Intervalos de confianza para una población normal

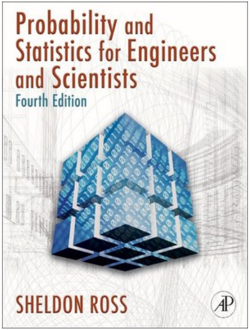
Estadística, Grado en Sistemas de Información

Constantino Antonio García Martínez

Universidad San Pablo Ceu

Contenidos

References



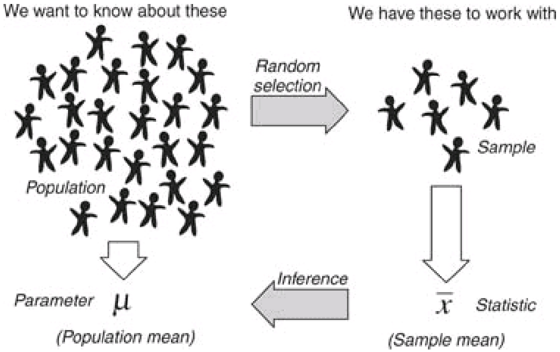
S. Ross. Introduction to Probability and Statistics for Engineers and Scientists. Chapter 7.

C.D. Barr, D.M. Diez, M. Çetinkaya-Rundel. OpenIntro Statistics. Chapters 4.

Introducción a la estadística

La estadística y el muestreo

La estadística trata de obtener conclusiones a partir de la observación de datos. Generalmente estos datos se obtienen a partir del **muestreo** de una **población**.



Un **estadístico S** es **una VARIABLE ALEATORIA calculada a partir de los valores de una muestra**

$$S = f(X_1, X_2, \ldots, X_n)$$

.

Intervalos de confianza para medias si la varianza es conocida

Consideremos el problema de **estimar un parámetro desconocido a partir de una muestra**. Por ejemplo...

```
## Loading required package: nlme

##
## Attaching package: 'nlme'

## The following object is masked from 'package:dplyr':
##
##     collapse

## This is mgcv 1.8-42. For overview type 'help("mgcv-package")'.
```

Ejercicio: CIs

El fichero “iq.csv” contiene los resultados de un test de cociente intelectual (CI) realizado a una muestra de estudiantes de cierta universidad (con cientos de miles de estudiantes matriculados). ¿Cuál es el cociente intelectual medio de los alumnos de la universidad? Ten en cuenta que los tests de cociente intelectual se diseñan para que la desviación estándar poblacional sea de 15 puntos.

Procederemos en pasos...

Intervalos de confianza para medias si la varianza es conocida

1) Buscar un estadístico para estimar el parámetro poblacional.

1) Buscar un estadístico para estimar el parámetro poblacional. Lo lógico es emplear la **media muestral** para estimar la **media poblacional** μ:

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Ejercicio: CIs, paso 1
Calcula la media muestral para el problema de los CIs.

```
## `geom_smooth()` using method = 'loess' and formula = 'y ~
## x'
```

1) Buscar un estadístico para estimar el parámetro poblacional. Lo lógico es emplear la **media muestral** para estimar la **media poblacional** μ:

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Ejercicio: CIs, paso 1
Calcula la media muestral para el problema de los CIs.

```
## `geom_smooth()` using method = 'loess' and formula = 'y ~
## x'
```

Con los datos del enunciado $\bar{x} = 99$ (cuando lo habitual es 100).

1) Buscar un estadístico para estimar el parámetro poblacional. Lo lógico es emplear la **media muestral** para estimar la **media poblacional** μ:

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Ejercicio: CIs, paso 1

Calcula la media muestral para el problema de los CIs.

```
## `geom_smooth()` using method = 'loess' and formula = 'y ~
## x'
```

Con los datos del enunciado $\bar{x} = 99$ (cuando lo habitual es 100).

¿Significa esto que los alumnos de esta universidad tienen un cociente intelectual bajo?

1) Buscar un estadístico para estimar el parámetro poblacional. Lo lógico es emplear la **media muestral** para estimar la **media poblacional** μ:

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Ejercicio: CIs, paso 1
Calcula la media muestral para el problema de los CIs.

```
## `geom_smooth()` using method = 'loess' and formula = 'y ~
## x'
```

Con los datos del enunciado $\bar{x} = 99$ (cuando lo habitual es 100).

