TEMA 7: ESTADÍSTICA INFERENCIAL

Introducción:

Para estudiar una cierta característica de una población se recurre a una muestra de ella, es decir, un subconjunto que sea representativo de dicha población.

En este tema veremos como realmente pueden ser muy útiles para hacerse una idea global del comportamiento de la población. Veremos que características debe tener la muestra para que sea útil y algunos métodos para extraer muestras.

1. Muestras estadísticas. Tipos de muestreo:

- Población: conjunto de todos los individuos objetos de un estudio estadístico.
- Muestra: Subconjunto de dicha población. Su estudio sirve para "inferir" características de toda la población.
  Motivos para recurrir a una muestra:
  - Población muy numerosa: españoles con derecho a voto.
  - Población difícil de controlar: clientes en uno grandes almacenes.
  - Proceso de medición destructivo: vida media de una bombilla.
  - Rapidez en obtener resultados.

  ¿Se obtienen resultados razonables?: Ver fotos página 263.

- Muestreo: La elección de la muestra se llama muestreo. Debe tener en cuenta el tamaño y como se realiza la selección.

  Tipos: - Muestreo aleatorio: Los individuos de la muestra se eligen al azar.
  - Simple: Se enumeran los N individuos y se escogen n al azar.
  - Sistemático: se toma h = N/n. Se elige un individuo de los h primeros, y los restantes, de h en h.
  - Estratificado: Si la población se divide en estratos, se eligen de cada estrato un número de individuos de cada estrato proporcional al número total de cada uno de ellos.

  - Muestreo no aleatorio: No lo estudiaremos.

- Técnicas para obtener la muestra: En todos los tipos anteriores las extracciones se pueden realizar con o sin reemplazamiento.
  Ejs: - 200 españoles obtenidos de las listas del censo por ordenador: Con reemplazamiento.
  - 10 Números extraídos de una bolsa: Sin reemplazamiento.

2. Repaso de la distribución Normal:

- Uso de la tabla N(0, 1): $\Phi(K) = P[\, Z < k]$.
- Tipificación:

$$X \approx N(\mu,\sigma) \Rightarrow Z = \frac{X-\mu}{\sigma} \approx N(0,1)$$

3. Estimación puntual.

Trata de estudiar características de una población a través de los parámetros poblacionales (media, mediana, desviación típica, proporción,…). Cuando estos parámetros están referidos a una muestra de la población se llaman “estadísticos”.
Para inferir conclusiones sobre los parámetros poblacionales a partir de los estadísticos es necesario conocer como se distribuyen éstos últimos.

3.1. Distribución muestral de medias.

Consideremos que una variable estadística X de una cierta población de media μ y desviación típica σ.
Elegimos muestras de tamaño n, y consideramos la distribución de medias obtenidas en esas muestras (distribución de medias muestrales). Dicha distribución se parece a una distribución normal tanto más cuanto mayor sea n.
Aumentando n, la media μ se mantiene y la desviación típica σ va disminuyendo, es más:

Si tenemos una población con distribución normal N(μ , σ), la distribución de las medias muestrales de tamaño n, $\overline{X}$:

- Tiene la misma media μ que la población: $\mu\left(\overline{X}\right) = \mu$
- Su desviación típica es: $\sigma\left(\overline{X}\right) = \sigma / \sqrt{n}$

Cuando la población no sigue una distribución normal, aplicamos el Teorema central del límite:

$$\boxed{Si\ n \geq 30,\ \overline{X} \approx N\left(\mu , \sigma / \sqrt{n}\right).}$$

Observación: siempre las muestras serán de tamaño $n \geq 30$.

Consecuencias del TCL:

- Control de la suma de todos los individuos de la muestra:

$$\overline{x} = \frac{\sum x_i}{n} \Rightarrow \sum x_i = n \cdot \overline{x} \Rightarrow \sum x_i \approx N\left(n \cdot \mu , \sigma \cdot \sqrt{n}\right)$$

3.2. Distribución muestral de proporciones.

Consideremos una población de tamaño N, y supongamos que la proporción de individuos de esa población que tiene una cierta característica C, es p y q la de individuos que no la tienen $(q = 1 - p)$ Consideramos todas las muestras de tamaño n que de dicha población pueden extraerse. Cada muestra determina un estadístico muestral de la variable estudiada llamada Distribución muestral de proporciones P_r.

