2} \frac{s}{\sqrt{n}} < \mu < \bar{X} + t_{n-1, \alpha/2} \frac{s}{\sqrt{n}}\right) = 1 - \alpha$$

I.C. para μ con confianza de $(1 - \alpha)\%:$ $(\bar{X} \pm t_{n-1, \alpha/2} \frac{s}{\sqrt{n}})$

⁴Como hemos visto, para muestras grandes la distribución t-student se puede aproximar para la normal así que en estos casos ($n > 30$) usaremos el estadístico Z como antes

Estimación por intervalos de confianza de la media de población normal con varianza desconocida

Ejemplo

Estimación por intervalos de confianza para la proporción

Si tenemos una v.a. $\tilde{B}(n, p)$, podemos construir un intervalo de confianza para el valor del parámetro p .

$$\hat{p} \rightarrow N(p, \sqrt{pq/n})$$

El estadístico Z es: $Z = \frac{\hat{p}-p}{\sqrt{\hat{p}\hat{q}/n}} \rightarrow N(0, 1)^5$

IC para del $(1 - \alpha)\%$: $\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}\hat{q}}{n}} + \frac{1}{2n}$

⁵no conocemos p

Estimación por intervalos de confianza para la proporción

Ejemplo