

# Estimación y contraste

Estadística, Grado en Sistemas de Información

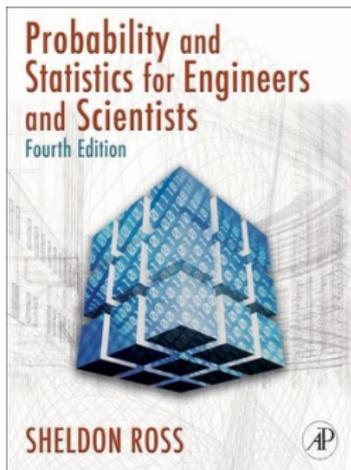
---

Constantino Antonio García Martínez

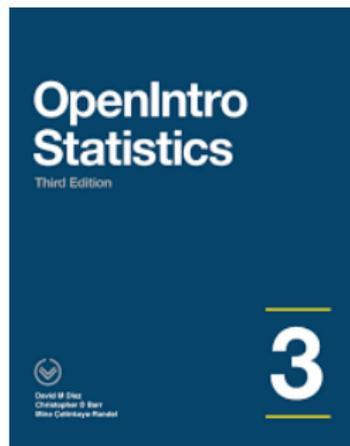
19 de noviembre de 2019

Universidad San Pablo Ceu

1. Definiciones
2. Intervalos de confianza
  - ICs para medias
  - ICs para proporciones
  - ICs para sumas y diferencias
  - ICs para la varianza
3. Estimaciones puntuales: método de máxima verosimilitud



S. Ross. Introduction to Probability and Statistics for Engineers and Scientists. Chapter 6.



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

## Definiciones

---

En estas transparencias afrontaremos el problema de **estimar un parámetro desconocido a partir de una muestra**. Algunas definiciones útiles:

- Un estadístico es **insesgado** si su esperanza es igual al parámetro poblacional que estima. En caso contrario, se dice que es **sesgado**.

### Ejemplo:

La cuasivarianza es un estadístico insesgado ya que  $\mathbb{E}[\hat{S}^2] = \sigma^2$ . En cambio, la varianza es un estadístico sesgado ya que  $\mathbb{E}[S^2] = \frac{n}{n-1}\sigma^2$ .

- Si dos estadísticos tienen la misma esperanza, preferiremos aquél con menos varianza y diremos que es un estimador **más eficiente/preciso**.
- Una estimación de un parámetro dada por un simple número es una **estimación puntual**. Una estimación de un parámetro dada por un intervalo es una **estimación por intervalos**.

### Ejemplo:

Una estimación puntual sería decir la anchura media de una pieza es de 6.03 cm. Una estimación interválica sería  $6.03 \pm 0.1$  cm.

Los intervalos nos dan información acerca de la precisión de la estimación, a veces llamada **fiabilidad**.

## Intervalos de confianza

---

Ilustramos el concepto de intervalo de confianza (IC) con un ejemplo.

### Ejemplo: Intervalos de Confianza

Sea un estadístico  $S$  con media  $\mu_s$  y  $\sigma_s$  con distribución Normal (o aproximadamente Normal, lo cuál es cierto para muchos estadísticos si  $n \geq 30$ ). Usando la regla del 68-95-99 esperamos encontrar  $S$  en los intervalos

$\mu_s \pm \sigma_s$  el 68.27 % de las veces,

$\mu_s \pm 2\sigma_s$  el 95.45 % de las veces,

$\mu_s \pm 3\sigma_s$  el 99.73 % de las veces.

## Ejemplo: (Continuación)

Dado que, por ejemplo,

$P(\mu_s - \sigma_s \leq S \leq \mu_s + \sigma_s) = P(S - \sigma_s \leq \mu_s \leq S + \sigma_s) = 0.6827$ , podemos esperar encontrar  $\mu_s$  en los intervalos

$\mu_s \pm \sigma_s$  el 68.27 % de las veces,

$\mu_s \pm 2\sigma_s$  el 95.45 % de las veces,

$\mu_s \pm 3\sigma_s$  el 99.73 % de las veces.

