MATEMATICAS (MAT ICCSS)

1º Bachillerato

ESTADISTICA DESCRIPTIVA

Departamento de Matemáticas

Ies Dionisio Aguado

1 ESTADÍSTICA DESCRIPTIVA

1.1 Definiciones

La estadística descriptiva es la técnica matemática que obtiene, organiza, presenta y describe un conjunto de datos con el propósito de facilitar el uso, generalmente con el apoyo de tablas, medidas numéricas o gráficas. Además, calcula parámetros estadísticos para describir los datos estudiados.La estadística tiene dos grandes ramas: Descriptiva e Inferencial.

- *Estadística Descriptiva* analiza las características de una población o muestra definiéndose unas propiedades acerca de su estructura y composición.
- *Estadística Inferencial* basándose en los resultados obtenidos de una muestra induce o estima las leyes reales de comportamiento de la población de la que proviene dicha muestra.

Elementos que intervienen en el estudio estadístico:

- *Población*
  - Son todos y cada uno de los elementos que se quieren analizar. Puede ser finita o infinita(en realidad las poblaciones infinitas no existen, pero cuando se trata de un número grande se trata como si lo fuera).
- Muestra
  - Es el conjunto menor de individuos accesible y limitado de la población sobre el que realizamos las mediciones o el experimento con la idea de obtener conclusiones generalizables a la población.
  - El individuo es cada uno de los componentes de la población y la muestra.
  - Al número de individuos que forman la muestra se llama tamaño muestral (n).
  - La muestra debe ser representativa de la población y con ello queremos decir que cualquier individuo de la población en estudio debe haber tenido la misma probabilidad de ser elegido.
- Muestreo: Es el proceso de selección de los individuos de la muestra. Este muestreo puede realizarse con diferentes técnicas:
  - Muestreo aleatorio simple Cada individuo tiene las mismas posibilidades de ser elegido para formar parte de la muestra.
  - Muestreo aleatorio estratificado Aseguras que la muestra tenga la misma proporción de una(s) variables que la población de la que procede.
  - Muestreo sistemático El proceso de selección se basa en alguna regla sistemática simple, por ejemplo, elegir uno de cada "n" individuos.
- *Característica*
  - Propiedad que se estudia de la población

1.2 Variables Estadísticas

1.2.1 *Variables* cuantitativas

- Son las variables que pueden medirse, cuantificarse o expresarse numéricamente. Las variables cuantitativas pueden ser de dos tipos:
  - Variables cuantitativas continuas, si admiten tomar cualquier valor dentro de un rango numérico determinado (peso, talla).
    - ➢ *Variable Discreta* si no admiten todos los valores intermedios en un rango. Suelen tomar solamente valores enteros (número de hijos, número de partos, número de hermanos, etc).

Ejemplo: Población Estudiantes de Matemáticas, Característica Edad de ellos, la característica se designa con letras mayúsculas X, Y, Z,...., los valores de esas edades son numéricos entonces es una variable cuantitativa y los valores que toman se denotarían X=$\{x_1, x_2, x_3, \ldots\ldots x_n\}$.

- *Dominio* de la variable son los valores que toma

- *Recorrido* de la variable es la diferencia entre el valor mayor y el menor de los que toma la variable.

- *Variable unidimensional* Estudia solo una característica de la población. Ejemplo: Estudiar el peso (X)

- *Variable bidimensional* Estudia dos características de una población. Ejemplo Estatura(X) y peso (Y)

1.2.2 *Variables cualitativas.*

- *Este tipo de variables representan una cualidad o atributo que clasifica a cada caso en una de varias categorías.*
  - *Dicotómicas : La situación más sencilla es aquella en la que se clasifica cada caso en uno de dos grupos (hombre/mujer, enfermo/sano, fumador/no fumador).*
  - *Ordinal (escalas ordinales): Se requiere de un mayor número de categorías (color de los ojos, grupo sanguíneo, profesión, etc). Las variable atributo* se designan con letras A, B, C, ...y sus valores $A = a_1, a_2, \ldots\ldots\ldots\ldots, a_n$.

