TEMA 57: USOS DE LA ESTADÍSTICA: ESTADÍSTICA DESCRIPTIVA Y ESTADÍSTICA INFERENCIAL. MÉTODOS BÁSICOS Y APLICACIONES DE CADA UNA DE ELLAS

TIEMPO: 74 — 74

Esquema

1) Introducción
   - 1.1) Definición
   - 1.2) Historia
     - 1.2.1) Prehistoria
     - 1.2.2) Antigüedad y Edad Media
     - 1.2.3) Edad Moderna y Contemporánea

2 Aplicaciones
   - 2.1) Medicina
   - 2.2) Economía
   - 2.3) Física de fluidos

3) Conceptos generales
   - 3.1) Definiciones básicas
     - 3.1.1) Población, individuo, muestra, característica
   - 3.2) Variables estadísticas
   - 3.2.1) Continua, discreta
   - 3.3) Frecuencias
   - 3.3.1) Absoluta, relativa
   - 3.4) Gráficos
     - 3.4.1) Histogramas, circulares, nubes de puntos

4) Parámetros asociados
   - 4.1) Definición
   - 4.2) Centralización: moda, media, mediana
   - 4.3) Dispersión: rangos, desviación típica, varianza, coeficiente de variación
   - 4.4) Otros

5) Encuestas y sondeos
   - 5.1) Sondeo: definición
   - 5.2) Encuesta: definición + a tener en cuenta
   - 5.3) Cuestionario: definición + a tener en cuenta

6) Método estadístico

6.1) Introducción

6.2) Pasos

6.2.1) Planteamiento

6.2.2) Construcción

6.2.3) Recogida de datos y presentación

6.2.4) Depuración y estimación

6.2.5) Refinamiento

6.2.6) Crítica y valoración

7) Estadística descriptiva e inferencial

7.1) Estadística descriptiva

7.1.1) Nacimiento

7.1.2) Evolución

7.2) Estadística inferencial

7.2.1) Definición

7.2.2) Muestra

7.2.3) Estimación

1) Introducción:

▷ Hoy en día podemos definir la Estadística como la ciencia que usa los conjuntos de datos para obtener, a partir de ellos, inferencias basadas en el cálculo de probabilidades.

Además, la Estadística ha ido evolucionando desde formas muy concretas (mera recopilación de datos) a otras más abstractas y rigurosas propias de las ciencias del s.XXI (test de hipótesis, teoremas en problemas de Big Data, estadística cuántica,...). Hagamos un repaso rápido por la Historia de la Estadística.

▷ Prehistoria: encontramos signos del conteo de animales o personas desde el 30.000 a.C. en forma de muescas y representaciones gráficas sobre pieles, huesos, rocas,...

▷ Antigüedad no europea: los babilonios recopilaban datos en tablas de arcilla para hacer recuentos de producciones agrícolas. Los egipcios llevaban registros sobre los niveles de crecida del Nilo y las superficies cultivables. Tanto en Israel como en China se hacían censos sobre poblaciones y sus territorios.

▷ Antigüedad europea: los griegos manejaban censos con el fin de llevar control sobre impuestos o, en Atenas por ejemplo, la población con capacidad de voto. Serían los romanos los primeros e realizar grandes estadísticas poblacionales, de superficie de territorios o de las rentas obtenidas en los mismos.

▷ Edad Media: Carlomagno inventarió en el s.VIII las propiedades de la Iglesia. Guillermo el Conquistador en el s.XI encarga censos e inventarios sobre las posesiones normandas en las Islas Británicas (en el conocido "Doomsday Book"). En el s.XV se hacen recuentos de los fuegos (hogares) de Castilla por orden de los Reyes Católicos.

▷ s.XVII: empiezan a formalizarse los trabajos en cuanto a metodología y medios empleados. Destacamos dos escuelas:

a) Escuela descriptiva alemana: fundada por Hermann Comring.

b) Escuela de los aritméticos políticos: fundada por John Grant y que estableció las primeras correlaciones entre variables.

▷ s.XVIII: el gran empujón lo dieron las compañías de seguros y su interés por establecer las cuotas anuales en base a los datos disponibles. En este período la Estadística seguía siendo la exposición de las características más notables de un Estado.