¿Significa esto que los alumnos de esta universidad tienen un cociente intelectual bajo?

La conclusión es precipitada... Lo que acabamos de hacer es una **estimación puntual**, y esta no incorpora una medida de incertidumbre. Podemos hacerlo mejor si descubrimos cuál es la distribución de $\bar{X}$.

2) Buscar la distribución en el muestreo para el estadístico.

Este problema puede ser muy complejo. Para simplificarlo, en este curso asumiremos...

- ... o bien **población infinita** (o muy grande comparada con la muestra)...
- ... o bien **muestreo aleatorio simple** (muestreo con reemplazamiento) (¿¿¡¡Por qué!!??).

Adicionalmente, suele ser necesaria hacer algún asunción acerca de las muestras X_i.

Ejercicio: CIs, paso 2
¿Cuál es la distribución de $\bar{X}$?

2) Buscar la distribución en el muestreo para el estadístico.

Este problema puede ser muy complejo. Para simplificarlo, en este curso asumiremos...

- ... o bien **población infinita** (o muy grande comparada con la muestra)...
- ... o bien **muestreo aleatorio simple** (muestreo con reemplazamiento) (¿¿¡¡Por qué!!??).

Adicionalmente, suele ser necesaria hacer algún asunción acerca de las muestras X_i.

Ejercicio: CIs, paso 2

¿Cuál es la distribución de $\bar{X}$?

$$\bar{X} = \mathcal{N}\left(\mu, \frac{\sigma^2}{n} = \frac{15^2}{100}\right).$$

3) Escribir matemáticamente la regla del 68-95-99 (o análoga) para el **nivel de confianza** deseado.

Ejercicio: CIs, paso 3a

Escribe matemáticamente la regla del 68-95-99 (tres ecuaciones, una para cada probabilidad) para nuestro estadístico $\bar{X}$.

3) Escribir matemáticamente la regla del 68-95-99 (o análoga) para el **nivel de confianza** deseado.

Ejercicio: CIs, paso 3a
Escribe matemáticamente la regla del 68-95-99 (tres ecuaciones, una para cada probabilidad) para nuestro estadístico $\bar{X}$.

De forma general, para un **nivel de significación** α:

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

siendo Z una VA estandarizada

$$Z = \frac{\bar{X} - \mu}{15/10} \Longrightarrow Z \sim \mathcal{N}(0, 1).$$

Ejercicio: CIs, paso 3b
Halla $z_{\alpha/2}$ para un nivel de significación del 98 %.

4) Centrándote en las inecuaciones del 68-95-99, despeja el parámetro poblacional y sustituye el estadístico por su valor observado.

Ejercicio: CIs, paso 4a

Despeja μ usando las inecuaciones del *paso 3* y empleando que $Z = \frac{\bar{X}-\mu}{15/10}$.

Ejercicio: CIS, paso 4b

Halla un **intervalo de confianza** al 98 % para μ empleando que $\bar{x} = 99$

4) Centrándote en las inecuaciones del 68-95-99, despeja el parámetro poblacional y sustituye el estadístico por su valor observado.

Ejercicio: CIs, paso 4a

Despeja μ usando las inecuaciones del *paso 3* y empleando que $Z = \frac{\bar{X}-\mu}{15/10}$.

Ejercicio: CIS, paso 4b

Halla un **intervalo de confianza** al 98 % para μ empleando que $\bar{x} = 99$

```
## [1]  95.51048 102.48952
```

Ejercicio: CIs

Halla un **intervalo de confianza** al 90 % para μ empleando que $\bar{x} = 99$

```
## [1]  96.53272 101.46728
```

Intervalos de confianza para la varianza muestral

Un estadístico razonable para estimar la **varianza poblacional** σ^2 es el siguiente...

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

Sin embargo, este estimador es **sesgado** ya que

$$\mathbb{E}[S^2] \neq \sigma^2.$$

Ejercicio: Estimador sesgado

Completa el siguiente código para verificar que S^2 es sesgado:

```
estimate_var <- function(x) {
    1 / length(???) * sum((x - ???) ^ 2)
}

replicate(
  5,
  mean(replicate(10000, {
    samples <- rnorm(4, sd = 2)
    estimate_var(samples)
  }))
)
```

```
## [1] 2.999511 2.987051 3.028164 2.995919 2.961770
```

Por ello definimos la **varianza muestral** como

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

¡Ten en cuenta que la función *var* de `R` implementa $\hat{S}^2$!