2 Etapas del análisis estadístico

1. Recogida de Datos
2. Ordenación de los mismos en tablas
3. Resumen de la información recogida a través de las medidas(*Descriptiva)*
4. Analizar los datos provenientes de una muestra para sacar conclusiones sobre la población de la que proviene la muestra *(Inferencial*).

3 Escalas de medida

- *Escala nominal* la característica estudiada se clasifica en una serie de características no numéricas y mutuamente excluyentes y no se puede establecer ningún orden entre ellos.

- *Escala ordinal* el carácter medido no es numérico pero puede establecerse algún tipo de orden. Ejemplo estudios de una persona.

- *Escala de intervalos* la característica puede cuantificarse numéricamente, estableciéndose intervalos entre dos operaciones. Ejemplo: Renta mensual que percibe una persona.

Análisis estadístico de distribuciones unidimensionales

- *Distribución unidimensional* está formada por los valores que toma la variable que se estudia acompañados de sus respectivas frecuencias.

4 Tablas de frecuencias

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

4.0.1 *Frecuencia absoluta* ($\mathbf{n_i}$)

- Se llama frecuencia absoluta del valor x_i al número de veces que aparece repetida la observación en la recopilación de datos. Se representa por n_i .
  - 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, n_i : número de veces que aparece cada valor
  -

xi	ni

4.0.2 *Frecuencia absoluta acumulada*

- $N_i = \sum_{j01}^{i} n_j$ es decir se suman las frecuencias anteriores a un valor dado, por tanto la acumulada al final coincide con la población N.

4.0.3 *Frecuencia relativa* ($\mathbf{f_i}$)

- Es el cociente entre la frecuencia absoluta y el número total de observaciones, por tanto la frecuencia relativa está siempre entre cero y uno. $f_i = \frac{n_i}{n}$

4.0.4 *Frecuencia relativa acumulada*

- $F_i = \sum_{j01}^{i} f_j$ es decir se suman las frecuencias anteriores a un valor dado, por tanto la acumulada al final coincide 1.

4.1 Distribución de datos

- *Distribución por datos no agrupados* es cuando se especifican todos y cada uno de los valores de la variable.

- *Distribución por datos agrupados* Cuando en una distribución estadística el número de valores que toma la variable es muy grande los valores de la variable se miden en intervalos , la amplitud del intervalo es la diferencia entre el extremo superior e inferior del intervalo y la suma de las amplitudes de todos los intervalos es igual al recorrido (diferencia entre el valor mayor y el menor de la distribución).
  - Como se hace:
    - ➢ 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.
- ≻ Es deseable que los extremos de los intervalos no coincidan con ninguno de los datos. Para ello, puede convenir que los extremos de los intervalos tengan una cifra decimal más que los datos.

- *Marca de clase de un intervalo* es la semisuma de los extremos del intervalo y es el valor que sustituye a todo el intervalo $x_i = \frac{l_{i-1}+l_i}{2}$ siendo el intervalo $[l_{i-1}, l_i]$.

4.2 Representaciones gráficas

Las representaciones gráficas tienen que estar hechas para que el simple impacto visual nos dé información de la distribución

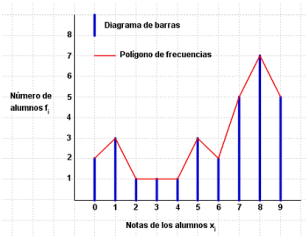
En distribuciones cuantitativas si los datos no están agrupados, se emplea el *diagrama de barras*, si están agrupados el *histograma.*, si la distribución es cualitativa se suele emplear el *diagrama de sectores.*

4.3 Graficos para variables cuantitativas discretas

4.3.1 *Diagrama de barras*

Datos sin agrupar y las barras proporcionales a las frecuencias.

- 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: n_i, fr 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.

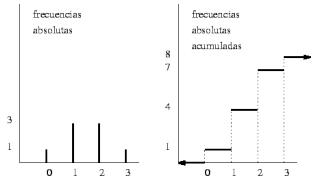


4.3.2 *Diagrama de barras acumuladas:*

- 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 %

- Se levanta para cada valor de la X una barra que representa la frecuencia acumulada de dicho valor. Si unimos mediante una poligonal los puntos más altos de cada barra obtenemos el polígono de frecuencias acumuladas.