▷ s.XIX: es en este siglo cuando, debido al incremento del volumen de datos, se le empiezan a incorporar métodos numéricos y una teoría que la convertirá en ciencia propia dentro de las Matemáticas.

a) Los orígenes teóricos son debido a los trabajos de Gauss y Laplace sobre la teoría de errores en las mediciones.

b) Destacamos las contribuciones de Galton y Pearson.

▷ s.XX: la Estadística se convierte en una teoría puramente matemática usando el Cálculo de Probabilidades y todos los trabajos sobre inferencia, test de hipótesis, estimaciones de parámetros, intervalos de confianza,... que le dan la forma de ciencia abstracta en cuanto a la teoría pero tremendamente práctica en cuanto a las aplicaciones que tiene hoy en día.

2) Aplicaciones:

▷ La Estadística tiene una multitud de aplicaciones en la actualidad más allá de aquellas con las que nació (censos, demografía, seguros,...). Algunas de ellas son:

▷ **Medicina**: el estudio de nuevos fármacos y sus posibles efectos beneficiosos/perjudiciales requieren de toda una serie de test de hipótesis, intervalos de confianza, estimación de parámetros y el manejo de los "errores Tipo I" y "errores Tipo II".

▷ **Economía**:algunos instrumentos de medida del riesgo usan la Estadística. Uno de los más famosos es el VaR ("Value at Risk") que, mediante históricos de datos, calcula la probabilidad, dado un margen de confianza, de sufrir unas ciertas pérdidas.

▷ **Física de fluidos**: en medios donde se pueden generar turbulencias se utilizan métodos estadísticos para generar las simulaciones numéricas de cómo evolucionará en el tiempo dicho sistema.

3) Conceptos generales:

▷ Recordemos algunos conceptos generales de la Estadística que intervendrán en el desarrollo del tema:

1) Población: es el conjunto sobre el que se realiza el estudio estadístico.

2) Individuo: cada una de las unidades elementales sobre las que se realiza el estudio.

3) Muestra: es un subconjunto representativo de la población.

4) Tamaño: es el número de elementos de la población o la muestra.

5) Características: es el aspecto, rasgo o cualidad que se estudia en cada individuo de la población. Las hay de dos tipos dependiendo de los valores que puedan tomar:

   5.1) Cuantitativa: sus valores son numéricos (peso, edad, altura,...).

   5.2) Cualitativa: sus valores no son numéricos (color de ojos, sexo, nombre,...).

▷ A las características cuantitativas se les pueden asociar variables estadísticas que pueden ser:

a) Discretas: cuando sus valores son un conjunto numerable (número de hijos, años de edad,...).

b) Continuas: cuando sus valores son un intervalo (altura de las personas, tiempo de vida de un coche,...).

▷ En las variables discretas suele tener importancia el uso de las "frecuencias":

a) Frecuencia absoluta: es el número de veces que se repite cada valor de la variable.

b) Frecuencia relativa: es el cociente entre la frecuencia absoluta y el total de datos.

▷ Los gráficos nos permiten ver, rápidamente, las relaciones entre las variables y sus frecuencias. Los gráficos más empleados son:

1) Histogramas: relacionan cada valor de la variable con su frecuencia absoluta/relativa.

2) Diagramas circulares: cada sector es proporcional a la frecuencia relativa de ese valor de la variable.

3) Nubes de puntos: no están relacionados con las frecuencias pero son muy útiles cuando el número de datos es muy elevado y se quieren deducir patrones generales como correlaciones, agrupamientos, fenómenos especiales,...

4) Parámetros estadísticos:

▷ Dada una distribución estadística existen unas medidas inherentes a ella que sirven para exponer sus aspectos más destacados. A estas medidas las llamamos parámetros estadísticos.

Los parámetros estadísticos sirven también para comparar las distribuciones y extraer conclusiones en relación a las variables empleadas. Como se ven con detalle en temas posteriores, aquí haremos una introducción a los dos grandes grupos: de centralización y de dispersión.

▷ **Parámetros de centralización**: son los que miden los valores centrales representativos de la distribución. Los más importantes son la moda, la media y la mediana.

