ESTADÍSTICA

DOS RAMAS DE LA ESTADÍSTICA

INTRODUCCIÓN

La estadística tiene por objeto el desarrollo de técnicas para el conocimiento numérico de un conjunto de datos empíricos (recogidos mediante experimentos o encuestas).

Caracteres y variables: Caracteres son los aspectos que deseamos estudiar. Cada carácter puede tomar distintos valores o modalidades. Una variable estadística recorre todos los valores de un cierto carácter.

Clasificación de las variables estadísticas:
- *Cualitativas:* No toman valores numéricos
- *Cuantitativas discretas:* Toman valores numéricos aislados

Cuantitativas continuas: Pueden tomar todos los valores de un intervalo.

Población: Es el conjunto de todos los elementos cuyo conocimiento nos interesa y serán objeto de nuestro estudio.

Muestra: Es un subconjunto, extraído de la población, cuyo estudio sirve para inferir características de toda la población.

Individuo: Es cada uno de los elementos que forman la población o la muestra.

DOS RAMAS DE LA ESTADÍSTICA

- **La estadística descriptiva:** Trata de "describir" y analizar algunos caracteres de los individuos de un grupo dado, sin extraer conclusiones para un grupo mayor. Para este estudio, se siguen estos pasos:
  - Selección de caracteres que interese estudiar.
  - Análisis de cada carácter: diseño de la encuesta o del experimento y recogida de datos.
  - Clasificación y organización de los resultados en tablas de frecuencias.
  - Elaboración de gráficos, si conviene, para divulgarlos a un público amplio (no experto).
  - Obtención de parámetros: valores numéricos que resumen la información obtenida.

- **La estadística inferencial:** Trabaja con muestras y pretende, a partir de ellas, "inferir" características de toda la población. Es decir, se pretende tomar como generales propiedades que solo se han verificado para casos particulares. En ese proceso hay que operar con mucha cautela: ¿Cómo se elige la muestra?, ¿Qué grado de confianza se puede tener en el resultado obtenido?

ESTADÍSTICA DESCRIPTIVA

TABLAS DE FRECUENCIAS

DEFINICIÓN

Las **tablas de frecuencias** sirven para ordenar y organizar los datos estadísticos. Con ellas, una masa amorfa de datos pasa a ser una colección ordenada y perfectamente inteligible.

Con los datos se construye la tabla de frecuencias:

- En la primera columna, la variable x_i, con todos sus posibles valores
- En la segunda columna, la correspondiente frecuencia, f_i: número de veces que aparece cada valor.

x_i	f_i

FRECUENCIAS RELATIVAS

Cuando se desea comparar varias distribuciones similares con distinto número de elementos, se debe recurrir a las **frecuencias relativas**. Estas vienen dadas en "tanto por uno" (f_r) o en "tantos por ciento" (%). Si N es el número de individuos: $fr_i = \frac{f_i}{N}$ $\% = 100.fr_i = \frac{100.f_i}{N}$

FRECUENCIAS ACUMULADAS

En una distribución de frecuencias, se llama **frecuencia acumulada**, F_i, correspondiente al valor i-ésimo, x_i, a la suma de la frecuencia de ese valor con todas las anteriores: $F_i = f_1 + f_2 + ... + f_i$ $Fr_i = F_i / N$ % acum. $= Fr_i \times 100$

TABLAS CON DATOS AGRUPADOS

Cuando en una distribución estadística el número de valores que toma la variable es muy grande, conviene elaborar una tabla de frecuencias agrupándolos en intervalos. Para ello:

- Se localizan los valores extremos, a y b, y se halla su diferencia, r = b–a
- Se decide el número de intervalos que se quiere formar, teniendo en cuenta la cantidad de datos que se poseen. El número de intervalos no debe ser inferior a 6 ni superior a 15.
- Se toma un intervalo, r', de longitud algo mayor que el recorrido r y que sea múltiplo del número de intervalos, con objeto de que estos tengan una longitud entera.
- Se forman los intervalos de modo que el extremo inferior del primero sea algo menor que a y el extremo superior del último sea algo superior a b.

El punto medio de cada intervalo se llama **marca de clase**. Es el valor que representa a todo el intervalo para el cálculo de algunos parámetros.

Cuando se elabora una tabla con datos agrupados, se pierde algo de información (pues en ella se ignora cada valor concreto, que se difumina dentro de un intervalo). A cambio, se gana en claridad y eficacia.

GRÁFICOS ESTADÍSTICOS

GRAFICOS PARA VARIABLES CUALITATIVAS

Diagrama de barras:

- En el eje de las X : Se representan los valores de la variable
- En el eje de las Y : Se representan los valores de la frecuencia: f, f_r ó %
- Se levanta para cada valor de la X una barra que representa la frecuencia de dicho valor.

Si unimos mediante una poligonal los puntos más altos de cada barra obtenemos **el polígono de frecuencias**.