4.4 Graficos para variables cuantitativas continuas

4.4.1 Intervalos con 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 de cada uno de dichos rectángulos el polígono de frecuencias.

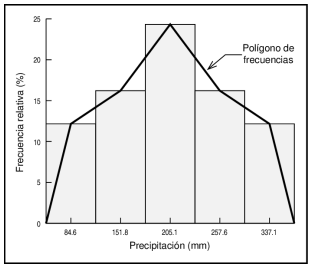
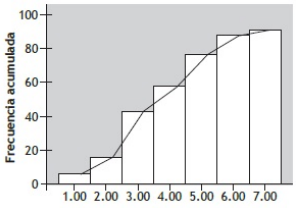


Diagrama de barras acumuladas: - En el eje de las X : Se representan los valores de la variable

- 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.



4.4.2 Intervalos diferente amplitud

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..

Densidad de frecuencia $d_i = f_i/a_i$

Diagrama en escalera para datos no agrupados se utiliza para las frecuencias acumuladas , son histogramas en los que en el eje vertical se acumulan las frecuencias absolutas, por eso se llaman en escalera.

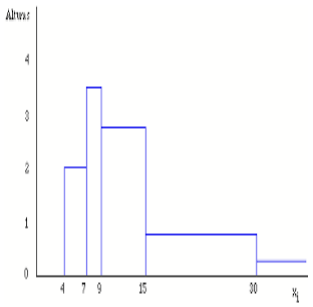


Diagrama de barras acumulado para datos agrupados

4.5 Graficos para variables cualitativas

- Gŕafico de sectores:El arco de cada porción se calcula usando la regla de tres

- Pictogramas:Expresan con dibujos alusivos al tema de estudio las frecuencias de las modalidad es de la variable. Estos gráficos se hacen representado a diferentes escalas un mismo dibujo,

Parámetros estadísticos

5 *Medidas de posición*

Se trata de resumir la información en un único número.Las medidas de posición pueden ser:

5.1 Centrales

5.1.1 *Media aritmética*

Se suman de todos los valores de la variable ponderados por sus frecuencias absolutas y dividido todo ello por el número total de observaciones

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i f_i}{N}$$

La ***media aritmética*** es siempre el ***centro de gravedad*** de la distribución y es siempre un valor que entra dentro del campo de variación de la variable.

Si los datos están agrupados en intervalos se toma la marca de clase de cada intervalo para su cálculo.

Propiedades

1.- Cuando a los valores de la variable se les suma una constante, la nueva media es la antigua más la constante.

2.- Si a los valores de la variable se les multiplica por una constante, la nueva media es la antigua multiplicada por la constante.

Como consecuencia de las dos anteriores si a los valores de una variable se les multiplica por constante y se les suma un número, la media aritmética queda multiplicada por la constante y sumado el número.

4.- La media aritmética se puede hacer siempre con variables cuantitativas y es perfecta, pero tiene un inconveniente que es que si los valores son muy extremos (desviados del resto), puede desvirtuarse la situación y hacerla poco representativa, debido a este problema, a veces se hace *la media truncada* que es quitar los extremos y hacer la media de los que quedan.

5.1.2 Media Geométrica

Es la raíz n-ésima del producto de los valores de la variable elevado cada uno de ellos a su frecuencia absoluta

$$G = \sqrt[n]{x_1^{f_1} \cdot x_2^{f_2} \cdot x_3^{f_3} \ldots \cdot x_n^{f_n}}$$

Tiene el problema de que su cálculo es muy complicado sobre todo si N es grande.

5.1.3 *Mediana*

Para datos no agrupados

- Se calcula primero el 50 % de la población N/2, se lleva ese valor a la columna de frecuencias absolutas acumuladas.
- Si el valor no está en la columna de acumuladas, se toma como valor de la mediana el de la variable correspondiente al siguiente.
- Si el valor si está en la columna de acumuladas, se toma como mediana la media aritmética del valor de la variable y el siguiente.