Las relaciones entre los parámetros de población y los estadísticos de P_r son.

$$\mu(P_r) = p \qquad y \qquad \sigma(P_r) = \sqrt{\frac{pq}{n}}$$

Si las muestras son de tamaño n>30 entonces dicha proporción sigue una distribución normal

$$\boxed{P_r \approx N\left(p, \sqrt{\frac{pq}{n}}\right)}$$

4. Estimación por intervalos.

La estimación puntual se utiliza poco, puesto que carecemos de datos que indiquen la fiabilidad del dato muestral que hemos tomado. La estimación por intervalos consiste en dar dos valores entre los cuales estimamos que se encontrará el parámetro poblacional con una probabilidad prefijada de antemano.

Llamamos intervalo de confianza al intervalo que, con una cierta probabilidad, contiene al parámetro que se está estimando.

Llamamos nivel de confianza a la probabilidad de que el intervalo de confianza contenga al verdadero valor del parámetro. Se suele expresar en porcentaje: **(1-α) %**. Se denota por N_c. Esto es: N_c **= (1-α) %.** A cada nivel de confianza le corresponde un $Z_{\alpha/2}$ llamado valor crítico correspondiente a la distribución normal N(0, 1) y que cumple:

$$P\left[|Z| \leq Z_{\alpha/2}\right] = N_c$$

CALCULO DEL VALOR CRÍTICO

$$P\left[|Z| \leq Z_{\alpha/2}\right] = N_c \rightarrow P\left[-Z_{\alpha/2} \leq Z \leq Z_{\alpha/2}\right] = N_c$$

$$P\left[Z \leq Z_{\alpha/2}\right] - P\left[Z \leq -Z_{\alpha/2}\right] = N_c$$

$$2 \cdot P\left[Z \leq Z_{\alpha/2}\right] - 1 = N_c$$

$$de\ donde \quad P\left[\leq Z \leq Z_{\alpha/2}\right] = \frac{1+N_c}{2}$$

PRINCIPALES VALORES CRÍTICOS

N_c	Z_c
0’90	1’645
0’95	1’96
0’99	2’575

4.1. Construcción del intervalo de confianza.

Parámetros	Intervalos de confianza I_c
Media	$\left(\mu - Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}, \mu + Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \right)$
Proporción	$\left(p - Z_{\alpha/2} \cdot \sqrt{\frac{p \cdot q}{n}},\ p + Z_{\alpha/2} \cdot \sqrt{\frac{p \cdot q}{n}} \right)$

5. Tamaño de las muestras

Los objetivos de un buen tamaño muestral son evitar el gsto excesivo de dinero al tomar muchos elementos maestrales y conseguir resultados fiables, delimitando el error máximo que queramos admitir.

5.1 Estimación de una media.

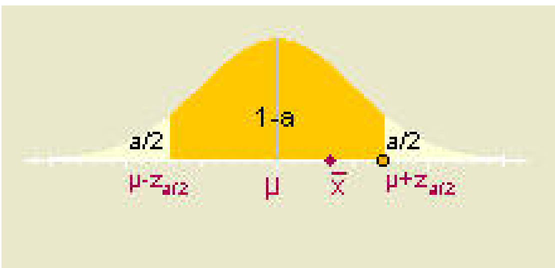
El error máximo que se puede cometer, al nivel de confianza N_c, es:

$$E = Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \quad \text{de donde} \quad n = \frac{Z_{\alpha/2}^{\ 2} \cdot \sigma^2}{E^2}$$

5.1 Estimación de una proporción.

El error máximo que se puede cometer, al nivel de confianza N_c, es:

$$E = Z_{\alpha/2} \cdot \sqrt{\frac{p \cdot q}{n}} \quad \text{de donde} \quad n = \frac{Z_{\alpha/2}^{\ 2} \cdot p \cdot q}{E^2}$$



6. Contraste de hipótesis.

Una **hipótesis estadística** es una afirmación respecto a alguna característica de una población. **Contrastar** una hipótesis es comparar las predicciones con la realidad que observamos. Si dentro del margen de error que nos permitimos admitir, hay coincidencia, aceptaremos la hipótesis y en caso contrario la rechazaremos.