Diagrama de sectores: Se dibuja un círculo y los porcentajes correspondientes a cada valor (Para dibujar los sectores conviene hacerlo a partir del % acumulado, pues facilita el trabajo)

GRAFICOS PARA VARIABLES CUANTITATIVAS DISCRETAS

Diagrama de barras y polígono de frecuencias: Como en las cualitativas

Diagrama de barras acumuladas: Como en los diagramas de barras pero en el eje OY se toman los valores de las frecuencias acumuladas (F_i, Fr_i o % acum).
Si unimos mediante una poligonal los puntos más altos de cada barra obtenemos el **polígono de frecuencias acumuladas.**

Diagrama de sectores: Como en las cualitativas

GRAFICOS PARA VARIABLES CUANTITATIVAS CONTINUAS

SI TODOS LOS INTERVALOS TIENEN LA MISMA AMPLITUD

Histograma :

- En el eje de las X : Se representan los valores de la variable
- En el eje de las Y : Se representan los valores de la frecuencia: f, f_r ó %
- Se levanta para cada valor del intervalo de la X un rectángulo de altura la frecuencia de dicho intervalo.

Si unimos mediante una poligonal los puntos medios más altos de cada uno de dichos rectángulos obtenemos **el polígono de frecuencias**.

Diagrama de barras acumuladas:

- En el eje de las X : Se representan los valores de la variable
- En el eje de las Y : Se representan los valores de la frecuencia acumulada: F, F_r ó %a
- Se levanta para cada valor del intervalo de la X un rectángulo de altura la frecuencia acumulada de dicho valor.

Si unimos mediante una poligonal las diagonales de dichos rectángulos obtenemos **el polígono de frecuencias acumuladas**.

Diagrama de sectores: Como en las cualitativas

SI LOS INTERVALOS NO SON TODOS DE LA MISMA AMPLITUD

En los histogramas, en el eje de las Y, en vez de representar la frecuencia se representa la densidad de frecuencia : $d_i = f_i/a_i$ siendo a_i la amplitud de dicho intervalo, para que así la frecuencia coincida con el área del rectángulo.

Los histogramas acumulados y los diagramas de sectores iguales.

PARÁMETROS DE CENTRALIZACIÓN Y DISPERSIÓN

Las definiciones siguientes sirven tanto para datos aislados como para datos agrupados en intervalos:

- Si los datos son aislados: los x_i son los valores que toma la variable
- Si los datos están agrupados en intervalos: los x_i son las marcas de clase.

MODA

Mo = El x_i que tenga mayor frecuencia

MEDIA

$$\bar{x} = \frac{\sum f_i.x_i}{\sum f_i} = \frac{\sum f_i.x_i}{N}$$

VARIANZA

$$Var = \frac{\sum f_i.(x_i - \bar{x})^2}{N} = \frac{\sum f_i.x_i^2}{N} - \bar{x}^2$$

DESVIACIÓN TÍPICA

$\sigma = \sqrt{varianza}$

COEFICIENTE DE VARIACIÓN

$C.V. = \frac{\sigma}{\bar{x}}$ Sirve para comparar las dispersiones de poblaciones heterogéneas, pues indica la variación relativa.

MEDIDAS DE POSICIÓN PARA DATOS AISLADOS

MEDIANA

Si los individuos de una población están colocados en orden creciente según la variable que estudiamos, el que ocupa el valor central se llama individuo mediano, y su valor, la mediana: Me

La **mediana**, Me, está situada de modo que antes de ella está el 50% de la población y, detrás, el otro 50%

Si el número de individuos es par, la mediana es el valor medio de los dos centrales.

CUARTILES

Si un lugar de partir la totalidad de los individuos en dos mitades, lo hacemos en cuatro partes iguales (todas ellas con el mismo número de individuos), los dos nuevos puntos de partición se llaman **cuartiles**.

Cuartil inferir: Q_1, es un valor de la variable que deja por debajo de él al 25 % de la población, y por encima, al 75%

Cuartil superior: Q_3, deja debajo el 75% y encima el 25%

En realidad existiría uno cuartil, Q_2, que coincide con la mediana.

También se suelen llamar, primer cuartil, segundo cuartil = mediana y tercer cuartil.

CENTILES O PERCENTILES

Si partimos la población en 100 partes y señalamos el lugar que deja debajo k de ellas, el valor de la variable correspondiente a es lugar se designa por p_k y se denomina centil k o percentil k.

La mediana es Me = p_{50} y los cuartiles $Q_1 = p_{25}$, Q_{75}

Si en vez de dividir en 100 partes dividimos sólo en 10, obtenemos los deciles:
$D_2 = p_{20}$

OBTENCIÓN PERCENTILES EN TABLAS DE FRECUENCIAS

Para hallar el percentil p_k en una tabla de frecuencias, se obtienen los porcentajes acumulados. El percentil p_k es el valor para el cual la frecuencia acumulada correspondiente supera el k%.

En el caso de que una de ellas coincida con k%, se toma como p_k el valor intermedio entre ese valor de x y el siguiente.

