INFERENCIA ESTADÍSTICA

Tema 5: MUESTREO Y DISTRIBUCIONES MUESTRALES.

1 Introducción.

`Objetivo de la Inferencia Estadística:`

Estudiar una característica entre los individuos de una población de referencia

`¿Cómo lo conseguiremos?`

Estudiando la característica de interés entre los individuos que integran una **muestra** que es un subconjunto representativo de la población

Pretendemos generalizar las conclusiones que se obtengan estudiando la muestra a toda la población y dar una medida de la confianza de nuestras conclusiones.

Para ello debemos de **seleccionar al azar los individuos que integran la muestra**.

`Ejemplos:`

1. Altura de los jóvenes españoles.
2. Porporción de hogares españoles conectados a la red.

2 Planteamiento del problema de Inferencia Estadística. Muestra aleatoria simple.

`Objetivo`: *Estudiar una característica entre los individuos de una población de referencia.*

Consideramos el experimento aleatorio: "Seleccionamos un individuo al azar de la población" y definimos la variable aleatoria:

X=Valor de la característica de interés en ese individuo concreto

$\equiv$ **variable aleatoria poblacional**

Suponemos que la distribución de probabilidad de X es conocida $f_\theta(x)$ y depende de un parámetro θ (**parámetro poblacional**) que es desconocido.

Escuela Técnica Superior de Ingeniería Agronómica.
Universidad Politécnica de Cartagena.

¿Cómo podemos obtener valores para θ?

Repitiendo el experimento n veces de manera independiente y definiendo:

$$\begin{aligned} X_1 &= \text{Valor de } X \text{ obtenido en la realización 1 del experimento} \\ X_2 &= \text{Valor de } X \text{ obtenido en la realización 2 del experimento} \\ &\vdots \\ X_n &= \text{Valor de } X \text{ obtenido en la realización n del experimento} \end{aligned}$$

Las variables $(X_1, X_2, ..., X_n)$ son independientes y sus distribuciones de probabilidad coinciden con la distribución de probabilidad de X. Decimos que $(X_1, X_2, ..., X_n)$ es una **muestra aleatoria simple de tamaño n** (abreviadamente **m.a.s.**) de la distribución de X. Llamaremos a $(X_1, X_2, ..., X_n)$ las **variables muestrales.**

Se denotará por $(x_1, x_2, ..., x_n)$ a los valores de la muestra $(X_1, X_2, ..., X_n)$ para una realización concreta de la muestra y se denominará **realización de la muestra**.

`Ejemplos de planteamiento del problema de inferencia:`

1. Altura de los jóvenes españoles $\Longrightarrow$

   $$X\text{=Altura de los jóvenes españoles} \sim N(\mu, \sigma)$$

   Entonces el parámetro poblacional sería:

   $$\theta = (\mu, \sigma^2) \text{ o bien } \theta = (\mu, \sigma)$$

2. Porporción de hogares españoles conectados a la red $\Longrightarrow$

   $$X = \begin{cases} 1, & \text{si el hogar está conectado a la red con probabilidad } p \\ 0, & \text{en otro caso con probabilidad } 1-p \end{cases} \Rightarrow X \sim b(p)$$

   Entonces el parámetro poblacional sería:

   $$\theta = p$$

3 Estadístico. Estimador. Distribuciones muestrales.

Definición:

Dada la v.a. poblacional X con distribución de probabilidad $f_\theta(x)$ y una m.a.s. de tamaño n, $(X_1, X_2, ..., X_n)$, se denominará **estadístico** a cualquier función de la muestra que sea independiente del parámetro θ:

$$T(X_1, X_2, ..., X_n)$$

Ejemplos de estadísticos:

$$T(X_1, X_2, ..., X_n) = \sum_{i=1}^{n} X_i$$

$$T(X_1, X_2, ..., X_n) = X_n - X_1$$

$$T(X_1, X_2, ..., X_n) = \frac{\sum_{i=1}^{n} X_i}{n} \equiv \textbf{media muestral } (\overline{X})$$

$$T(X_1, X_2, ..., X_n) = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n-1} \equiv \textbf{varianza muestral } (S^2)$$

Se llama **estimador puntual del parámetro** θ a un estadístico que sirva para hacer inferencia sobre el parámetro.

