Tema 4

Introducción a la Inferencia. Estimación.

Introducción a la Inferencia. Distribuciones.

- Conceptos básicos de Inferencia.
- Distribuciones más importantes en el muestreo de poblaciones normales.
- Problemas que trata la Inferencia Estadística.
- Estimación:
  - Estimación puntual
  - Estimación por intervalo

Conceptos de Inferencia

- **Población:** Conjunto de entes en los se estudia una característica modelada por **X**.
- **Individuo:** Cada elemento de la población. X toma en cada individuo un **valor numérico concreto x**.
- **Parámetro:** Elemento desconocido de la distribución de carácter numérico (sea unidimensional o n-dimensional). Se representa por la **letra griega θ (theta)**. La distribución de X está determinada:
  - en variables discretas por $\{P(X = x_i; \theta)\}_i$.
  - en variables continuas por $f(x; \theta)$.

Ejemplo

- Una empresa eléctrica fabrica focos que tienen una duración distribuida de forma normal. Una muestra de 30 focos tiene una duración promedio de 780 horas, con una desviación de 40 horas.

Ejemplo conceptos de Inferencia

- **Población**: X = Duración, en horas, de un foco producido por la fábrica.
- **Individuo**: cada uno de los focos producidos, más concretamente el <u>valor de X en ese foco</u>, por ejemplo 782 horas de duración.
- La distribución de esta población (normal) tiene **dos parámetros**: media (μ) y desviación típica (σ) desconocidos.

Conceptos de Inferencia

- <u>**Muestra:**</u> Subconjunto finito de la población.

  Es un conjunto de **n variables: X_1, X_2,..., X_n**. Representa los valores de la población en los n individuos seleccionados **antes de ser elegidos.**

  Si se tienen individuos concretos seleccionados, **dejan de ser variables y pasan a ser valores numéricos**, denotados por x_1, x_2,..., x_n y denominados <u>**valor o realización muestral**</u>**.**

Conceptos de Inferencia

- <u>**Tamaño (de la población o de la muestra):**</u> número de individuos de la misma, se denota por n.
- <u>**Muestra aleatoria simple:**</u> Muestra en la que cada elemento se selecciona con independencia de los demás. Se representa cada observación por medio de una variable aleatoria X_i con distribución igual que X, obteniendo así una colección de variables **X_1, X_2, ..., X_n independientes y con la misma distribución que X.**

Ejemplo conceptos de Inferencia

- La **muestra** serían las **variables** X_1, ..., X_{30}, (**tamaño 30**) que representan la duración de 30 focos seleccionados aleatoriamente (muestra aleatoria simple), sin especificar en concreto de cuáles se trata.
- Una vez elegidos los 30 focos y comprobada su duración, tendríamos el **valor muestral**, que serían 30 **valores** numéricos, por ejemplo, 782 horas, 800 horas, ..., 650 horas.

Estadístico

- **<u>Estadístico:</u>** es una función de la muestra.

  Es una **variable aleatoria** $T= T(X_1, X_2, ..., X_n)$, cuyos **valores** particulares, t, son las correspondientes funciones de los valores muestrales: $t = T(x_1, x_2, ..., x_n)$.
- Estadísticos más usuales:
  - Media muestral.
  - Proporción muestral.
  - Varianza muestral (desviación muestral).

Media muestral

- El estadístico más frecuente es la v.a. **media muestral**:

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

que toma como valores las medias aritméticas de los valores muestrales $x_1, x_2, ..., x_n$, esto es:

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}.$$

Proporción muestral

- En particular, si las variables $\mathbf{X_i}$ **son de Bernoulli** (sólo valores 0 ó 1), la media muestral recibe el nombre de **proporción muestral**, y es la v.a.:

$$\hat{p} = \frac{\sum_{i=1}^{n} X_i}{n},$$

que toma como valores las proporciones de éxitos obtenidos en muestras concretas.