Ejercicio: Varianza muestral insesgada

Demuestra que $\hat{S}^2$ es insesgado con simulaciones (usa el anterior ejercicio como punto de partida).

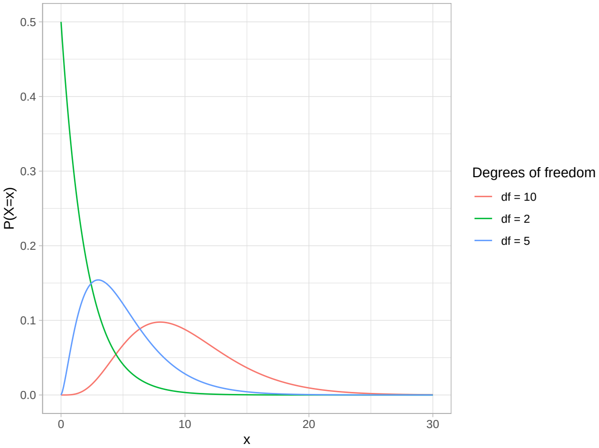
Si se toman n muestras de una **población normal**, entonces la VA

$$\frac{(n-1)\hat{S}^2}{\sigma^2}$$

tiene distribución **Chi-cuadrado** $n-1$ **grados de libertad**.

$$\frac{(n-1)\hat{S}^2}{\sigma^2} \sim \chi^2_{n-1}$$

Intervalos de confianza para la varianza muestral



Ejercicio: Chips

Una máquina fabrica cierta pieza de un ordenador. El tamaño deseado de la pieza es de 5 cm. En el proceso de fabricación siempre hay circunstancias que no se pueden controlar, por lo que el tamaño de la pieza varía aleatoriamente. El proceso de fabricación está diseñado de forma que el tamaño de cualquier pieza $X_i \sim \mathcal{N}(5, \sigma^2)$, aunque todavía no se conoce σ^2. Para ello, se dispone de los datos almacenados en "pieces.csv". Halla un intervalo de confianza al 99 % para σ^2.

```
## [1] 0.0005653365 0.0022549273
```

Intervalos de confianza para la media cuando la varianza es desconocida

Distribución de la media cuando la varianza es desconocida

Hemos visto que

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1)$$

pero, ¿qué hacer si σ es desconocida?

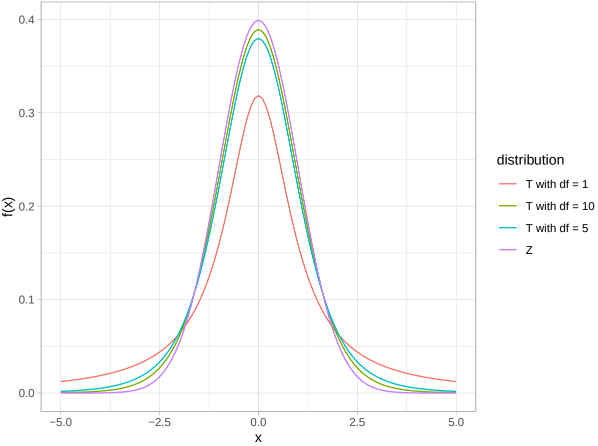
Teorema
Si n muestras se toman de una **población normal**, la VA

$$\frac{\bar{X} - \mu}{\hat{S}/\sqrt{n}}$$

se distribuye como una **T de Student con** $n-1$ **grados de libertad**.

$$\frac{\bar{X} - \mu}{\hat{S}/\sqrt{n}} \sim \mathcal{T}_{n-1}$$

Distribución de la media cuando la varianza es desconocida



Distribución de la media cuando la varianza es desconocida

Ejercicio:

Nueva York es conocida como "la ciudad que nunca duerme". En cierta encuesta se preguntó a una muestra aleatoria de 25 neoyorquinos cuánto tiempo dormían por la noche. Los datos se encuentra en "new_york.csv". ¿Apoyan los datos la afirmación de que los neoyorquinos duermen menos de 8 horas por noche en promedio? Usa un nivel de confianza del 96 %.