Esto es:

- Sean $x_1, x_2, x_3, \ldots, x_n$ los datos de una muestra "ordenada en orden creciente" y designando la mediana como M_e, distinguimos dos casos:

- a) Si "n es impar", la mediana es el valor que ocupa la posición $(n+1)/2$ una vez que los datos han sido ordenados (en orden creciente o decreciente), porque éste es el valor central. Es decir: $M_e = x_{(n+1)/2}$.

- Por ejemplo, si tenemos 5 datos, que ordenados son:$x_1 = 3, x_2 = 6, x_3 = 7, x_4 = 8, x_5 = 9 =>$ El valor central es el tercero: $x_{(5+1)/2} = x_3 = 7$. Este valor, que es la mediana de ese conjunto de datos, deja dos datos por debajo (x_1, x_2) y otros dos por encima de él (x_4, x_5).

- b) Si "n es par", la mediana es la media aritmética de los dos valores centrales. Cuando n es par, los dos datos que están en el centro de la muestra ocupan las posiciones n/2 y (n/2)+1. Es decir: $M_e = (x_{\frac{n}{2}} + x_{\frac{n}{2}+1})/2$.

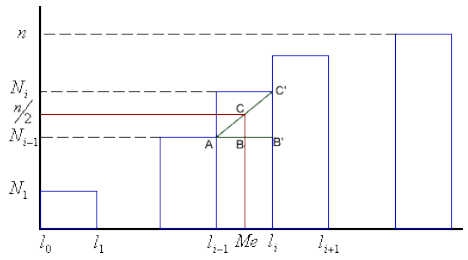
Por ejemplo, si tenemos 6 datos, que ordenados son: $x_1 = 3, x_2 = 6, x_3 = 7, x_4 = 8, x_5 = 9, x_6 = 10$. Aquí dos valores que están por debajo del $x_{\frac{6}{2}} = x_3 = 7$ y otros dos que quedan por encima del siguiente dato x_ $\frac{6}{2} + 1 = x_4 = 8$.

Por tanto, la mediana de este grupo de datos es la media aritmética de estos dos datos: $M_e = \frac{x_3+x_4}{2} = \frac{7+8}{2} = 7,5$.

Para datos agrupados en intervalos

- Se calcula como antes la mitad de la población, y se lleva ese valor a la columna de frecuencias absolutas acumuladas.

- Si el valor no está en la columna, se toma como intervalo al que pertenece la Mediana el siguiente al valor de N/2, y después de situarnos en el intervalo por la hipótesis de uniformidad hacemos una proporción entre la amplitud del intervalo, los elementos que tiene y la amplitud que correspondería a la diferencia entre N/2 y la frecuencia acumulada anterior valor que añadiríamos al extremo inferior del intervalo.



- Si el valor sí está en la columna de frecuencias acumuladas, se toma como Mediana el extremo superior del intervalo correspondiente.

- También se puede hallar gráficamente con el diagrama correspondiente a las frecuencias absolutas acumuladas.

- Al tratar con datos agrupados, si $\frac{n}{2}$ coincide con el valor de una frecuencia acumulada, el valor de la mediana coincidirá con la abscisa correspondiente. Si no coincide con el valor de ninguna abscisa, se calcula a través de semejanza de triángulos en el histograma o polígono de frecuencias acumuladas, utilizando la siguiente equivalencia:

$$\frac{N_i - N_{i-1}}{l_i - l_{i-1}} = \frac{\frac{n}{2} - N_{i-1}}{p} \Rightarrow p = \frac{\frac{n}{2} - N_{i-1}}{N_i - N_{i-1}}(l_i - l_{i-1})$$

- Donde N_i y N_{i-1} son las frecuencias absolutas acumuladas tales que $N_{i-1} < \frac{n}{2} < N_i, l_{i-1}$ y l_i son los extremos, interior y exterior, del intervalo donde se alcanza la mediana y

$$M_e = l_{i-1} + p$$

  - es la abscisa a calcular, la mediana. Se observa que $l_i - l_{i-1}$es la amplitud de los intervalos seleccionados para el diagrama.