Ejemplo media muestral

- En el ejemplo el estadístico media muestral sería:

$$\overline{X} = \frac{X_1 + \ldots + X_{30}}{30},$$

y un valor concreto de este estadístico sería:

$$\overline{x} = \frac{782 + 800 + \ldots + 650}{30} = 780 \quad \text{horas.}$$

Varianza muestral

- Otro estadístico de uso frecuente es la v.a. **varianza muestral:**

$$S^2 = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{n-1},$$

que toma como valores las varianzas de los valores muestrales x_1, x_2, ..., x_n, esto es:

$$s^2 = \frac{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}{n-1}.$$

Desviación muestral

- A la v.a. raíz cuadrada de la varianza muestral se le llama **desviación muestral:**

$$S = \sqrt{\frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{n-1}},$$

que toma como valores las desviaciones de los valores muestrales x_1, x_2, ..., x_n, esto es:

$$s = \sqrt{\frac{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}{n-1}}.$$

Ejemplo varianza muestral

- En el ejemplo el estadístico varianza muestral sería:

$$S^2 = \frac{\sum_{i=1}^{30}\left(X_i - \overline{X}\right)^2}{29},$$

y un valor concreto de este estadístico sería:

$$\frac{(782-780)^2 + (800-780)^2 + \ldots + (650-780)^2}{29} = 40^2 \text{ horas}^2,$$

por lo que el valor de la desviación muestral S es s=40 horas.

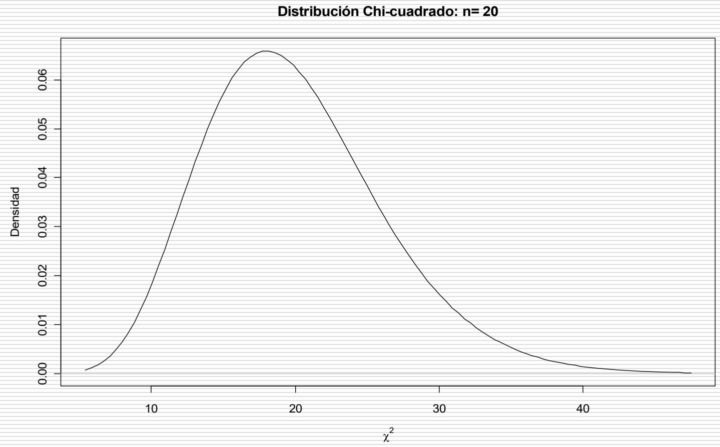
Muestreo en poblaciones Normales

- Cuando la variable X que representa la población tiene distribución **N(μ, σ)** (o si se manejan muestras con $n \geq 30$) se emplean con frecuencia, en problemas inferenciales, **estadísticos cuya distribución está relacionada**, aparte de con la propia distribución normal, con otras distribuciones de tipo continuo:
  - Distribución χ^2 de Pearson.
  - Distribución t de Student.

Distribución χ^2 de Pearson

- Es un modelo de tipo continuo.
- Se lee distribución "ji-dos" o "chi-cuadrado".
- La distribución χ^2 de Pearson con n grados de libertad, $\boldsymbol{\chi^2_n}$, se construye como suma de los cuadrados de n variables independientes con distribución normal N(0,1).
- Su representación gráfica depende de los grados de libertad n.

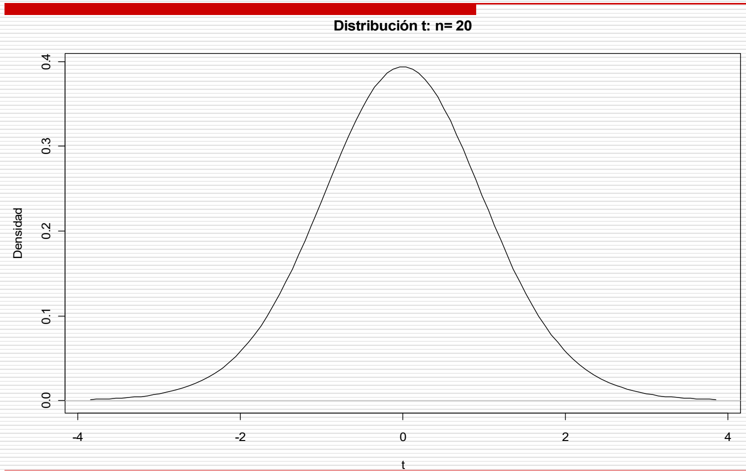
Grafica de χ^2 de Pearson



Distribución t de Student

- Es un modelo de tipo continuo.
- La distribución t de Student con n grados de libertad, $\mathbf{t_n}$, es el cociente entre una variable N(0,1) y la raíz cuadrada de una variable χ^2, independiente de la normal, dividida por su número de grados de libertad.
- Su representación gráfica depende de los grados de libertad n, pero se asemeja bastante a la campana de Gauss. En particular, es **simétrica** respecto al eje de ordenadas.