Llamamos a estos intervalos **Intervalos de confianza** para  $\mu_s$  del 68.27 %, 95.45 % y 99.73 %, respectivamente. A cada uno de estos porcentajes se le conoce como **nivel de confianza** y a su complementario **nivel de significación**. Al cambiar el nivel de confianza cambiará el “número de desviaciones estándar” a la izquierda y derecha de  $S$ . Este valor se conoce como **valor crítico** y lo denotaremos con  $z_c$  para las distribuciones normales.

## Intervalos de confianza

---

ICs para medias

### Con varianza conocida

Asumimos que la población es Normal o bien, si la población no es normal que el muestreo es grande  $n \geq 30$ .

- Muestreo con reemplazamiento o población infinita:

$$\bar{x} - z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}$$

- Muestreo sin reemplazamiento de una población finita de tamaño N:

$$\bar{x} - z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \leq \mu \leq \bar{x} + z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

### Ejercicio:

Se quiere transmitir un valor analógico  $\mu$  a través de un medio que introduce ruido  $\epsilon \sim \mathcal{N}(0, \sigma^2 = 4)$ . Para reducir el error, se transmite 9 veces el mismo valor. ¿IC al 95% para  $\mu$  si se recibe 5, 8.5, 12, 15, 7, 9, 7.5, 6.5, 10.5 ?

## Ejercicio:

Se quiere transmitir un valor analógico  $\mu$  a través de un medio que introduce ruido  $\epsilon \sim \mathcal{N}(0, \sigma^2 = 4)$ . Para reducir el error, se transmite 9 veces el mismo valor. ¿IC al 95% para  $\mu$  si se recibe 5, 8.5, 12, 15, 7, 9, 7.5, 6.5, 10.5 ?

```
samples = c(5, 8.5, 12, 15, 7, 9, 7.5, 6.5, 10.5)
n = length(samples)
x_hat = mean(samples)
sigma = sqrt(4)
z_c = qnorm(0.975)
c(x_hat - z_c * sigma / sqrt(n), x_hat + z_c * sigma / sqrt(n))

## [1] 7.693357 10.306643
```

## Con varianza desconocida

Asumimos que la población es Normal y que el muestreo es con reemplazamiento o la población infinita.

- $n < 30$ :

$$\bar{x} - t_{n-1, 1-\alpha/2} \frac{\hat{s}}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{n-1, 1-\alpha/2} \frac{\hat{s}}{\sqrt{n}}$$

- $n \geq 30$ . Si la población es grande podemos aproximar la T de Student por una Normal:

$$\bar{x} - z_{1-\alpha/2} \frac{\hat{s}}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{1-\alpha/2} \frac{\hat{s}}{\sqrt{n}}$$

### Ejercicio:

Repita el problema de la transmisión si la varianza es desconocida

## Ejercicio:

Repite el problema de la transmisión si la varianza es desconocida

```
## Method 1
samples = c(5, 8.5, 12, 15, 7, 9, 7.5, 6.5, 10.5)
n = length(samples)
x_hat = mean(samples)
s = sd(samples)
t_c = qt(0.975, df = n - 1)
c(x_hat - t_c * s / sqrt(n), x_hat + t_c * s / sqrt(n))

## [1] 6.630806 11.369194

## Method 2
test = t.test(samples, conf.level = 0.95)
test$conf.int

## [1] 6.630806 11.369194
## attr(,"conf.level")
## [1] 0.95
```

## Intervalos de confianza

---

### ICs para proporciones

### ICs para proporciones

Asumimos que obtenemos un muestra grande  $n \geq 30$  de una población binomial en donde  $p$  es la probabilidad de éxito.