El valor concreto que tomará el estimador al trabajar con una muestra concreta y, por lo tanto, la solución particular a nuestro problema, se denomina **estimación puntual del parámetro** y se denota:

$$\hat{\theta} = T(x_1, x_2, ..., x_n)$$

Estadísticos más usuales:

$$T(X_1, X_2, ..., X_n) = \frac{\sum_{i=1}^{n} X_i}{n} \equiv \textbf{media muestral } (\overline{X}) \text{ (será el estimador de } \mu)$$

$$T(X_1, X_2, ..., X_n) = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n-1} \equiv \textbf{varianza muestral } (S^2) \text{ (será el estimador de } \sigma^2)$$

Definición:

La distribución de probabilidad del estadístico $T(X_1, X_2, ..., X_n)$ se denomina **distribución muestral** o **distribución en el muestreo**.

4 Distribuciones Muestrales de los Estadísticos más usuales

4.1 El estadístico media muestral

Consideremos una v.a. poblacional X que sigue una distribución de probabilidad con $E(X) = \mu$ y $Var(X) = \sigma^2$.

Si queremos estimar la media poblacional μ, parece razonable escoger una muestra y calcular la media de esta muestra.

Sea $(X_1, X_2, ..., X_n)$ una m.a.s. de la v.a. $X \Rightarrow$

Las variables muestrales $X_1, X_2, ..., X_n$ son independientes, $E(X_i) = \mu$ y $Var(X_i) = \sigma^2$, $\forall i \Rightarrow$

Definimos el estadístico **MEDIA MUESTRAL** como sigue:

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

Se tiene que:

$$E(\overline{X}) = \mu$$

$$Var(\overline{X}) = \frac{\sigma^2}{n}$$

sin más que aplicar las propiedades vistas de la esperanza y la varianza para variables independientes.
CUANDO LA VARIANZA σ^2 ES CONOCIDA, su distribución muestral viene dada por los dos resultados siguientes:

Teorema de la aditividad de la distribución normal:

Sea $(X_1, X_2, ..., X_n)$ una m.a.s. de una v.a. $X \sim N(\mu, \sigma)$, entonces

$$\boxed{\overline{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}}) \implies Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)}$$

Teorema central del límite:

Sea $(X_1, X_2, ..., X_n)$ una m.a.s. de una v.a. X tal que $E(X) = \mu$ y $Var(X) = \sigma^2$, entonces la variable aleatoria

$$\overline{X} \text{ se aproxima a la distribución } N(\mu, \frac{\sigma}{\sqrt{n}}) \text{ cuando } n \longrightarrow \infty.$$

Esto es,

$$\boxed{\overline{X} \underset{n \geq 30}{\approx} N(\mu, \frac{\sigma}{\sqrt{n}}) \implies Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \underset{n \geq 30}{\approx} N(0,1)}$$

Nota: En general, para cualquier distribución, la variable media muestral se puede aproximar por la distribución $N(\mu, \frac{\sigma}{\sqrt{n}})$ cuando $n \geq 30$, entonces $Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \underset{n\geq 30}{\approx} N(0,1)$.

CUANDO LA VARIANZA σ^2 ES DESCONOCIDA, la estimamos a partir de los datos mediante la

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

A la raíz cuadrada $S = \sqrt{S^2}$ se le llama **desviación típica o estándar muestral**.

Entonces para conocer la distribución de probabilidad de la variable aleatoria $\frac{\overline{X} - \mu}{S/\sqrt{n}}$ tenemos el siguiente resultado:

Teorema de Fisher:

Sea $(X_1, X_2, ..., X_n)$ una m.a.s. de una v.a. $X \sim N(\mu, \sigma)$, entonces:

- Los estadísticos $\overline{X}$ y S^2 son independientes.
- La variable aleatoria

$T = \frac{\overline{X} - \mu}{S/\sqrt{n}}$ sigue una **distribución t de Student con n-1 grados de libertad**

Esto es, $T = \frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$

Nota: En general, para cualquier distribución, la variable media muestral se puede aproximar por la distribución de una $N(\mu, \frac{S}{\sqrt{n}})$ cuando $n \geq 30$, entonces $\frac{\overline{X} - \mu}{S/\sqrt{n}} \underset{n\geq 30}{\approx} N(0,1)$.