Grafica de t de Student



Distribuciones de estadísticos

- Si el muestreo se realiza sobre una población con $X \sim$ **N(μ, σ)** (o se maneja **un tamaño de muestra n≥30**), se pueden obtener las distribuciones de estadísticos relacionados con la media muestral, proporción muestral y varianza muestral.

Distribuciones media muestral

- Relacionados con la media muestral:

$$T = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}} \sim N(0,1) \quad \text{y} \quad T = \frac{\overline{X} - \mu}{S / \sqrt{n}} \sim t_{n-1}.$$

- Para el caso particular de proporción muestral:

$$T = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0,1) \quad \text{y} \quad T = \frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \sim N(0,1).$$

Distribución varianza muestral

- Relacionada con la varianza muestral se tiene que :

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

Problemas que soluciona la Inferencia

- Se parte del desconocimiento total o parcial de la distribución de la población X (si ésta se encontrara especificada en su totalidad carecería de sentido cualquier tipo de inferencia).
- Esta ignorancia puede ser relativa fundamentalmente a dos aspectos:
  - Por el desconocimiento de un parámetro θ del que depende la distribución de la variable X→**Estimación.**
  - Sobre la "certeza" o "falsedad" de una hipótesis realizada sobre algún aspecto de la distribución de variable X →**Contraste de hipótesis.**

Estimación

- No se conoce el valor de un parámetro θ del que depende la distribución de la variable X a través de su función de probabilidad o densidad y se pretende **estimarlo:**
  - asignando un valor al parámetro desconocido, por medio de la **estimación puntual**, con la exigencia de ciertas "garantías" de que la asignación es, por lo menos, "razonablemente buena".
  - determinando un intervalo en el que, nuevamente con algunas "garantías", se encuentre el valor desconocido; es lo que se denomina la **estimación por intervalo.**

Contraste de hipótesis

- Se pretende decidir si puede considerarse "cierta" o "falsa" una proposición (hipótesis) por medio de algún test estadístico, efectuando un contraste de hipótesis:
  - Si la hipótesis formulada se refiere al valor que toma un parámetro de la distribución, se habla de **contrastes paramétricos**.
  - En caso contrario nos referimos a **contrastes no paramétricos.** En ellos se desconoce si las condiciones de la distribución de X o las de obtención de la muestra son las adecuadas para formular los modelos que resuelvan los problemas paramétricos anteriores.

Observación a la Inferencia

- En el estudio particular de cada procedimiento deben explicarse claramente qué **"garantías"** (medidas en términos de probabilidad o con la presentación de propiedades "deseables") se dan de las conclusiones expuestas.
- El **riesgo de error** es más fuerte de lo que en general se cree. Además, en el mejor de los casos, las conclusiones con sus limitaciones serán válidas mientras se esté trabajando con **la muestra aleatoria**, pero dejarán de serlo cuando se descienda a la **muestra particular**.

Estimación puntual

- En la estimación puntual se utiliza un **estimador** que da lugar, a partir de los datos de la muestra, a un valor "aproximado" del parámetro desconocido de la distribución. Este valor "aproximado" recibe el nombre de **estimación del parámetro**.
- Hay que tener presente que, en el caso de la estimación puntual, nos planteamos una de estas situaciones:
  - Se realiza un muestreo sobre una población X con parámetros μ (media) y σ (desviación típica), con X población normal o con un tamaño de muestra $n \geq 30$.
  - Se realiza un muestreo (con un tamaño de muestra $n \geq 30$) sobre una población X que sigue una distribución de Bernoulli de parámetro p.

Estimador

- Nos proponemos dar un valor "aproximado" del parámetro o parámetros desconocidos. Para ello utilizamos un estimador, que es un estadístico que cumple unas propiedades deseables entre las que destacamos:
  - ser centrado (la media del estimador coincide con el parámetro a estimar).
  - tener varianza lo más pequeña posible.

