4 Intervalos de confianza

- Nos interesa dar una medida de la precisión de la estimación que hemos hecho del parámetro. Daremos un rango de valores entre los que debería encontrarse el verdadero valor del parámetro.
- *Intervalo de confianza*: rengo de valores entre los que posiblemente se encuentre el verdadero valor del parámetro.
- Un intervalo de confianza tiene una *probabilidad* de 1 (si contiene al parámetro) ó 0 (si no lo contiene), y una *confianza* del 95% de contener al parámetro.

4.1 Intervalos de confianza para una población normal

- Intervalo de confianza para la media de una población normal con varianza conocida: Sea $X \sim \mathcal{N}(\mu, \sigma^2)$ con varianza σ^2 conocida. Entonces $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$, con lo que $P(-z_{\alpha/2} < \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} < z_{\alpha/2}) = 1-\alpha$. Si despejamos μ obtenemos lo siguiente (siendo $\overline{x}$ una realización particular de $\overline{X}$):

$$IC(\mu)_{(1-\alpha)100\%} = [\overline{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}]$$

- Intervalo de confianza para la media de una población normal con varianza desconocida: Sea $X \sim \mathcal{N}(\mu, \sigma^2)$ con varianza σ^2 desconocida. Entonces $\frac{\overline{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}$, luego $P(-t_{n-1,\alpha/2} < \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} < t_{n-1,\alpha/2}) = 1-\alpha$. Si $\overline{x}$ es una realización particular de $\overline{X}$ y s una realización particular de S, obtenemos:

$$IC(\mu)_{(1-\alpha)100\%} = [\overline{x} - t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}, \overline{x} + t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}]$$

- Intervalo de confianza para la varianza de una población normal: Sea $X \sim \mathcal{N}(\mu, \sigma^2)$. Como sabemos que $\frac{n-1}{\sigma^2}S^2 \sim \mathcal{X}^2_{n-1}$, tenemos que $P(-\mathcal{X}^2_{n-1,1-\alpha/2} < \frac{n-1}{\sigma^2}S^2 < \mathcal{X}^2_{n-1,\alpha/2}) = 1-\alpha$, luego, si s es una realización particular de S

$$IC(\sigma^2)_{(1-\alpha)100\%} = [\frac{n-1}{\mathcal{X}^2_{n-1,\alpha/2}}s^2, \frac{n-1}{\mathcal{X}^2_{n-1,1-\alpha/2}}s^2]$$

4.2 Intervalos de confianza para dos poblaciones normales

- Extracción de muestras: *datos pareados* (las muestras se eligen por pares y cada par de variables corresponde al mismo elemento de la muestra) o *muestras independientes* (las muestras de cada población son independientes entre sí)
- Intervalo de confianza para la diferencia de medias de dos poblaciones normales basado en datos pareados: Sean $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$ e $Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$. Entonces la variable $D = X - Y$ es $\mathcal{N}(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2 - 2\cdot cov(X,Y))$, con lo que $\frac{\overline{D}-(\mu_1-\mu_2)}{S_d/\sqrt{n}} \sim t_{n-1}$ y tenemos que $\frac{\overline{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}$, luego $P(-t_{n-1,\alpha/2} < \frac{\overline{D}-(\mu_1-\mu_2)}{S_d/\sqrt{n}} < t_{n-1,\alpha/2}) = 1-\alpha$. Si $\overline{d}$ es una realización particular de $\overline{D}$ y s_d una realización particular de S_d tenemos:

$$IC(\mu_1-\mu_2)_{(1-\alpha)100\%} = [\overline{d} - t_{n-1,\alpha/2}\frac{s_d}{\sqrt{n}}, \overline{d} + t_{n-1,\alpha/2}\frac{s_d}{\sqrt{n}}]$$

- Intervalo de confianza para la diferencia de medias de dos poblaciones normales con varianzas conocidas basado en muestras independientes: Sean $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$ e $Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$, con varianzas $\sigma_1^2 y \sigma_2^2$ conocidas. Sabemos que $\overline{X} - \overline{Y} \sim \mathcal{N}(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})$, donde n_i es el tamaño de la muestra i, con lo que $P(-z_{\alpha/2} < \frac{\overline{X}-\overline{Y}-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} < z_{\alpha/2}) = 1-\alpha$. De esta manera:

$$IC(\mu_1-\mu_2)_{(1-\alpha)100\%} = [\overline{x} - \overline{y} - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, \overline{x} - \overline{y} + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}]$$