4.2 El estadístico proporción muestral

Supongamos que los individuos de una población determinada pueden presentar o no una cierta característica y queremos estimar la **proporción de unidades que en la población presentan dicha característica**, p.
La variable aleatoria poblacional la definimos como sigue:

$$X = \begin{cases} 1, & \text{si el individuo presenta la característica} \\ 0, & \text{en otro caso} \end{cases} \Longrightarrow X \sim b(p)$$

Sea $(X_1, X_2, ..., X_n)$ una m.a.s. de la v.a. $X \Rightarrow$

Las variables muestrales $X_1, X_2, ..., X_n$ son independientes, $E(X_i) = p$ y $Var(X_i) = p(1-p)$, $\forall i \Rightarrow$

Definimos el estadístico **PROPORCIÓN MUESTRAL** como sigue:

$$\boxed{\widehat{p} = \frac{\sum_{i=1}^{n} X_i}{n}}$$

Se tiene que:

$$E(\widehat{p}) = p$$

$$Var(\widehat{p}) = \frac{p(1-p)}{n}$$

por las propiedades de la esperanza y la varianza de variables aleatorias independientes ya citadas.

Y, aplicando el teorema central del límite, tenemos que su distribución muestral es:

$$\boxed{\widehat{p} \underset{n\geq 30}{\approx} N\left(p, \sqrt{\frac{p(1-p)}{n}}\right) \Rightarrow Z = \frac{\widehat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \underset{n\geq 30}{\approx} N(0,1)}$$

5 Introducción a los gráficos de control

Conocer las distribuciones muestrales de algunos estadísticos como la media muestral, proporción muestral y varianza muestral nos permite proporcionar procedimientos de control estadísticos de la calidad de procesos industriales.

Los **gráficos de control** son técnicas estadísticas que permiten comprobar de manera continua que se mantiene constante la calidad de una producción y nos alerta cuando ésta se altera y, de esta manera, se puede intervenir para resolver el problema.

Los gráficos de control distinguen entre la **variabilidad natural** de un proceso y la **variabilidad adicional** debido a modificaciones de los medios productivos: maquinaria, operarios, condiciones ambientes, etc..., que sí modifican el adecuado rendimiento del proceso.

Los gráficos de control combinan la descripción gráfica y numérica de los datos con la utilización de las distribuciones muestrales de los estadísticos media muestral, proporción muestral y varianza muestral.

5.1 Generalidades de los gráficos de control

La **población** de referencia es el **proceso productivo**, esto es, todos los artículos que se producirían en un proceso industrial bajo las mismas condiciones.

Todos los artículos fabricados son **muestras** de dicha población.

Se escoge una **característica** que representa la calidad de un artículo y se observa en todas las muestras.

Tipos:

- Cuando la característica es una variable cuantitativa: El proceso de vigila con el valor medio de la variable y con su variabilidad:
  - **Gráfico de control para medias.**
  - **Gráfico de control para varianzas.**
- Cuando la característica es una variable cualitativa: El proceso de vigila con la proporción de artículos que no cumplen las especificaciones y se clasifican como defectuosos:
  - **Gráfico de control para proporciones.**

El procedimiento es muy sencillo, se toman muestras de tamaño n del proceso a intervalos regulares de tiempo y se calculan las $\overline{x}$ o las $\widehat{p}$ y se dibujan en un gráfico.

Gráfico de control de la media muestral

`Ejemplo`: **Un productor de ordenadores quiere controlar la tensión que pasa por la rejilla de hilos de cobre detrás de la pantalla del ordenador. Si ésta es muy alta dañará el ensamblaje que hay detrás de la pantalla y si ésta es muy baja el usuario no podrá trabajar bien. La tensión ideal es de 275 (mV) y se sabe que, en condiciones normales de producción, los valores de la tensión en los monitores producidos se distribuyen según una distribución aproximadamente normal con desviación típica de $\sigma = 43$ (mV). Para controlar la producción se escogen cada hora 4 pantallas y se mide la tensión en sus rejillas, calculándose a continuación la media de los cuatro valores obtenidos. Los datos se presentan en la tabla siguiente:**

Muestra n^o	$\overline{x}$	Muestra n^o	$\overline{x}$
1	269.5	11	264.7
2	297.0	12	307.7
3	269.6	13	310.0
4	283.3	14	343.3
5	304.8	15	328.1
6	280.4	16	342.6
7	233.5	17	338.8
8	257.4	18	340.1
9	317.5	19	374.6
10	327.4	20	336.1

Formalicemos el contexto:

► La v.a. poblacional es

$$X = \text{Tensión de las rejillas } \sim N(\mu, \sigma = 43)$$

y cuando el **proceso está bajo control**

$$X = \text{Tensión de las rejillas } \sim N(\mu = 275, \sigma = 43)$$

► Cada hora se eligen muestras de tamaño $n = 4$ y se calcula la media muestral de la tensión de las cuatro pantallas observadas.