Estimadores usuales

Parámetro que se estima	Estimador
Media poblacional (μ)	Media muestral $\overline{X} = \dfrac{\sum_{i=1}^{n} X_i}{n}$
Varianza poblacional (σ^2)	Varianza muestral $S^2 = \dfrac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n-1}$
Proporción poblacional (p)	Proporción muestral $\hat{p} = \dfrac{\sum_{i=1}^{n} X_i}{n}$

Ejemplo y nota

- En el ejemplo:
  - Estimación puntual de la media μ: 780 horas
  - Estimación puntual de la desviación típica σ: 40 horas.
- Hay que señalar que con la estimación puntual es muy improbable que obtengamos el verdadero valor buscado, por lo que parece razonable dar un conjunto mayor de posibles valores para el parámetro. Ese conjunto vendrá dado mediante un intervalo, lo que examinaremos en el siguiente apartado.

Intervalo de confianza

- Se acota el valor del parámetro entre dos valores con alguna **garantía** expresada en términos de **probabilidad**. Los **valores considerados deben ser aleatorios** para que tenga sentido hablar de una probabilidad.
- Dicha **probabilidad** recibe el nombre de **coeficiente de confianza** (o nivel de confianza). Se representa por $\mathbf{1-\alpha}$ y suele ser un valor próximo a uno.

Cálculo de un intervalo de confianza

- El cálculo del intervalo de confianza para un parámetro θ a un nivel $1-\alpha$ se realiza siguiendo estos pasos:

1. Seleccionar el estadístico T a utilizar y conocer su distribución.
2. Calcular los valores $\lambda_{\alpha/2}$ y $\lambda_{1-\alpha/2}$, tales que:
$$1-\alpha = P(\lambda_{\alpha/2} \le T \le \lambda_{1-\alpha/2}).$$
3. Operar hasta llegar a una expresión del tipo:
$$P(\theta \in [L_1, L_2]) = 1-\alpha.$$
4. Particularizar a los valores muestrales.

Tipos de intervalos de confianza

- Calcularemos intervalos de confianza en los siguientes casos:
  - Para la **media** de una población.
  - Para la **varianza (desviación típica)** de una población.
  - Para la **proporción** de una población Bernoulli.

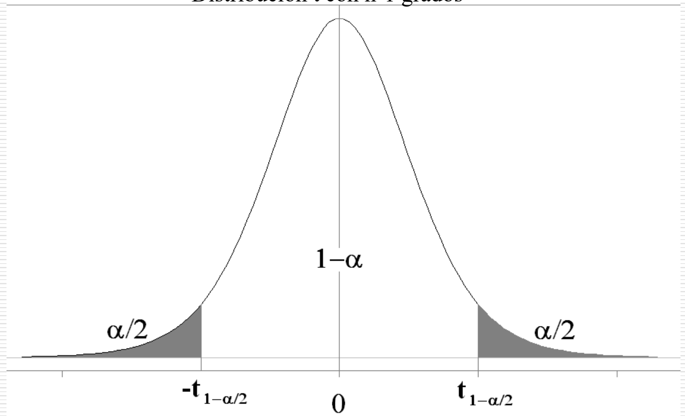
Intervalo para la media I

- $X \sim N(\mu,\sigma)$ o tengo una muestra con $n \ge 30$, con media (μ) y desviación (σ) son desconocidas. Para calcular el **intervalo de confianza al nivel $1-\alpha$ para la media μ** realizo los siguientes pasos:

1. Estadístico:
$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}.$$
2. Calcular los valores $t_{\alpha/2}$ y $t_{1-\alpha/2}$, tales que:
$$1-\alpha = P(t_{\alpha/2} \le t_{n-1} \le t_{1-\alpha/2}).$$
   - En realidad (por la simetría de la t): $t_{\alpha/2} = -t_{1-\alpha/2}$.

