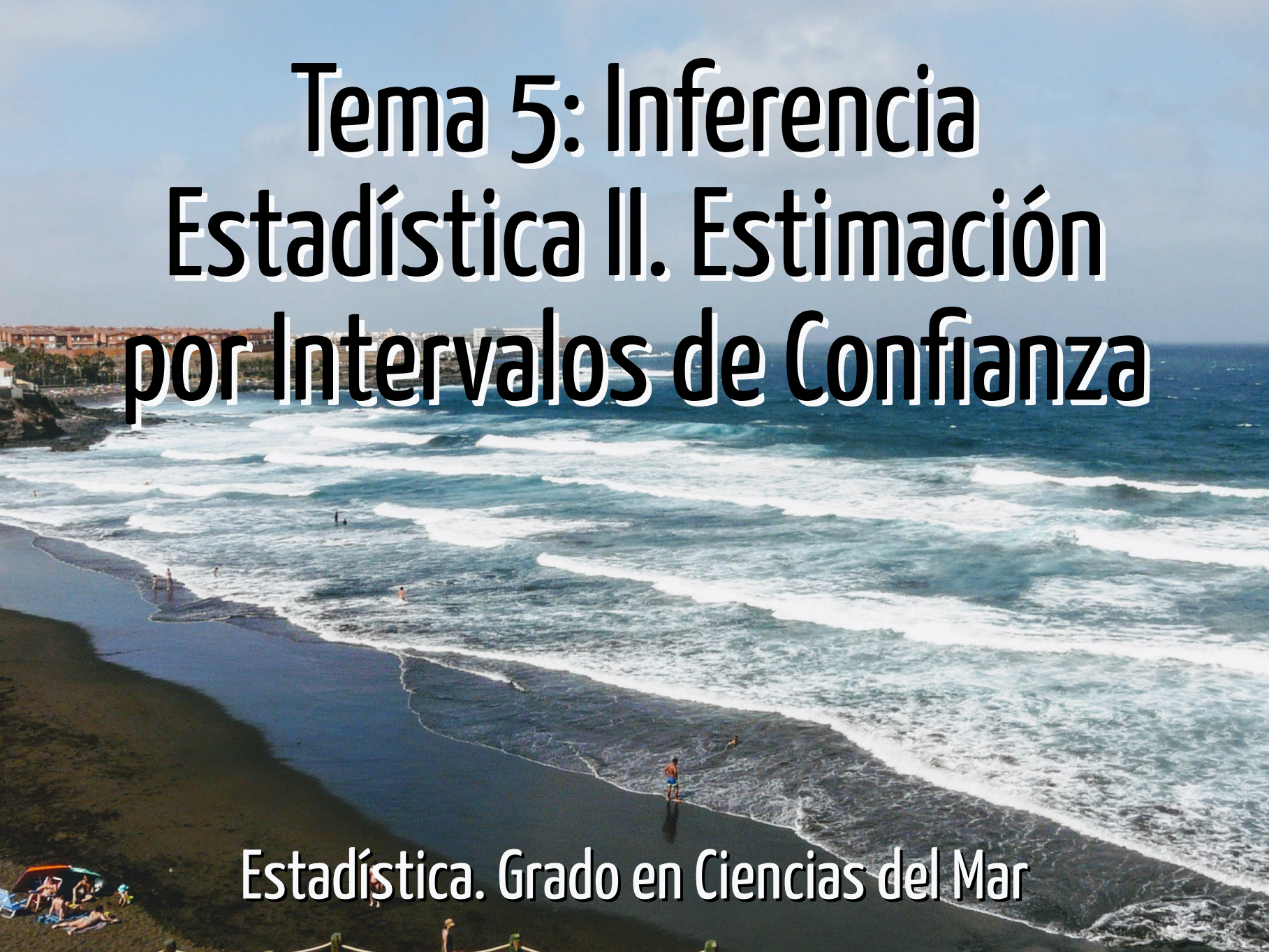
Tema 5: Inferencia Estadística II. Estimación por Intervalos de Confianza



Estadística. Grado en Ciencias del Mar

Intervalo de confianza

Sean:

- X una variable aleatoria cuya distribución queda caracterizada por un parámetro desconocido θ
- $\omega = \{X_1, X_2, \ldots, X_n\}$ una muestra aleatoria de observaciones de dicha variable.

