TEMA 3:

INTRODUCCIÓN A LA TEORÍA DE MUESTRAS

ÍNDICE

- **INTRODUCCIÓN**
  - **Definiciones**
  - **Tipos de muestreo**
- **ESTADÍSTICOS**
  - **Ejemplos**
  - **Media muestral**
  - **Varianza y cuasivarianza muestral**
- **DISTRIBUCIONES DE LOS ESTADÍSTICOS**
  - **Cambio de variable**
  - **Función generatriz de momentos**
  - **Muestreo artificial (Método de Montecarlo)**
  - **Aproximación asintótica (T. C. L.)**
- **TAMAÑO DE LA MUESTRA**

INTRODUCCIÓN

Comienzo de la parte inferencial del curso

INFERENCIA: Dados unos valores de una población,

¿qué se puede decir del modelo?

- Comparar resultados con modelos reales
- Obtener conclusiones

Esquema de trabajo:

Partiendo de una población en la que se ha definido un modelo probabilístico de los estudiados hasta ahora se desea averiguar alguna característica (parámetro) desconocida.

Se extrae una muestra de tal población y con los resultados obtenidos se realizan inferencias sobre el modelo (valores del parámetro).

ENFOQUE PARAMÉTRICO

INTRODUCCIÓN

INFERENCIA ESTADÍSTICA (Temas 3, 4, 5, 6 y 7)

- Nociones básicas de muestreo
- Resultados teóricos acerca del muestreo

Tema 3

ENFOQUE PARAMÉTRICO

- Problema: Dada una población, la existencia de parámetros (valores poblacionales) desconocidos
- Soluciones:
  - Dar un único valor (Estimación puntual) **Tema 4**
  - Dar un intervalo (Estimación por intervalos) **Tema 5**
  - Creencia a priori de que el parámetro tiene algún valor o rango de valores, y confirmarlo mediante técnicas estadísticas (Test de hipótesis) **Tema 6**
  - Comparación de dos poblaciones **Tema 7**

ENFOQUE NO PARAMÉTRICO

- Otras características del modelo que es necesario confirmar (Test no paramétricos)

INTRODUCCIÓN

CONCEPTOS BÁSICOS

POBLACIÓN: Conjunto o colección de elementos sobre los cuales se quiere estudiar alguna característica.

PARÁMETRO: Característica de la población cuyo valor estamos interesados en estimar.

VARIABLE: Característica observable de un elemento de la población.

MUESTRA: Subconjunto representativo de una población.

MUESTRA ALEATORIA: Muestra extraída mediante mecanismos aleatorios.

MUESTREO: Procedimiento utilizado para la selección de los elementos de la muestra. Si la muestra es aleatoria, dicho procedimiento es probabilístico.

DATOS: Valores observados de una variable en los elementos de la muestra.

INTRODUCCIÓN

CONCEPTOS BÁSICOS

ESTADÍSTICO: Característica de la muestra

MODELO: Representación abstracta simplificada de la realidad.

MODELO PROBABILÍSTICO: Modelo matemático que describe el proceso de obtención de los datos de la muestra. Utiliza el Cálculo de Probabilidades como herramienta matemática para describirlo.

TEORÍA DE MUESTRAS: Parte del Cálculo de Probabilidades dedicada al estudio matemático de mecanismos de extracción de muestras aleatorias. Analiza, en particular, de qué tamaño se debe escoger y como dependen del azar y de los parámetros poblacionales los resultados obtenidos.

INTRODUCCIÓN

En este tema, intentaremos responder a las siguientes cuestiones:

- ¿Por qué tomar muestras?
- ¿Cómo se toman muestras?
- ¿Qué tipo de variables son las muestras?
- ¿Qué hacemos con la muestra?
- ¿Cuántos datos tomamos en la muestra?

¿POR QUÉ TOMAMOS MUESTRAS?

- Poblaciones infinitas
- Costes de la toma de muestras
- Destrucción de las unidades estudiadas

INTRODUCCIÓN

EJEMPLO INTRODUCTORIO

POBLACIÓN: Infinitas piezas que produce la máquina.