Intervalo para la media (gráfico)

Distribución t con n-1 grados



Intervalo para la media II

3. Operar para tener $P(\mu \in [L_1,L_2])=1-\alpha$. Para ello sustituimos en la expresión de 2. la distribución por el estadístico y obtenemos:

$$1-\alpha = P\left(-t_{1-\alpha/2} \le \frac{\overline{X}-\mu}{S/\sqrt{n}} \le t_{1-\alpha/2}\right) = P\left(\overline{X}-t_{1-\alpha/2}\, S/\sqrt{n} \le \mu \le \overline{X}+t_{1-\alpha/2}\, S/\sqrt{n}\right) =$$

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

intervalo simétrico respecto a la media muestral y con amplitud $2 \cdot t_{1-\alpha/2}\, S/\sqrt{n}$.

4. Particularizar a los valores muestrales.

Comentarios a la interpretación de I.C.

- En el intervalo aleatorio obtenido en el paso 3. tiene sentido afirmar que el parámetro μ pertenece a este intervalo con probabilidad $1-\alpha$, ya que los extremos del intervalo son variables aleatorias.
- En el intervalo numérico obtenido en el paso 4. **no tiene sentido** hablar de **probabilidad** de pertenencia a este intervalo (ya no hay v.a.) ya que el parámetro estará con probabilidad 1 o no estará entre los dos números que limitan el intervalo.

Interpretación nivel de confianza

- El **nivel de confianza** $1-\alpha$ se interpreta como la **proporción de intervalos numéricos, construidos a partir del intervalo aleatorio, que contienen el verdadero valor del parámetro**.
- Es decir, si obtenemos "muchos" intervalos con distintas muestras, aproximadamente un porcentaje del $(1-\alpha)100$ % contiene el valor desconocido del parámetro.

Error máximo de la estimación

- A la cantidad

$$\varepsilon = t_{1-\alpha/2}\, S/\sqrt{n}$$

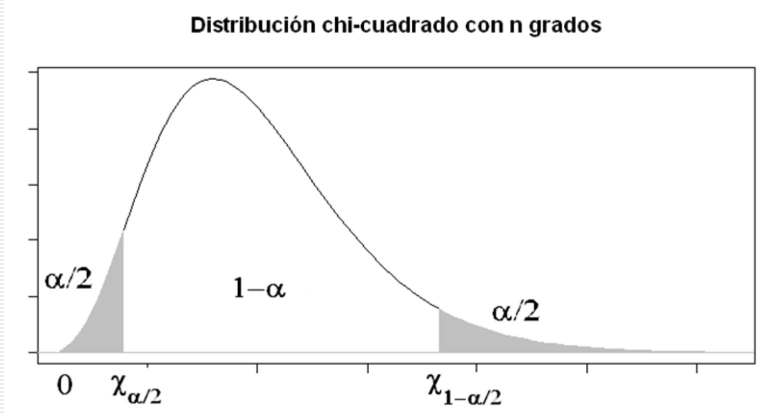
se le denomina **error máximo** cometido en la estimación; está relacionado, tal como se aprecia, con el nivel de confianza y con el tamaño de la muestra:
  - Manteniendo fijo el tamaño de la muestra, al aumentar el nivel de confianza aumenta el error
  - Manteniendo fijo el nivel de confianza, al aumentar n disminuye el error
  - Manteniendo fijo el error, al aumentar el tamaño de la muestra aumenta el nivel de confianza

Intervalo para la varianza I

- $X \sim N(\mu,\sigma)$ o tengo una muestra con $n \geq 30$, con media (μ) y desviación (σ) son desconocidas. Para calcular el **intervalo de confianza al nivel 1-α para la varianza σ ²** realizo los siguientes pasos:

1. Estadístico: $T = \dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$.
2. Calcular los valores $\chi_{\alpha/2}$ y $\chi_{1-\alpha/2}$, tales que:
$$1-\alpha = P(\chi_{\alpha/2} \leq t_{n-1} \leq \chi_{1-\alpha/2}).$$

Intervalo para la varianza (gráfico)



Intervalo para la varianza II

3. Operar para tener $P(\sigma^2 \in [L_1,L_2])=1-\alpha$. Para ello sustituimos en la expresión de 2. la distribución por el estadístico y obtenemos:
$$1-\alpha = P\left(\chi_{\alpha/2} \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi_{1-\alpha/2}\right) =$$
$$= P\left(\frac{(n-1)S^2}{\chi_{1-\alpha/2}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi_{\alpha/2}}\right) =$$
$$= P\left(\sigma^2 \in \left[\frac{(n-1)S^2}{\chi_{1-\alpha/2}}, \frac{(n-1)S^2}{\chi_{\alpha/2}}\right]\right).$$
4. Particularizar a los valores muestrales.