- La hipótesis emitida se suele designar por **H_o** y se llama **Hipótesis nula** porque parte del supuesto que la diferencias entre el valor verdadero del parámetro y su valor hipotético es debida al azar, es decir no hay diferencia.
- La hipótesis contraria se designa por **H_1** y se llama **Hipótesis alternativa**

Los contrastes pueden ser **unilaterales** o **bilaterales** (también llamados de una o dos colas) según establezcamos las hipótesis, si las definimos en términos de igual y distinto estamos ante una hipótesis bilateral, si suponemos una dirección (en términos de mayor o menor) estamos ante uno unilateral.

Se trata pues, de extraer conclusiones a partir de una muestra aleatoria y significativa, que permitan aceptar o rechazar una hipótesis previamente emitida, sobre el valor de un parámetro desconocido de la población. El método que seguiremos es el siguiente:

1. **Enunciar** la hipótesis
2. **Elegir** un **nivel de significación a** y construir la **zona de aceptación**, intervalo fuera del cual sólo se encuentran el a100% de los casos más raros. A la zona de rechazo la llamaremos **región crítica**, y su área es el nivel de significación.
3. **Verificar** la hipótesis extrayendo una muestra cuyo tamaño se ha decidido en el paso anterior y obteniendo de ella el correspondiente estadístico (media o proporción en nuestro caso).
4. **Decidir**. Si el valor calculado en la muestra cae dentro de la zona de aceptación se acepta la hipótesis y si no se rechaza.

Aquí nos vamos a limitar a estudiar hipótesis sobre la media y sobre la proporción en una población. En cada caso se trabaja con un contraste bilateral y otro unilateral.

POSIBLES ERRORES EN EL CONTRASTE DE HIPÓTESIS

El contraste de hipótesis no establece la verdad de la hipótesis, sino un criterio que nos permite decidir si una hipótesis se acepta o se rechaza, o el determinar si las muestras observadas difieren significativamente de los resultados esperados. En este proceso podemos incurrir en dos tipos de errores según sea la situación real y la decisión que tomemos.

Si rechazamos una hipótesis cuando debiera ser aceptada, cometemos un **error de tipo I**, mientras que si la aceptamos debiendo ser rechazada diremos que hemos cometido un **error de tipo II**. Minimizar los errores no es una cuestión sencilla, un tipo suele ser más grave que otro y los intentos de disminuir uno suelen producir el aumento del otro. La única forma de disminuir ambos a la vez es aumentar el tamaño de la muestra.

	Ho verdadera	Ho falsa
DECISIÓN: Mantener Ho	Decisión correcta	Decisión incorrecta **Error de tipo II**
DECISIÓN: Rechazar Ho	Decisión incorrecta **Error de tipo I**	Decisión correcta

6.1. Contraste de hipótesis para la media

Queremos contrastar una hipótesis acerca del valor de la media poblacional a partir de los resultados de una muestra. El proceso que seguimos es:

Contraste bilateral

$H_0: \mu = \mu_o$ $\quad H_1: \mu \neq \mu_o$	***1) Establecer la hipótesis***
buscamos z_c tal que $P(-Z_{\alpha/2} \leq z \leq Z_{\alpha/2}) = 1-\alpha$  $\left(\mu - Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}, \mu + Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right)$	Las medias muestrales se distribuyen $\bar{x} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ ***2) Elegir el nivel de significación α y determinar la zona de aceptación a partir del*** Intervalo de confianza
$\bar{x} \in \left(\mu - Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}, \mu + Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right) \rightarrow aceptamos\ H_0$ $\bar{x} \notin \left(\mu - Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}, \mu + Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right) \rightarrow rechazamos\ H_0$	***3) Verificación*** ***4) Decisión***