- Intervalo de confianza para la diferencia de medias de dos poblaciones normales con varianzas desconocidas pero iguales basado en muestras independientes: Sean $X \sim \mathcal{N}(\mu_1, \sigma^2)$ e $Y \sim \mathcal{N}(\mu_2, \sigma^2)$, con varianzas desconocidas e iguales a σ^2. En este caso $\overline{X} - \overline{Y} \sim \mathcal{N}(\mu_1 - \mu_2, \sigma^2 \frac{n_1+n_2}{n_1 n_2})$, con lo que $\frac{\overline{X}-\overline{Y}-(\mu_1-\mu_2)}{\sqrt{\sigma^2 \frac{n_1+n_2}{n_1 n_2}}} \sim \mathcal{N}(0,1)$.

  Para crear el intervalo de confianza sustituiremos el valor de σ^2 por una ponderación de las dos cuasi-varianzas muestrales $S^2 = \frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$, ya que de esta manera obtenemos que $\frac{\overline{X}-\overline{Y}-(\mu_1-\mu_2)}{\sqrt{S^2 \frac{n_1+n_2}{n_1 n_2}}} \sim t_{n-2}$.

  Por tanto, $P(-t_{n-2,\alpha/2} < \frac{\overline{X}-\overline{Y}-(\mu_1-\mu_2)}{\sqrt{S^2 \frac{n_1+n_2}{n_1 n_2}}} < t_{n-2,\alpha/2}) = 1-\alpha$, y así

$$IC(\mu_1 - \mu_2)_{(1-\alpha)100\%} = [\overline{x} - \overline{y} - t_{n-2,\alpha/2}\sqrt{s^2 \frac{n_1+n_2}{n_1 n_2}}, \overline{x} - \overline{y} + t_{n-2,\alpha/2}\sqrt{s^2 \frac{n_1+n_2}{n_1 n_2}}]$$

4.3 Intervalos de confianza para muestras grandes

- Supondremos que el tamaño de la muestra es suficientemente grande y así poder utilizar el Teorema Central del Límite.
- Intervalo de confianza para una proporción: como ya hemos visto, la proporción muestral $\hat{p}$ coincide con la media muestral $\bar{x}$, luego por el TCL tenemos que $\frac{\hat{p}-p}{\sqrt{\hat{p}(1-\hat{p})/n}} \sim \mathcal{N}(0,1)$, con lo que

$$IC(p)_{(1-\alpha)100\%} = [\hat{p} - z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}, \hat{p} + z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}]$$

- Intervalo de confianza para la diferencia de dos proporciones: por el TCL tenemos que $\frac{\hat{p_1}-\hat{p_2}-(p_1-p_2)}{\sqrt{\hat{p_1}(1-\hat{p_1})/n_1+\hat{p_2}(1-\hat{p_2})/n_2}} \sim \mathcal{N}(0,1)$, luego

$$IC(p_1 - p_2)_{(1-\alpha)100\%} = [\hat{p_1} - \hat{p_2} \pm z_{\alpha/2}\sqrt{\hat{p_1}(1-\hat{p_1})/n_1 + \hat{p_2}(1-\hat{p_2})/n_2}]$$

- Intervalo de confianza para la media de cualquier distribución: por el TCL tenemos que

$$IC(\mu)_{(1-\alpha)100\%} = [\overline{x} - z_{\alpha/2}\frac{s}{\sqrt{n}}, \overline{x} + z_{\alpha/2}\frac{s}{\sqrt{n}}]$$

- Intervalo de confianza para un parámetro cualquiera de cualquier distribución: Sea $\hat{\theta}_{MV}$ el estimador de máxima verosimilitud de θ, entonces, independientemente de la distribución y siendo $L(\hat{\theta})$ la función soporte, tenemos que

$$IC(\theta)_{(1-\alpha)100\%} = [\hat{\theta} - z_{\alpha/2}\sqrt{\frac{-1}{L''(\hat{\theta})}}, \hat{\theta} + z_{\alpha/2}\sqrt{\frac{-1}{L''(\hat{\theta})}}]$$

4.4 Determinación del tamaño muestral

- Hemos visto que el intervalo de confianza para la media μ de una población normal con varianza conocida es $IC(\mu)_{(1-\alpha)100\%} = [\overline{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}]$.
- Así, el tamaño muestral necesario para garantizar una longitud determinada para el intervalo de confianza es $n = (\frac{z_{\alpha/2}\sigma}{A})^2$, siendo A la amplitud del intervalo (def. la mitad de la longitud del intervalo)