- Muestreo con reemplazamiento o población infinita:

$$\hat{p} - z_{1-\alpha/2} \sqrt{\frac{\hat{p}\hat{q}}{n}} \leq p \leq \hat{p} + z_{1-\alpha/2} \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

- Muestreo sin reemplazamiento de una población finita de tamaño  $N$ :

$$\hat{p} - z_{1-\alpha/2} \sqrt{\frac{\hat{p}\hat{q}}{n}} \sqrt{\frac{N-n}{N-1}} \leq \mu \leq \hat{p} + z_{1-\alpha/2} \sqrt{\frac{\hat{p}\hat{q}}{n}} \sqrt{\frac{N-n}{N-1}}$$

## Intervalos de confianza

---

ICs para sumas y diferencias

### ICs para sumas y diferencias

Si  $S_1$  y  $S_2$  son dos estadísticos con distribución (aproximadamente) normal, tenemos que los ICs para la suma de los parámetros poblaciones estimados por  $s_1$  y  $s_2$  vendrán dados por los límites:

$$s_1 + s_2 \pm z_{1-\alpha/2} \sqrt{\sigma_{S_1}^2 + \sigma_{S_2}^2},$$

mientras que los ICs para la diferencia de parámetros poblaciones vendrán dados por

$$S_1 - S_2 \pm z_{1-\alpha/2} \sqrt{\sigma_{S_1}^2 + \sigma_{S_2}^2},$$

### Ejemplo:

Diferencia de medias

$$\bar{x}_1 - \bar{x}_2 - z_c \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \leq \mu_1 - \mu_2 \leq \bar{x}_1 - \bar{x}_2 + z_c \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

## ICs para sumas y diferencias

Dos instrumentos de electrocardiografía se han testeado para determinar el voltaje al que fallan. Los fallos registrados fueron:

Aparato A

36	54
44	52
41	37
53	51
38	44
36	35
34	44

Aparato B

52	60
64	44
38	48
68	46
66	70
52	62

¿IC al 95% de la diferencia de medias, si las medias son normales con  $\sigma_A^2 = 40$  y  $\sigma_B^2 = 100$

## ICs para sumas y diferencias

Dos instrumentos de electrocardiografía se han testeado para determinar el voltaje al que fallan. Los fallos registrados fueron:

Aparato A

36	54
44	52
41	37
53	51
38	44
36	35
34	44

Aparato B

52	60
64	44
38	48
68	46
66	70
52	62

¿IC al 95% de la diferencia de medias, si las medias son normales con  $\sigma_A^2 = 40$  y  $\sigma_B^2 = 100$

```
a = c(36, 44, 41, 53, 38, 36, 34, 54, 52, 37, 51, 44, 35, 44)
n = length(a)
b = c(52, 64, 38, 68, 66, 52, 60, 44, 48, 46, 70, 62)
m = length(b)

s = sqrt(40 / n + 100 / m)
diff_mean = mean(a) - mean(b)
c(diff_mean - qnorm(0.975) * s, diff_mean + qnorm(0.975) * s)

## [1] -19.604124 -6.491114
```

## ICs para sumas y diferencias Normales con varianza desconocida

Si la varianza es desconocida, podemos hallar los ICs suponiendo:

- La población es normal.
- Las varianzas de las poblaciones son iguales:  $\sigma_1^2 = \sigma_2^2 = \sigma^2$ .

En tal caso,

$$\hat{\sigma}_p^2 = \frac{(n-1)\hat{S}_1^2 + (m-1)\hat{S}_2^2}{n+m-2},$$

verifica que  $(n+m-2)\hat{S}_p/\sigma^2 \sim \chi_{n+m-2}^2$ , y por tanto

$$\frac{\bar{X} \pm \bar{Y} - (\mu_1 \pm \mu_2)}{\hat{S}_p \sqrt{1/n + 1/m}} \sim t_{n+m-2}.$$

Se sigue que el IC al  $100(1-\alpha)$  para  $\bar{X} + \bar{Y}$  es

$$\left( \bar{x} + \bar{y} - t_{n+m-2, 1-\alpha/2} \hat{S}_p \sqrt{1/n + 1/m}, \bar{x} + \bar{y} + t_{n+m-2, 1-\alpha/2} \hat{S}_p \sqrt{1/n + 1/m} \right)$$