Contraste unilateral H_0: $\mu \geq \mu_0$ H_1: $\mu < \mu_0$	***1) Establecer la hipótesis***	**Contraste unilateral** H_0: $\mu \leq \mu_0$ H_1: $\mu > \mu_0$
buscamos Z_α tal que $P(Z \geq Z_\alpha) = 1-\alpha$  $\left(\mu - Z_\alpha \cdot \frac{\sigma}{\sqrt{n}}, +\infty\right)$	Las medias muestrales se distribuyen $\bar{x} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ ***2) Elegir el nivel de significación α y determinar la zona de aceptación a partir del*** Intervalo de confianza	buscamos Z_α tal que $P(Z \leq Z_\alpha) = 1-\alpha$  $\left(-\infty, \mu + Z_\alpha \cdot \frac{\sigma}{\sqrt{n}}\right)$
$\bar{x} \in \left(\mu - Z_\alpha \cdot \frac{\sigma}{\sqrt{n}}, +\infty\right) \rightarrow aceptamos\ H_0$ $\bar{x} \notin \left(\mu - Z_\alpha \cdot \frac{\sigma}{\sqrt{n}}, +\infty\right) \rightarrow rechazamos\ H_0$	***3) Verificación*** ***4) Decisión***	$\bar{x} \in \left(-\infty, \mu + Z_\alpha \cdot \frac{\sigma}{\sqrt{n}}\right) \rightarrow aceptamos\ H_0$ $\bar{x} \notin \left(-\infty, \mu + Z_\alpha \cdot \frac{\sigma}{\sqrt{n}}\right) \rightarrow rechazamos\ H_0$

6.2. Contraste de hipótesis para la proporción

Queremos contrastar una hipótesis acerca de la proporción en una población a partir de los resultados de una muestra. El proceso que seguimos es:

Contraste bilateral

H_o: $p = p_o$ H_1: $p \neq p_o$	***1) Establecer la hipótesis***
buscamos z_c tal que $P(-Z_{\alpha/2} \leq Z \leq Z_{\alpha/2}) = 1-\alpha$  $\left(p - Z_{\alpha/2}\cdot\sqrt{\frac{pq}{n}}, p + Z_{\alpha/2}\cdot\sqrt{\frac{pq}{n}}\right)$	Las medias muestrales se distribuyen $p_r \sim N\left(p, \sqrt{\frac{pq}{n}}\right)$ ***2) Elegir el nivel de significación α y determinar la zona de aceptación a partir del*** Intervalo de confianza
$p \in \left(p - Z_{\alpha/2}\cdot\sqrt{\frac{pq}{n}}, p + Z_{\alpha/2}\cdot\sqrt{\frac{pq}{n}}\right) \rightarrow aceptamos\ H_0$ $p \notin \left(p - Z_{\alpha/2}\cdot\sqrt{\frac{pq}{n}}, p + Z_{\alpha/2}\cdot\sqrt{\frac{pq}{n}}\right) \rightarrow rechazamos\ H_0$	***3) Verificación*** ***4) Decisión***

Contraste unilateral H_0: $p \geq p_0$ H_1: $p < p_0$	***1) Establecer la hipótesis***	**Contraste unilateral** H_0: $p \leq p_0$ H_1: $p > p_0$
buscamos Z_α tal que $P(Z \geq Z_\alpha)=1-\alpha$  $\left(p - Z_\alpha \cdot \sqrt{\frac{pq}{n}}, +\infty\right)$	Las proporciones muestrales se distribuyen $p_r \sim N\left(p, \sqrt{\frac{pq}{n}}\right)$ ***2) Elegir el nivel de significación α y determinar la zona de aceptación a partir del*** Intervalo de confianza	buscamos Z_α tal que $P(Z \leq Z_\alpha)=1-\alpha$  $\left(-\infty, p + Z_\alpha \cdot \sqrt{\frac{pq}{n}}\right)$
$p \in \left(p - Z_\alpha \cdot \sqrt{\frac{pq}{n}}, +\infty\right) \rightarrow aceptamos\ H_0$ $p \notin \left(p - Z_\alpha \cdot \sqrt{\frac{pq}{n}}, +\infty\right) \rightarrow rechazamos\ H_0$	***3) Verificación*** ***4) Decisión***	$p \in \left(-\infty, p + Z_\alpha \cdot \sqrt{\frac{pq}{n}}\right) \rightarrow aceptamos\ H_0$ $p \notin \left(-\infty, p + Z_\alpha \cdot \sqrt{\frac{pq}{n}}\right) \rightarrow rechazamos\ H_0$