MODELO: Una variable aleatoria Bernoulli X=Pieza defectuosa X≈ Be(p)

MUESTRA: Número de piezas que se seleccionan para comprobar si son defectuosas.

PROBLEMA: No sabemos el porcentaje de piezas defectuosas, por lo tanto "p" es el parámetro desconocido. Queremos hacer inferencias acerca del verdadero valor de p

NOTA: Nunca se sabe el valor real de p

Solo es posible estimarlo

NOTA: Con la información muestral se quiere conocer toda la población

TIPOS DE MUESTREO

¿CÓMO SE TOMAN MUESTRAS?

TIPOS DE MUESTREO ALEATORIO

1) Muestreo aleatorio simple con reposición

2) Muestreo aleatorio simple sin reposición

3) Muestreo Sistemático

4) Muestreo Estratificado

5) Muestreo por conglomerados

6) Otros tipos de muestreo (polietápico, MUM,...)

TIPOS DE MUESTREO

MUESTREO ALEATORIO SIMPLE CON REPOSICIÓN

Se TOMA un elemento al azar, se ESTUDIA y se DEVUELVE antes de extraer el siguiente.

Todos los elementos tienen la misma probabilidad de ser extraídos.

Las extracciones son independientes.

EJEMPLOS

Una m.a.s. de tamaño n (X_1, ..., X_n) es una v.a. n-dimensional donde todas las X_i:

- Son independientes
- Tienen la misma distribución
- Su distribución coincide con la de población

TIPOS DE MUESTREO

MUESTREO ALEATORIO SIMPLE SIN REPOSICIÓN

Se TOMA un elemento al azar, se ESTUDIA y NO se DEVUELVE para futuras extracciones.

NO todos los elementos tienen la misma probabilidad de ser extraídos.

Las extracciones NO son independientes.

EJEMPLOS

Una m.a. sin reposición de tamaño n (X_1, ..., X_n) es una v.a. n-dimensional donde todas las X_i:

- Son dependientes
- Tienen la misma distribución
- Su distribución coincide con la de población

TIPOS DE MUESTREO

MUESTREO ALEATORIO SISTEMÁTICO

Se divide la población en k subconjuntos, donde k es el cociente entre el tamaño poblacional y el tamaño muestral: $k = N/n$

Se selecciona al azar un elemento en el primer subconjunto y se toma en el mismo orden en el resto de subconjuntos. Por ejemplo, el elemento 3 y se tomará el tercero de todos los demás subconjuntos.

En general, se toma el elemento i y el resto de la muestra será el i+k, i+2k, i+3k, …, i+(n-1)k.

EJEMPLOS

TIPOS DE MUESTREO

MUESTREO ALEATORIO SISTEMÁTICO

El muestreo sistemático presenta ventajas con respecto al muestreo aleatorio simple por su sencillez en la selección de las unidades para la muestra.

- Si el orden en la población es aleatorio, la muestra sistemática proporciona resultados comparables a una muestra aleatoria simple.
- Si los elementos de la población están ordenados de forma que los más próximos tienden a ser más semejantes que los alejados, el muestreo sistemático tiende a ser más preciso que el muestreo aleatorio simple.
- El riesgo se este tipo de muestreo está en los casos en que se dan periodicidades en la población ya que al elegir a los miembros de la muestra con una periodicidad constante (k) podemos introducir una homogeneidad que no se da en la población.

TIPOS DE MUESTREO

MUESTREO ALEATORIO ESTRATIFICADO

El muestreo estratificado consiste en dividir a la población en grupos disjuntos, llamados **estratos**, de forma que, dentro de cada grupo el comportamiento con respecto a la característica en estudio sea homogéneo (se puede estratificar, por ejemplo, según la profesión, el municipio de residencia, el sexo, el estado civil, etc.). Dentro de cada estrato se seleccionan los elementos concretos que formarán parte de la muestra mediante muestreo aleatorio simple.