► Sabemos que, cuando el **proceso está bajo control**,

$$\overline{X} \sim N(\mu = 275, \frac{\sigma}{\sqrt{n}}) = N(\mu = 275, \frac{43}{\sqrt{4}})$$

De donde,

$$P(\mu - 3\frac{\sigma}{\sqrt{n}} \leq \overline{X} \leq \mu + 3\frac{\sigma}{\sqrt{n}}) = 0.997$$

Con lo cual, es poco probable obtener valores de $\overline{x}$ más allá de 3 desviaciones típicas de la media.

<u>Señal de alarma</u>:

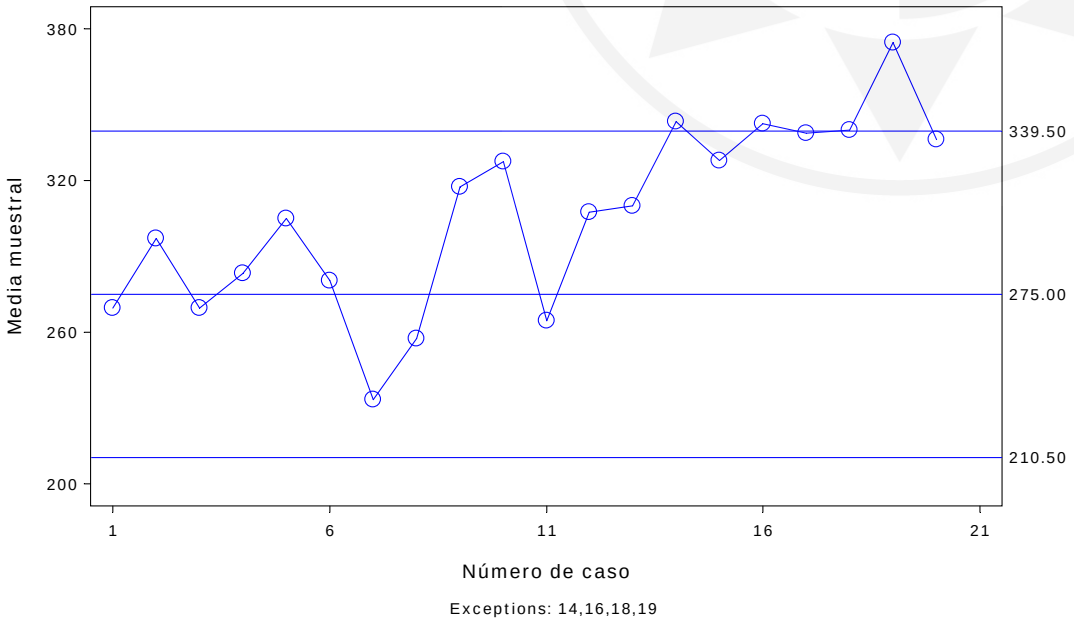
- Algún valor de $\overline{x}$ más allá de 3 desviaciones típicas de la media.

Gráfico de control para $\overline{x}$:

Calculemos primeros los límites de control:

$$\begin{array}{l}\text{Límite de control superior}=\mu+3\dfrac{\sigma}{\sqrt{n}}=275+3\times\dfrac{43}{\sqrt{4}}=339.5\\ \text{Línea central}=\mu=275\\ \text{Límite de control inferior}=\mu-3\dfrac{\sigma}{\sqrt{n}}275+3\times\dfrac{43}{\sqrt{4}}=210.5\end{array}$$

Y el gráfico quedaría:



Observamos que a partir de la muestra 11 ha aumentado considerablemente la tensión de las pantallas, luego se tendría que parar el proceso y hacer una revisión al sistema de producción para detectar el problema.

Gráfico de control de la proporción muestral

En algunas situaciones la calidad de la producción se vigila a través de la proporción de defectuosos producidos.