El IC al  $100(1-\alpha)$  para  $\bar{X} - \bar{Y}$  es

$$\left( \bar{x} - \bar{y} - t_{n+m-2, 1-\alpha/2} \hat{S}_p \sqrt{1/n + 1/m}, \bar{x} - \bar{y} + t_{n+m-2, 1-\alpha/2} \hat{S}_p \sqrt{1/n + 1/m} \right)$$

### Ejercicio:

Se emplean dos procedimientos para producir baterías. Las capacidades de ambos métodos dan como resultado (en Amperios hora):

Aparato A		Aparato B	
140	132	144	134
136	142	132	130
138	150	136	146
150	154	140	128
152	136	128	131
144	142	150	137
		130	135

¿IC al 90% de la diferencia de medias? Asume normalidad y que las varianzas de ambos métodos son iguales.

# ICs para medias con varianza desconocida

## Ejercicio:

Se emplean dos procedimientos para producir baterías. Las capacidades de ambos métodos dan como resultado (en Amperios hora):

Aparato A		Aparato B	
140	132	144	134
136	142	132	130
138	150	136	146
150	154	140	128
152	136	128	131
144	142	150	137
		130	135

¿IC al 90% de la diferencia de medias? Asume normalidad y que las varianzas de ambos métodos son iguales.

```
a = c(140, 132, 136, 142, 138, 150, 150, 154, 152, 136, 144, 142)
b = c(144, 134, 132, 130, 136, 146, 140, 128, 128, 131, 150, 137, 130, 135)
n = length(a)
m = length(b)

sp = sqrt(((n - 1) * var(a) + (m - 1) * var(b)) / (n + m - 2))
diff_mean = mean(a) - mean(b)
c(diff_mean - qt(0.95, df = n + m - 2) * sp * sqrt(1 / n + 1 / m),
  diff_mean + qt(0.95, df = n + m - 2) * sp * sqrt(1 / n + 1 / m))

## [1] 2.498164 11.930407
```

## Intervalos de confianza

---

ICs para la varianza

### ICs para la varianza de una población Normal

Dado que  $nS^2/\sigma^2 = (n-1)\hat{S}^2/\sigma^2 \sim \chi_{n-1}^2$  tenemos

$$(n-1)\hat{S}^2/\chi_{c_1, n-1}^2 \leq \sigma^2 \leq (n-1)\hat{S}^2/\chi_{c_2, n-1}^2,$$

donde para un nivel de significación de  $\alpha$  seleccionamos  $c_1 = 1 - \alpha/2$  y  $c_2 = \alpha/2$ .

### Ejemplo: IC del 95 %

Usamos  $c_2 = 0.025$  y  $c_1 = 0.975$ , por lo que el IC es:

$$(n-1)\hat{S}^2/\chi_{0.975, n-1}^2 \leq \sigma^2 \leq (n-1)\hat{S}^2/\chi_{0.025, n-1}^2$$

### Ejercicio:

Un proceso de fabricación de arandelas está diseñado para que hay poca variación en su grosor. Estima la varianza del proceso si en una muestra se obtiene que los grosores son: 0.123 , 0.124 , 0.126 , 0.120 , 0.130 , 0.133 , 0.125 , 0.128 , 0.124 , 0.126 (en cms). Utiliza un IC al 90% de confianza para tu estimación.

### Ejercicio:

Un proceso de fabricación de arandelas está diseñado para que hay poca variación en su grosor. Estima la varianza del proceso si en una muestra se obtiene que los grosores son: 0.123 , 0.124 , 0.126 , 0.120 , 0.130 , 0.133 , 0.125 , 0.128 , 0.124 , 0.126 (en cms). Utiliza un IC al 90% de confianza para tu estimación.

```
x = c(0.123, 0.124, 0.126, 0.120, 0.130, 0.133, 0.125, 0.128, 0.124, 0.126)
n = length(x)
s_2 = var(x)
# CIs in cm^2
c(
  (n - 1) * s_2 / qchisq(0.95, df = n - 1),
  (n - 1) * s_2 / qchisq(0.05, df = n - 1)
)

## [1] 7.264032e-06 3.696115e-05
```