En ocasiones las dificultades que plantean son demasiado grandes, pues exige un conocimiento detallado de la población (tamaño geográfico, sexos, edades,...).

EJEMPLOS

TIPOS DE MUESTREO

MUESTREO ALEATORIO ESTRATIFICADO

La distribución de la muestra en función de los diferentes estratos se denomina afijación, y puede ser de diferentes tipos:

- **Afijación Simple o uniforme**: A cada estrato le corresponde igual número de elementos muestrales.

- **Afijación Proporcional**: La distribución se hace de acuerdo con el tamaño de la población en cada estrato.

- **Afijación de mínima varianza**: se obtiene asignando a cada estrato un tamaño de la muestra proporcional al tamaño del estrato y a su variabilidad.

TIPOS DE MUESTREO

MUESTREO ALEATORIO POR CONGLOMERADOS

En el **muestreo por conglomerados** la unidad muestral es un grupo de elementos de la población que forman una unidad, a la que llamamos conglomerado. Las unidades hospitalarias, los departamentos universitarios, una caja de determinado producto, etc., son conglomerados naturales. En otras ocasiones se pueden utilizar conglomerados no naturales como, por ejemplo, las urnas electorales.

El muestreo por conglomerados consiste en **seleccionar aleatoriamente un cierto numero de conglomerados** (el necesario para alcanzar el tamaño muestral establecido) y en **investigar después todos los elementos pertenecientes a los conglomerados elegidos.**

EJEMPLOS

TIPOS DE MUESTREO

MUESTREO ALEATORIO POR CONGLOMERADOS

Aunque en el muestreo estratificado las unidades elementales también se encuentran agrupadas, el procedimiento de selección para un muestreo por conglomerados es distinto.

- En **muestreo estratificado** se selecciona una muestra dentro de cada uno de los grupos. En **muestreo por conglomerados** se selecciona una muestra de grupos y se estudian todas las unidades que lo contienen.

- En **muestreo estratificado** los grupos se definen de forma que los individuos sean homogéneos con respecto a la característica en estudio. En el **muestreo por conglomerados** se supone que los elementos dentro de los conglomerados se comportan de forma heterogénea.

TIPOS DE MUESTREO

MUESTREO DE UNIDADES MONETARIAS (MUM)

Este **tipo de muestreo es especifico en auditoría**, y viene a solucionar el problema que plantea la selección aleatoria de partidas contables que no tienen (evidentemente) el mismo importe económico.

En un muestreo estrictamente aleatorio se "primaría" la inspección de las numerosas partidas pequeñas irrelevantes dejando sin inspección las importantes y cuantiosas.

Para solucionarlo el **MUM plantea la selección aleatoria no de asientos o partidas sino de unidades monetarias** (ordenadas y numeradas) de tal manera que el defecto anterior se subsana al tener una partida cuantiosa más probabilidades de ser elegida pues contiene más unidades monetarias.

TIPOS DE MUESTREO

MUESTREO POLIETÁPICO

Cuando los elementos de la población se encuentran agrupados de forma natural en conglomerados, el **muestreo bietápico** consiste en seleccionar una muestra aleatoria de éstos y dentro de cada conglomerado, seleccionar al azar y de forma independiente, una muestra de unidades elementales.

Si dentro de los conglomerados las unidades están a su vez agrupadas en conglomerados, se selecciona una muestra aleatoria de los primeros y se aplica muestreo bietápico de forma independiente en cada uno de los conglomerados seleccionados. Es un diseño de **muestreo trietápico**.