Entonces $[\theta_1(\omega), \theta_2(\omega)]$, donde $\theta_1(\omega)$ y $\theta_2(\omega)$ son variables aleatorias que dependen de la muestra, es un **intervalo de confianza** a nivel $1-\alpha$ para el parámetro θ si:

$$P(\theta \in [\theta_1(\omega), \theta_2(\omega)]) = 1 - \alpha$$

Ejemplo:

Ya hemos visto que si $X \approx N(\mu, \sigma)$, y $\bar{X}$ es la media aritmética de una muestra aleatoria de observaciones $\omega = \{X_1, X_2, \ldots, X_n\}$ de X:

$$\frac{\bar{X} - \mu}{s/\sqrt{n}} \approx t_{n-1}$$

y de aquí es posible deducir que *antes de tomar la muestra*:

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

Obviamente, los extremos del intervalo $\theta_1(\omega) = \bar{X} - t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}$ y $\theta_2(\omega) = \bar{X} + t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}$ son variables aleatorias porque su valor no se conoce hasta que no se haya tomado la muestra.

Interpretación del término "confianza"

- **Antes de tomar la muestra** tiene sentido hablar de la probabilidad de que el intervalo, una vez que se construya, contenga al parámetro desconocido.
- **Después de tomar la muestra y calcular el intervalo** éste contendrá o no contendrá al parámetro, pero no tiene sentido hablar de probabilidad porque el experimento ya se ha realizado.
- En la práctica, si la probabilidad *a priori* de capturar el parámetro en el intervalo es $1-\alpha$, si tomáramos muchas muestras de tamaño n y para cada una construyéramos el intervalo $[\theta_1(\omega), \theta_2(\omega)]$, podríamos esperar que el $100(1-\alpha)\%$ de dichos intervalos contuviese al parámetro θ.
- Pero como en realidad solo tenemos una muestra y construimos un único intervalo **no sabemos** si este intervalo contiene o no a θ, pero *confiamos* en que sea uno de entre el $100\,(1-\alpha)\,\%$ de intervalos que contienen al parámetro. De ahí que valoremos nuestra confianza en $1-\alpha$

Ejercicio

- Podemos simular la obtención de una muestra de tamaño $n = 40$ de una variable $X \approx N\left(\mu = 80, \sigma = 10\right)$ mediante el siguiente código R:

```
muestra=rnorm(40,80,10)
```

- Podemos también obtener un intervalo de confianza para μ a partir de esta muestra mediante:

```
t.test(muestra)$conf.int
```

```
## [1] 79.08081 85.72290
## attr(,"conf.level")
## [1] 0.95
```

- Como vemos, en este caso el valor del parámetro μ ha quedado contenido en el intervalo.

- Vamos a repetir el proceso.

Ejercicio

- El siguiente código repite 100 veces el proceso de generar una muestra de tamaño 40 y construir el intervalo de confianza correspondiente; los intervalos resultantes se guardan en el objeto `intervalos`:

```
intervalos <- replicate(100,t.test(rnorm(40,80,10))$conf.int)
```

- Podemos ver los primeros intervalos:

```
intervalos[,1:6]
```

```
##          [,1]     [,2]     [,3]     [,4]     [,5]     [,6]
## [1,] 77.81020 78.95325 75.62774 78.74789 77.05240 76.54074
## [2,] 85.56283 85.63792 82.90436 84.47362 83.33463 82.39630
```

- O usar la siguiente sintaxis para verlos todos de una forma más cómoda:

```
array(sprintf("[%.2f, %.2f]", intervalos[1,], intervalos[2,]),
      dim=c(20,5))
```

Ejercicio

Los 100 intervalos correspondientes a las 100 muestras de tamaño 40 son los siguientes. ¿Contienen todos a la media poblacional $\mu = 80$?:

[77.81, 85.56]	[76.29, 83.25]	[78.33, 85.28]	[77.94, 84.21]	[77.37, 83.32]	[75.94, 82.04]
[78.95, 85.64]	[81.20, 86.73]	[77.46, 83.72]	[75.98, 82.34]	[77.67, 83.53]	[77.40, 83.32]
[75.63, 82.90]	[78.38, 84.38]	[77.12, 83.92]	[76.32, 82.70]	[75.36, 81.92]	[78.05, 84.15]
[78.75, 84.47]	[75.47, 81.86]	[76.11, 82.33]	[75.22, 81.65]	[78.75, 85.43]	[76.03, 81.27]
[77.05, 83.33]	[75.44, 81.27]	[75.30, 82.36]	[74.22, 80.83]	[75.85, 82.27]	[79.01, 86.39]
[76.54, 82.40]	[73.07, 78.58]	[78.33, 83.57]	[76.15, 81.94]	[74.86, 81.09]	[77.21, 84.06]
[77.81, 84.30]	[77.24, 83.31]	[76.21, 83.06]	[77.51, 83.16]	[77.50, 83.24]	[77.57, 83.79]
[78.39, 84.66]	[75.46, 81.43]	[74.39, 81.75]	[76.77, 82.01]	[75.30, 80.62]	[77.47, 83.62]

Ejercicio

Podemos comprobar que no todos los intervalos contienen al parámetro $\mu = 80$:

[77.81, 85.56]	[76.29, 83.25]	[78.33, 85.28]	[77.94, 84.21]	[77.37, 83.32]	[75.94, 82.04]
[78.95, 85.64]	**[81.20, 86.73]**	[77.46, 83.72]	[75.98, 82.34]	[77.67, 83.53]	[77.40, 83.32]
[75.63, 82.90]	[78.38, 84.38]	[77.12, 83.92]	[76.32, 82.70]	[75.36, 81.92]	[78.05, 84.15]
[78.75, 84.47]	[75.47, 81.86]	[76.11, 82.33]	[75.22, 81.65]	[78.75, 85.43]	[76.03, 81.27]
[77.05, 83.33]	[75.44, 81.27]	[75.30, 82.36]	[74.22, 80.83]	[75.85, 82.27]	[79.01, 86.39]
[76.54, 82.40]	**[73.07, 78.58]**	[78.33, 83.57]	[76.15, 81.94]	[74.86, 81.09]	[77.21, 84.06]
[77.81, 84.30]	[77.24, 83.31]	[76.21, 83.06]	[77.51, 83.16]	[77.50, 83.24]	[77.57, 83.79]
[78.39, 84.66]	[75.46, 81.43]	[74.39, 81.75]	[76.77, 82.01]	[75.30, 80.62]	[77.47, 83.62]

Ejercicio

- R puede utilizarse para contar "de forma automática" cuantos de estos intervalos contienen al parámetro, en este caso, al valor de la media poblacional $\mu = 80$:

```
contieneMu <- which(intervalos[1,]<=80 & 80<=intervalos[2,])
length(contieneMu)
```

```
## [1] 94
```

- En este ejemplo, 94 de 100 intervalos (el 94%) contienen al verdadero valor del parámetro.
- Si repetimos el proceso con otras muestras distintas veremos que siempre obtenemos valores de cobertura (proporción de intervalos que contienen al parámetro) próximos al 95%

Intervalos de confianza más habituales

Esperanza μ de una variable X con distribución normal

$$\mu \in \left[\bar{X} - \frac{S}{\sqrt{n}} t_{n-1,\alpha/2}, \bar{X} + \frac{S}{\sqrt{n}} t_{n-1,\alpha/2} \right]$$

Varianza σ^2 de una variable X con distribución normal

$$\sigma^2 \in \left[\frac{(n-1)\, S^2}{\chi^2_{n-1,\alpha/2}}, \frac{(n-1)\, S^2}{\chi^2_{n-1,1-\alpha/2}} \right]$$

Cociente de varianzas σ_1^2/σ_2^2 de variables normales.

$$\frac{\sigma_1^2}{\sigma_2^2} \in \left[\frac{S_1^2/S_2^2}{F_{n_1-1,n_2-1,\alpha/2}}, \frac{S_1^2/S_2^2}{F_{n_1-1,n_2-1,1-\alpha/2}} \right]$$

Diferencia de medias de variables normales independientes

$$\mu_1 - \mu_2 \in \left[(\bar{X}_1 - \bar{X}_2) \pm t_{n,\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \right]$$

$$s_p = \sqrt{\frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2}{n_1 + n_2 - 2}}$$

$$n = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)^2}{\left(\frac{s_1^2}{n_1} \right)^2 \frac{1}{n_1 - 1} + \left(\frac{s_2^2}{n_2} \right)^2 \frac{1}{n_2 - 1}}$$