MEDIDAS DE POSICIÓN PARA DATOS AGRUPADOS EN INTERVALOS

INTRODUCCIÓN

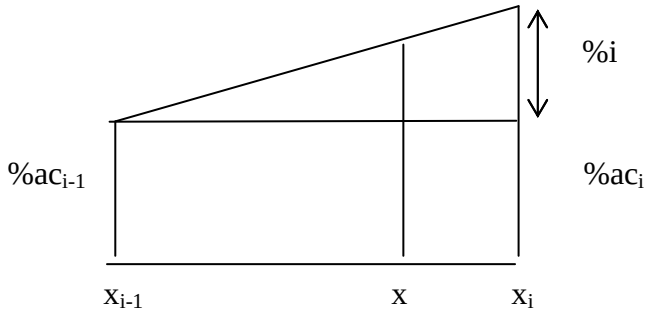
En las tablas de frecuencias con datos agrupados en intervalos se ha perdido el valor concreto de cada individuo. ¿Cómo saber, pues, dónde está la mediana o el percentil 20?

Teniendo en cuenta el convenio: En una tabla de frecuencias con datos agrupados en intervalos, suponemos que los datos de cada intervalo se reparten uniformemente en él.

Según esto, los valores de las frecuencias acumuladas deben asignarse a los extremos superiores de los intervalos, pues es al final de cada intervalo cuando se han contabilizado todos los individuos.

CÁLCULO

Procedemos como si los datos no estuviesen agrupados para hallar el intervalo correspondiente, y teniendo en cuenta el **polígono de porcentajes acumuladas** en dicho intervalo:

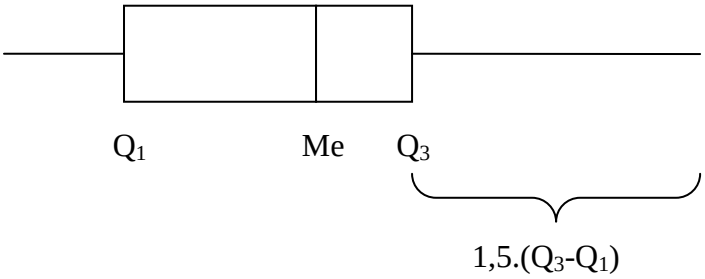


Aplicando semejanza de triángulos, obtendremos “x”.

DIAGRAMAS DE CAJA

Se construyen del siguiente modo:

- La caja abarca el intervalo Q_1, Q_3 (llamado recorrido intercuartílico) y en ella se señala expresamente el valor de la Mediana, Me.
- Los bigotes se trazan hasta abarcar la totalidad de los individuos, con la condición de que cada lado no se alargue más de una vez y media la longitud de la caja.
- Si uno (o más) de los individuos quedara por debajo o por encima de esta longitud, el correspondiente bigote se dibujará con esa limitación y se añadiría, mediante asterisco, el individuo en el lugar que le corresponde.



ESTADÍSTICA INFERENCIAL

¿POR QUÉ SE RECURRE A LAS MUESTRAS?

En la práctica, es muy frecuente tener que recurrir a una muestra para inferir datos de la población por alguno o varios de los siguientes motivos:
1 – La población es excesivamente numerosa.
2 – La población es muy difícil, o imposible, de controlar.
3 – El proceso de medición es destructivo o demasiado caro.
4 – Se desea conocer rápidamente ciertos datos de la población y se tardaría demasiado en consultar a todos.

TAMAÑO DE LA MUESTRA

Respecto del tamaño, es claro que si la muestra es demasiado pequeña, no podremos extraer de ella ninguna conclusión que valga la pena. Sin embargo, con muestras aparentemente muy pequeñas se consiguen estimaciones sorprendentemente buenas en la realidad.

LA MUESTRA HA DE ELEGIRSE AL AZAR

Al sustituir el estudio de la población por el de la muestra, se cometen errores. Pero con ellos contamos de antemano y pueden controlarse.

Sin embargo, si la muestra está mal elegida (**no es representativa**), se producen errores adicionales imprevistos e incontrolados (**sesgos**).

El proceso mediante el cual se confecciona la muestra se llama **muestreo**. ¿Cómo debe ser el muestreo para que nos proporcione una muestra representativa, no sesgada? La muestra ha de ser elegida al azar, es decir, el muestreo ha de ser aleatorio.

Se dice que un **muestreo es aleatorio** cuando los individuos de la muestra se eligen al azar, de modo que todos los individuos de la población tienen la misma probabilidad de ser elegidos.

El muestreo aleatorio es el único que garantiza la fiabilidad de las conclusiones que se obtengan.

CONCLUSIONES QUE SE OBTIENEN DE UNA MUESTRA

Las valoraciones numéricas se dan mediante intervalos, acompañados de una probabilidad (**nivel de confianza**).

Cuanto más amplio es el intervalo, mayor es el nivel de confianza que tendremos. Y, al contrario, si se quiere afinar mucho en las previsiones reduciendo el intervalo, perderemos confianza en los resultados.

El tamaño de la muestra también influye. Aumentándolo podremos:
- Mejorar el nivel de confianza manteniendo la amplitud del intervalo.
- Reducir la amplitud del intervalo manteniendo el nivel de confianza.