MUESTRA

En general, sea X una variable aleatoria definida sobre una población que tiene un parámetro desconocido (θ). Su función de densidad o cuantía depende, por lo tanto, de dicho parámetro $f(x / \theta)$

Se toma una m.a.s. (X_1, ..., X_n). Su distribución vendrá dada por una función de densidad o cuantía $g(x_1, ..., x_n / \theta)$ que se calcula como:

$$g\left(x_1,\ldots,x_n \middle| \theta\right) = \prod_{i=1}^{n} f\left(X = x_i \middle| \theta\right) \quad \text{o} \quad g\left(x_1,\ldots,x_n \middle| \theta\right) = \prod_{i=1}^{n} P\left\{X = x_i \middle| \theta\right\}$$

EJEMPLO

Sea X una variable Bernoulli que indica si una persona está de acuerdo con una política del gobierno, cuyo parámetro p es desconocido. Se toma una muestra de tamaño 3: (X_1, X_2, X_3). La distribución muestral viene dada por:

$$g\left(x_1, x_2, x_3\right) = p^{\sum_{i=1}^{3} X_i} \left(1 - p\right)^{3 - \sum_{i=1}^{3} X_i}$$

ESTADÍSTICOS

ESTADÍSTICO

Cualquier función T de la muestra totalmente especificada, es decir, que no dependa de parámetros desconocidos

$$T: \quad \Re^n \rightarrow \Re$$
$$(x_1,\dots,x_n) \mapsto T(x_1,\dots,x_n)$$

- Medida de información de la muestra
- Sirve para inducir valores de parámetros desconocidos
- Es una variable aleatoria

Su distribución de probabilidad dependerá de la muestra, es decir, será una transformación de la densidad o cuantía de la muestra: $g(x_1,\dots,x_n)$.

ESTADÍSTICOS

- **Estadísticos más usuales**:

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

  Varianza muestral $$S^2 = \frac{1}{n}\sum_{i=1}^{n} (X_i - \overline{X})^2$$

  Cuasivarianza muestral $$S_1^2 = \frac{1}{n-1}\sum_{i=1}^{n} (X_i - \overline{X})^2$$

ESTADÍSTICOS

Suponemos un m.a.s. $(X_1, \ldots, X_n)$ extraída de una población X cuyas características son:

$$E[X] = \mu \qquad Var[X] = \sigma^2$$

Vamos a calcular los momentos de los estadísticos

- Media muestral $E[\bar{x}] = \mu \quad Var[\bar{x}] = \sigma^2/n$
- Varianza muestral $E[s^2] = \frac{n-1}{n}\sigma^2$
- Cuasivarianza muestral $E[s_1^2] = \sigma^2$
- CASO PARTICULAR: Proporción $X \approx Be(p)$, recordar $p=E[X]$

Proporción muestral $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} x_i$

$$E[\hat{p}] = p \qquad Var[\hat{p}] = \frac{p(1-p)}{n}$$

ESTADÍSTICOS

¿Qué tipo de variables aleatorias son los estadísticos que hemos calculado hasta ahora?

¿CÓMO CONOCER SU DISTRIBUCIÓN?

Métodos exactos

1) **Método del cambio de variable**
2) **Método de la función generatriz**

Métodos aproximados

3) **Método de Monte Carlo o Muestreo Artificial**
4) **Aproximación asintótica**

ESTADÍSTICOS

3) Muestreo artificial

Este método consiste en simular muestras aleatorias de la población de partida, construir histogramas del estadístico objeto del estudio y a partir de ahí obtener aproximadamente su distribución.

Para ello se necesita simular variables aleatorias Para cada muestra:

1) Simular n valores de una U(0,1)
2) Utilizar que $Y= F(x) \approx U(0,1)$
3) Realizar la transformación inversa y obtener la muestra aleatoria de X
4) Calcular los estadísticos que se necesiten

- Repetir estos pasos con un número elevado de muestras
- Realizar los histogramas de todas las muestras

ESTADÍSTICOS

4) Aproximaciones asintóticas

En muchos estudios se toman muestras de tamaño elevado. Es conveniente conocer las distribuciones límite de algunos estadísticos cuando el tamaño muestral tiende a infinito.

Teorema Central del Límite (T.C.L.)

Dadas X_i i = 1, ..., n variables aleatorias t. q.