`Ejemplo`: **Un productor de reproductores de discos compactos usa el control estadístico de procesos para inspeccionar la calidad del circuito que contiene la mayoría de los componentes electrónicos del reproductor. La empresa se ha fijado como objetivo un 10% de defectuosos. Diariamente controla 400 circuitos y se anota la proporción de entre los 400 que fallan el test. En la tabla siguiente se presentan los datos correspondientes a los controles realizados en 16 días:**

Día n^o	$\widehat{p}$	Día n^o	$\widehat{p}$
1	0.1150	9	0.1000
2	0.1600	10	0.1600
3	0.1300	11	0.1675
4	0.1225	12	0.1225
5	0.1000	13	0.1375
6	0.1225	14	0.1975
7	0.1900	15	0.1525
8	0.1150	16	0.1675

Formalicemos el contexto:

► La v.a. poblacional es

$\widehat{p}$ = Proporción de reproductores defectuosos en una muestra de tamaño n $\approx N(p, \sigma = \sqrt{\frac{p(1-p)}{n}})$

y cuando el **proceso está bajo control**

$$\widehat{p} \approx N(0.1, \sigma = \sqrt{\frac{0.1 \times 0.9}{n}})$$

► Cada día se eligen muestras de tamaño $n = 400$ y se calcula la proporción de defectuosos en una muestra de 400 reproductores.

► Sabemos que, cuando el **proceso está bajo control**,

$$\widehat{p} \approx N(0.1, \sigma = \sqrt{\frac{0.1 \times 0.9}{n}})$$

De donde,

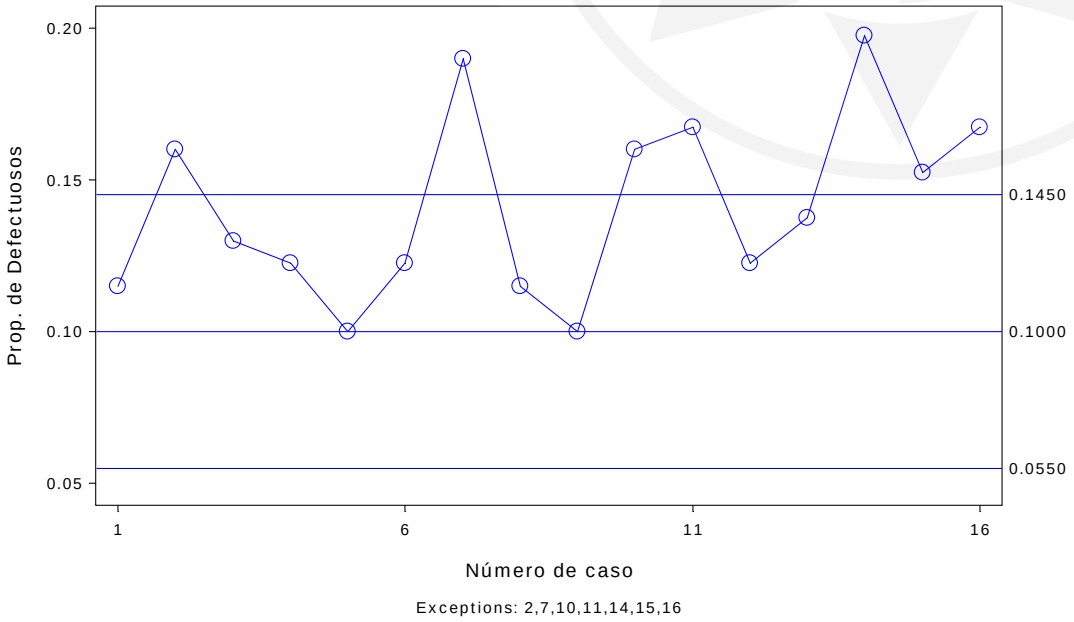
$$P(0.1 - 3\sqrt{\frac{0.1 \times 0.9}{n}} \leq \widehat{p} \leq 0.1 + 3\sqrt{\frac{0.1 \times 0.9}{n}}) = 0.997$$

Gráfico de control para $\widehat{p}$:

Calculemos primeros los límites de control:

$$\text{Límite de control superior}=p+3\sqrt{\frac{p(1-p)}{n}}=0.1+3\sqrt{\frac{0.1\times 0.9}{400}}=0.145$$
$$\text{Línea central}=p=0.1$$
$$\text{Límite de control inferior}=p-3\sqrt{\frac{p(1-p)}{n}}=0.1-3\sqrt{\frac{0.1\times 0.9}{400}}=0.055$$

Y el gráfico quedaría:



<u>Otra señal de alarma:</u>

- Nueve puntos consecutivos por encima o por debajo de la línea central u ojetivo.

Observamos que todos los valores del gráfico están por encima de la línea central, luego se tendría que parar el proceso y hacer una revisión al sistema de producción para detectar el problema.