▷ **Moda**: es el valor de la variable que más se repite. Es el único parámetro que vale tanto para cuantitativas como cualitativas. Una distribución puede tener más de una moda.
Por ejemplo:

Nota	Nº alumnos
5	2
6	5
7	5
8	3

En este caso las modas son 6 y 7.

▷ **Media o esperanza ($\overline{x}$, μ, $E[\mathbb{X}]$)**: es el promedio del conjunto de valores de la variable. La media es el centro de gravedad de la distribución. Cuando hay valores extremos puede no ser representativa o no existir.

$$\overline{x} = \mu = E[\mathbb{X}] = \sum_{j=1}^{N} x_j \cdot f_j = \frac{\sum_{j=1}^{N} x_j \cdot n_j}{N}$$

La media cumple una propiedad clave que es ser lineal: $E[a \cdot \mathbb{X} + b] = a \cdot E[\mathbb{X}] + b$, $\forall a, b \in \mathbb{R}$
Por ejemplo: en la tabla anterior: $E[\mathbb{X}] = \frac{5 \cdot 2 + 5 \cdot 5 + 7 \cdot 5 + 8 \cdot 3}{15} = 6,6$

▷ **Mediana**: es el valor central de los datos una vez ordenados de menor a mayor. Si el número de datos es impar, es aquel que ocupa el valor central. Si el número de datos es par, será la media de los valores centrales.
Por ejemplo:

Nota	Nº alumnos
5	2
6	2
8	1

La mediana es: ~~5~~ - ~~5~~ - 6 - ~~6~~ - ~~8~~ $\Longrightarrow$ la mediana es 6

▷ **Parámetros de dispersión**: son los que miden el grado de separación/dispersión de los valores de la variable respecto a una medida de centralización. Destacamos los rangos, la desviación típica. el coeficiente de variación, los de posición y los de forma.

▷ **Rangos**: miden la diferencia entre dos valores interesantes de la variable. Los más comunes son el Rango ($x_{\text{máx}} - x_{\text{mín}}$) y el Rango Intercuartílico ($x_{Q_3} - x_{Q_1}$).

▷ **Varianza (σ^2)**: es la media de los cuadrados de las desviaciones respecto a la media.

$$0 \leq \sigma^2 = \sum_{j=1}^{N} \frac{(x_j - \overline{x})^2 \cdot n_j}{N} = E[\mathbb{X}^2] - E[\mathbb{X}]^2$$

Dem. Usemos la definición de Esperanza matemática y sus propiedades.
$E[(\mathbb{X} - E[\mathbb{X}])^2] = E[\mathbb{X}^2 - 2\mathbb{X}E[\mathbb{X}] + E[\mathbb{X}]^2] = E[\mathbb{X}^2] - 2E[\mathbb{X}]^2 + E[\mathbb{X}]^2 = E[\mathbb{X}^2] - E[\mathbb{X}]^2$ □

▷ **Desviación típica (σ)**: como $\sigma^2 \geq 0$, tomamos $\sigma = +\sqrt{\sigma^2}$ y este parámetro es el que mejor mide si los datos están o no agrupados respecto a la media.
Propiedad: $\forall a$, b $\in \mathbb{R}$, $\sigma[a \cdot \mathbb{X} + b] = a \cdot \sigma[\mathbb{X}]$

▷ **Coeficiente de variación**: se utiliza para comparar la dispersión de la variable, sin unidades, y se define por: C.V. $= \dfrac{\sigma}{\overline{x}}$ (también llamado coeficiente de variación de Pearson). Cuanto más pequeño sea, más representativa es la media.

▷ **Parámetros de posición**: son los que, una vez ordenados los datos, nos indican cuántos elementos quedan a la izquierda/derecha de uno dado. Son los deciles, cuartiles y percentiles.

▷ **Parámetros de forma**: informan sobre el apuntamiento o simetría de la distribución.

5) Encuestas y sondeos:

▷ Antes de analizar los datos de un estudio hay que obtenerlos. Estos datos pueden venir de estadísticas oficiales, análisis de laboratorios,... o de un instrumento muy habitual como son las encuestas (y cuyo propósito es poner de manifiesto una determinada situación de una población). Veamos las tres partes involucradas.