- Xi son independientes
- Xi están idénticamente distribuidas
- La media y varianza son finitas $E[X] = \mu < \infty$ y $V[X] = \sigma^2 < \infty$

se verifica que: $1) \Rightarrow \dfrac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}} \xrightarrow[n\to\infty]{} N(0,1)$

$$2) \Rightarrow \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \xrightarrow[n\to\infty]{} N(0,1)$$

ESTADÍSTICOS

Es decir, la suma de variables se puede aproximar a una distribución $N(n\mu,\sigma\sqrt{n})$ y la media muestral se puede aproximar a una distribución $N(\mu,\sigma/\sqrt{n})$

- Nota: Se usa esta aproximación cuando $n > 30$
- La mayoría de estadísticos se expresan mediante una suma de variables aleatorias.

TAMAÑO DE LA MUESTRA

¿QUÉ TAMAÑO DE MUESTRA TOMAR?

Problema: Establecido el parámetro (θ) que se quiere estimar (media), el grado de precisión (error=e) necesario y el nivel de confianza deseado (1-α), ¿qué tamaño de muestra debemos tomar?

Planteamos la ecuación de probabilidad que hemos determinado con los datos anteriores:

$$P\left\{\left|T-\theta\right|<e\right\}=1-\alpha$$

Si 1-α=0,95 y e=5, la expresión anterior quiere decir que el error entre el estimador y el parámetro difiere en menos de 5 unidades con una probabilidad del 95%.

TAMAÑO DE LA MUESTRA

¿QUÉ TAMAÑO DE MUESTRA TOMAR?

La media muestral es una NORMAL de forma exacta o de forma aproximada (TCL)

Tenemos que: $\overline{X} \sim N(\mu, \sigma/\sqrt{n}) \Rightarrow P\left\{ \left| \overline{X} - \mu \right| < z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right\} = 1 - \alpha$

- Fijado el nivel de confianza, mediante las tablas de la distribución normal se calcula $z_{\alpha/2}$.
- Se obtiene el tamaño mediante $\boxed{n = \dfrac{z_{\alpha/2}^2 \sigma^2}{e^2}}$

■ **Nota 1:** El tamaño muestral (n) es directamente proporcional a la varianza de la población y al nivel de confianza, e inversamente proporcional al error que estemos dispuestos a cometer.

TAMAÑO DE LA MUESTRA

- Nota 2: Caso particular, proporciones X≈Be(p). Aplicando el TCL se obtiene

$$\hat{p} \approx N\left(p, \sqrt{p(1-p)/n}\right)$$

El tamaño muestral vendrá dado por:

$$\boxed{n = \frac{z_{\alpha/2}^2 p(1-p)}{e^2}}$$

- Nota 3: En caso de desconocer la varianza, ésta se sustituye por una estimación, habitualmente la cuasivarianza muestral (s_1^2) o por el peor valor posible. Por ejemplo, en el caso de la distribución Bernoulli la peor (máxima) varianza se alcanza cuando p = 1/2, es decir, la mayor incertidumbre posible es jugar al 50% entre éxito y fracaso.

M.A.S. SIN REPOSICIÓN

Es un muestreo sin reposición donde E[X] = μ y Var[X]=σ^2. Las variables X_i están idénticamente distribuidas pero son dependientes, por lo tanto, existe relación entre los datos muestrales.

- **Media muestral** $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
  - **Esperanza** $E\left[\overline{X}\right] = \mu$
  - **Varianza** $Var\left[\overline{X}\right] \approx \frac{\sigma^2}{n}(1 - f)$

donde f es la fracción de muestreo (f = n/N)

TAMAÑO DE LA MUESTRA

- **<u>Tamaño muestral:</u>** Para el caso de muestreo aleatorio sin reposición que se utiliza principalmente cuando muestreamos poblaciones finitas, se calcula lo que se denomina la "corrección para poblaciones finitas"

$$\boxed{n = \frac{n_{\infty}}{1 + \frac{n_{\infty}}{N}}}$$

donde n_{∞} es el tamaño calculado para el caso de muestreo aleatorio con reposición (utilizado para poblaciones infinitas).