Diferencia de medias de variables normales emparejadas

$$\mu_1 - \mu_2 \in \left[(\bar{X}_1 - \bar{X}_2) \pm t_{n-1,\alpha/2} \frac{S_D}{\sqrt{n}} \right]$$

donde:

$$S_D = \sqrt{S_1^2 + S_2^2 - 2S_{12}} = \sqrt{S_1^2 + S_2^2 - 2rS_1S_2}$$

Diferencia de medias de variables no normales independientes (sólo si $n_1 \geq 30, n_2 \geq 30$).

$$\mu_1 - \mu_2 \in \left[(\bar{X}_1 - \bar{X}_2) \pm z_{\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \right]$$

Diferencia de medias de variables no normales emparejadas (sólo si $n \geq 60$)

$$\mu_1 - \mu_2 \in \left[(\bar{X}_1 - \bar{X}_2) \pm z_{\alpha/2} \frac{S_D}{\sqrt{n}} \right]$$

siendo $S_D = \sqrt{S_1^2 + S_2^2 - 2S_{12}} = \sqrt{S_1^2 + S_2^2 - 2rS_1S_2}$

Intervalo de confianza para la proporción π de éxitos en una $B(n, \pi)$.

Sean n el tamaño de la muestra y N_E el número de éxitos observados.

- **Método de Agresti-Coull** (sólo si $n > 40$, $N_E \geq 5$, $n - N_E \geq 5$)

Llamando: $\tilde{N}_E = N_E + z^2_{\alpha/2}/2$, $\tilde{n} = n + \left(z_{\alpha/2}\right)^2$ y $\tilde{\pi} = \tilde{N}_E/\tilde{n}$:

$$\pi \in \left[\tilde{\pi} \pm z_{\alpha/2}\sqrt{\frac{\tilde{\pi}\,(1-\tilde{\pi})}{\tilde{n}}}\right]$$

- **Método de Clopper y Pearson** (cuando no se puede aplicar el anterior)

Llamando: $F_1 = F_{2(n-N_E+1),2N_E,\alpha/2}$ y $F_2 = F_{2(N_E+1),2(n-N_E),\alpha/2}$:

$$\pi \in \left[\frac{N_E}{(n-N_E+1)F_1 + N_E}, \frac{(N_E+1)F_2}{(n-N_E) + (N_E+1)F_2}\right]$$

Comparación de proporciones de dos variables binomiales X_1 y X_2

Se han realizado n_1 observaciones de X_1 con N_{E_1} éxitos y n_2 observaciones de X_2 con N_{E_2} éxitos. Sean $\hat{\pi}_1 = \frac{N_{E_1}}{n_1}$ y $\hat{\pi}_2 = \frac{N_{E_2}}{n_2}$ las proporciones observadas de éxitos. Si $n_1 \geq 30$, $n_2 \geq 30$:

- **Intervalo de confianza para la diferencia de proporciones**

$$\left[(\pi_1 - \pi_2) \pm \left(z_{\alpha/2}\sqrt{\frac{\hat{\pi}_1\left(1-\hat{\pi}_1\right)}{n_1} + \frac{\hat{\pi}_2\left(1-\hat{\pi}_2\right)}{n_2}} + \frac{1}{2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right)\right]$$

- **Intervalo de confianza para el cociente de proporciones**

$$\left(\frac{\pi_1}{\pi_2}\right) \in \left[\left(\frac{\hat{\pi}_1}{\hat{\pi}_2}\right) \cdot \exp\left(\pm z_{\alpha/2}\sqrt{\frac{(1-\hat{\pi}_1)}{n_1\hat{\pi}_1} + \frac{(1-\hat{\pi}_2)}{n_2\hat{\pi}_2}}\right)\right]$$

Intervalo de confianza para la esperanza μ de una distribución exponencial

$$\mu \in \left[\frac{2n\bar{X}}{\chi^2_{2n,\alpha/2}}, \frac{2n\bar{X}}{\chi^2_{2n,1-\alpha/2}} \right]$$

Intervalo de confianza para el parámetro λ de una distribución de Poisson.

$$\lambda \in \left[\frac{1}{2n}\chi^2_{n_1,1-\alpha/2}, \frac{1}{2n}\chi^2_{n_2,\alpha/2} \right]$$

siendo: $n_1 = 2T, \quad n_2 = 2(T+1), \quad T = \sum_{i=1}^{n} X_i$