▷ Sondeo: son un método de investigación para obtener información de un grupo de individuos previamente seleccionados.

▷ Encuesta: es una técnica para recoger información para su posterior estudio. Puede realizarse tanto sobre el total de la población como de una muestra.
Para su realización y elaboración habrá que tener en cuenta la información que se quiere obtener, la población/muestra objeto del estudio, el método a seguir, si su elaboración puede dar validez a las respuestas, la formulación del cuestionario, el trabajo de campo o el procesamiento y tabulación de los datos obtenidos.

▷ Cuestionario: será el conjunto de preguntas donde se recoge toda la información. Ha de favorecer la recogida de información y debe facilitar el estudio de los resultados.
Para su realización y elaboración se dividirá en secciones compuestas de preguntas que han de cumplir premisas como el ser concisas y concretas, la respuesta en una no debe influir en otras, han de ser fácilmente codificables, etc.

▷ Otro tema a parte son las técnicas a aplicar para facilitar el manejo de datos, la eliminación de los no significativos, el cálculo de las medidas a analizar, su relación con otros datos o los tipos de informes anexos (técnicos, de usuario,...).

6) Método estadístico:

▷ Una cuestión importante en la realización de experimentos es su fiabilidad. Hoy en día la técnica está suficientemente avanzada y apoyada en resultados potentes como la Ley de los Grandes Números o el Teorema Central del Límite como para dar validez más que sobrada a los experimentos estadísticos. Los pasos del método estadístico son:

▷ 1) Planteamiento del problema: se concretan la población, las variables y los métodos para medirlas.

▷ 2) Construcción del modelo: elegimos el modelo matemático que mejor se ajuste a la realidad objeto del estudio.

▷ 3) Recogida y presentación de datos: distinguiendo entre primarios (aquellos reunidos y registrados por el investigador en el momento de la investigación) y los secundarios (los recogidos en otro momento, informes publicados en revistas y, en general, los que se encuentran en bibliotecas o editoriales).

▷ 4) Depuración de datos: se trata de eliminar los posibles errores cometidos en las medidas.

▷ 5) Estimación de parámetros: a partir de los datos obtenidos, estimamos los parámetros poblacionales en los que estamos interesados.

▷ 6) Refinamiento del modelo: sin renunciar a la calidad requerida, eliminamos los parámetros superfluos.

▷ 7) Crítica y validación del modelo: se contrasta la compatibilidad entre la información empírica y la aportada por el modelo. Si hay incongruencias obvias, se replantearía todo desde el segundo paso.

▷ Según las etapas anteriores podemos dividir la Estadística en Estadística descriptiva (obtenemos los parámetros a partir de los datos del estudio) y Estadística inferencial (a partir de los parámetros de la muestra obtenemos los de la población global).

7) Estadística descriptiva e inferencial:

▷ A continuación bosquejaremos estas dos ramas pues se tratarán con más detalle en otros temas:

▷ Estadística descriptiva: esta rama trata del recuento, ordenación y clasificación de los datos y del cálculo de medias, desviaciones e índices a partir e los mismos. La mayor parte de sus deducciones son inmediatas a partir del material estadístico recopilado. Su fin es describir, no explicar. Es muy útil en áreas como la demografía, la economía o la biología.
La Estadística nació como descriptiva. Se necesitaría del descubrimiento de analogías y de ciertas distribuciones tipo para dar el salto al Cálculo de Probabilidades y desarrollar la rama de la Estadística inferencial.

▷ Estadística inferencial: es el proceso mediante el cual, a partir de los datos estadísticos de una muestra se estiman (infieren) los datos de la población total.
Un concepto clave es de "muestra", que es un subconjunto de la población que ha de ser representativo del total, tener un tamaño suficiente y con el menos error para que las conclusiones sean fiables (una muestra puede ser representativa para ciertos parámetros y no serlo para otros). En caso de que no sera representativa, decimos que está sesgada.
Otro concepto clave es del "estimación" (ya sea puntual, por intervalos,...) que es la técnica mediante la cual se plantean previsiones y conclusiones sobre una población en base a las leyes de la probabilidad y, después, se comprueban las hipótesis mediante los contrastes estadísticos adecuados.