Comentarios al intervalo de confianza

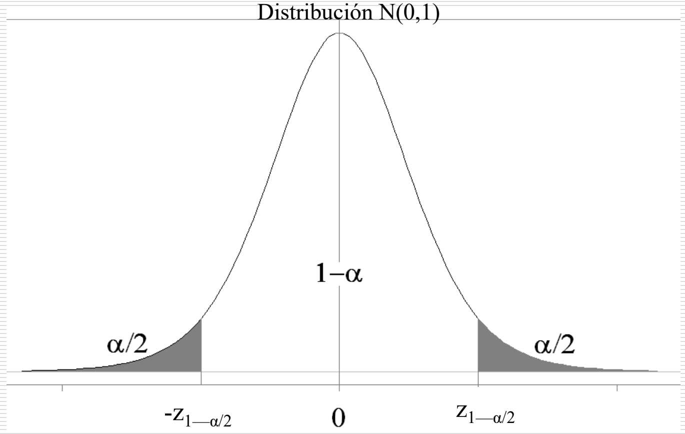
- Este intervalo **no es simétrico** respecto a la varianza muestral (la distribución chi-cuadrado tampoco lo es).
- Todas las apreciaciones hechas sobre el significado del nivel de confianza son válidas para todos los intervalos de confianza.

Intervalo para la proporción I

- X ~ B(p) y tengo una muestra con n≥30. Para calcular el **intervalo de confianza al nivel 1-α para p** realizo los siguientes pasos:

1. Estadístico: $T = \dfrac{\hat{p} - p}{\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}} \sim N(0,1).$
2. Calcular los valores $z_{\alpha/2}$ y $z_{1-\alpha/2}$, tales que:
   $$1-\alpha = P(z_{\alpha/2} \leq N(0,\ 1) \leq z_{1-\alpha/2}).$$
   - En realidad (por la simetría de la N): $z_{\alpha/2} = -z_{1-\alpha/2}$.

Intervalo para la proporción (gráfico)



Intervalo para la proporción II

3. Operar para tener $P(p \in [L_1, L_2]) = 1-\alpha$. Para ello sustituimos en la expresión de 2. la distribución por el estadístico y obtenemos:

$$1-\alpha = P\left(-z_{1-\alpha/2} \leq \frac{\hat{p}-p}{\sqrt{\frac{\hat{p}\cdot(1-\hat{p})}{n}}} \leq z_{1-\alpha/2}\right) = P\left(\hat{p} - z_{1-\alpha/2}\sqrt{\frac{\hat{p}\cdot(1-\hat{p})}{n}} \leq p \leq \hat{p} + z_{1-\alpha/2}\sqrt{\frac{\hat{p}\cdot(1-\hat{p})}{n}}\right) =$$
$$= P\left(p \in \left[\hat{p} - z_{1-\alpha/2}\sqrt{\frac{\hat{p}\cdot(1-\hat{p})}{n}}, \hat{p} + z_{1-\alpha/2}\sqrt{\frac{\hat{p}\cdot(1-\hat{p})}{n}}\right]\right),$$

   intervalo simétrico respecto a la proporción muestral y con amplitud
   $$2\cdot z_{1-\alpha/2}\sqrt{\frac{\hat{p}\cdot(1-\hat{p})}{n}}.$$
4. Particularizar a los valores muestrales.

Error máximo y tamaño de muestra

- Se denomina **error máximo** cometido en la estimación a la cantidad
  $$\varepsilon = z_{1-\alpha/2}\sqrt{\frac{\hat{p}\,(1-\hat{p})}{n}}.$$
- En la expresión anterior se puede despejar n, y obtener el **tamaño muestral** mínimo para estimar p con un cierto error máximo y para un determinado nivel de confianza:
  $$n = \frac{z_{1-\alpha/2}^2\,\hat{p}(1-\hat{p})}{\varepsilon^2}.$$