### ICs para la ratios de varianzas de poblaciones Normales

Para dos muestras de dos **poblaciones normales** de tamaño  $m$  y  $n$  hemos visto que  $\frac{\hat{\sigma}_1^2/\sigma_1^2}{\hat{\sigma}_2^2/\sigma_2^2}$  tiene distribución F con  $m - 1$ ,  $n - 1$  grados de libertad. Por tanto:

$$\frac{1}{F_{c_1}} \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} \leq \frac{\sigma_1^2}{\sigma_2^2} \leq \frac{1}{F_{c_2}} \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2},$$

donde para un nivel de significación de  $\alpha$  seleccionamos  $c_1 = 1 - \alpha/2$  y  $c_2 = \alpha/2$ .

### Ejemplo: IC del 98 %

Usamos  $c_2 = 0.01$  y  $c_1 = 0.99$ , por lo que el IC es:

$$\frac{1}{F_{0.99}} \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} \leq \frac{\sigma_1^2}{\sigma_2^2} \leq \frac{1}{F_{0.01}} \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2}$$

### Ejercicio:

Dos muestras de tamaño 16 y 10 se obtienen de dos poblaciones normales. Si su varianzas muestrales son 24 y 18 encuentra el IC al 98 % para el ratio de varianzas.

### Ejercicio:

Dos muestras de tamaño 16 y 10 se obtienen de dos poblaciones normales. Si su varianzas muestrales son 24 y 18 encuentra el IC al 98% para el ratio de varianzas.

```
# Convert variances to cuasi-variances
s2_a = 16 / 15 * 24
s2_b = 10 / 9 * 18
ratio = s2_a / s2_b
print(c(ratio / qf(0.99, 16 - 1, 10 - 1), ratio / qf(0.01, 16 - 1, 10 - 1)))

## [1] 0.2579564 4.9853288

# When data is available, we may use var.test
x = rnorm(16); y = rnorm(10)
x = x / sd(x) * sqrt(s2_a); y = y / sd(y) * sqrt(s2_b)
test_result = var.test(x, y, conf.level = 0.98)
print(test_result$conf.int)

## [1] 0.2579564 4.9853288
## attr(,"conf.level")
## [1] 0.98
```

## **Estimaciones puntuales: método de máxima verosimilitud**

---

Aunque los ICs son muy útiles, a veces es conveniente tener una estimación puntual del parámetro poblacional. Para ello empleamos el método de máximo verosimilitud (*maximum likelihood*, debido a Fisher):

### Máxima verosimilitud

Tenemos  $X_1, X_2, \dots, X_n$  observaciones independientes de una población con densidad  $f(X, \theta)$ . Para estimar  $\theta$  construimos la función de verosimilitud:

$$\mathcal{L} = f(x_1, \theta)f(x_2, \theta) \cdots f(x_n, \theta),$$

y simplemente buscamos el parámetro  $\theta$  que maximice  $\mathcal{L}$  (tenemos un problema de optimización).

Dado que la función log es estrictamente monótona, el máximo de  $\mathcal{L}$  es el mismo que  $\log \mathcal{L}$ . Así pues, por conveniencia maximizamos  $\log \mathcal{L}$  calculando

$$\frac{\partial \log \mathcal{L}}{\partial \theta} = \frac{\partial}{\partial \theta} \sum_{i=1}^n \log f(x_i, \theta) = 0.$$

## Estimación de máxima verosimilitud

Sea una población de variables independientes tomadas de una población exponencial de parámetro desconocido  $\lambda$ . Estima  $\lambda$  por el método de máxima verosimilitud.

## Estimación de máxima verosimilitud

Sea una población de variables independientes tomadas de una población exponencial de parámetro desconocido  $\lambda$ . Estima  $\lambda$  por el método de máxima verosimilitud.

```
samples = rexp(1000, 3)

minus_loglk = function(param, samples) {
  -sum(dexp(samples, param, log = TRUE))
}
opt = optim(par = 1, minus_loglk, samples = samples,
           method = "Brent", lower = 0, upper = 10000)
opt$par

## [1] 3.323755

# Theoretical

1 / mean(samples)

## [1] 3.323755
```