5.1.4 *Moda*

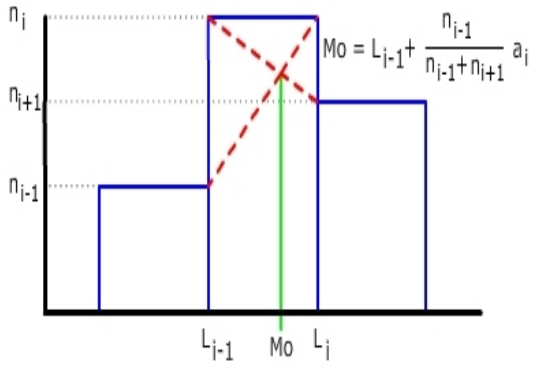
Es el valor de la variable que más veces se repite. En algunos casos existen varias modas, pero normalmente es una, si son dos se llama bimodal.

Para datos no agrupados La moda es el valor de la variable correspondiente a la mayor frecuencia absoluta.

Para datos agrupados en intervalos Se halla la densidad de frecuencia de cada uno de los intervalos (di) y el de mayor densidad de frecuencia se selecciona como intervalo modal, para determinar el valor de la Moda, se aplica la siguiente fórmula, basada en la proporcionalidad:

$$Mo = L_{i-1} + \frac{f_i - f_{i-1}}{f_i - f_{i-1} + f_i - f_{i+1}} \cdot a_i$$

$a_i$es la amplitud del intervalo modal
$f_i$es la frecuencia del intervalo modal
f_{i+1}es la frecuencia del intervalo siguiente al modal
f_{i-1}es la frecuencia del intervalo anterior al modal



Si los intervalos tienen todos la misma amplitud el intervalo modal es el de mayor frecuencia absoluta.

6 Medidas de posición no centrales

6.1 Cuantiles

Son medidas de posición que no tiene porqué ser central. Hay varios tipos de cuantiles:

1. *Cuartiles* Son valores de la variable que dividen a la distribución en cuatro partes iguales, por lo tanto los cuartiles son tres C_1 que deja por detrás de él al 25 % de la población, C_2 que divide a la población en dos partes iguales y C_3 que deja dtrás de él al 75 % de la población.

2. *Deciles* Son valores e la variable que dividen a la distribución en diez partes iguales, por lo tanto los deciles son nueve, D_1 deja al 10 % antes, D_2 al 20 % y así sucesivamente hasta D_9 que deja al 90 % antes y al 10 % después de él.

3. *Percentiles.-* Son valores de la variable que dividen a la distribución en cien partes iguales, por lo tanto los percentiles son 99.

En realidad tanto cuartiles como deciles se calculan con el correspondiente percentil.

$$D_1 = P_{10} \quad D_9 = P_{90} \quad C_1 = P_{25} \quad C_2 = D_5 = P_{50} = M_E \ .$$

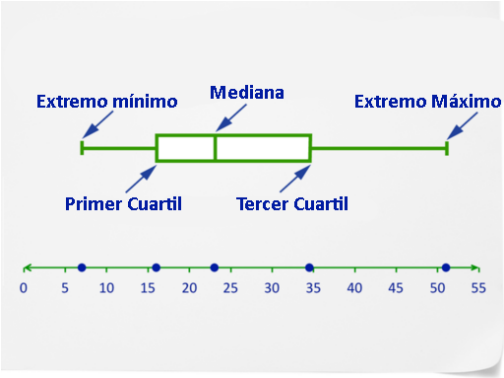
Para calcular cualquiera de ellos se utiliza por lo tanto el mismo procedimiento que el descrito en el cálculo de la Mediana.

El significado de cada una de las medidas de posición es muy claro: separa a los individuos de la población en dos fracciones cuyas proporciones respecto del total (25 %-75 %, 50 %-50 %, 10 %-90 %) quedan explícitas. Para la interpretación conjunta utilizaremos los diagramas de caja y bigotes

6.1.1 DIagrama de caja y bigotes

Se construyen del siguiente modo:

- La caja abarca el intervalo Q1, Q3 (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.



7 *Medidas de dispersión*

Las medidas de dispersión nos indican el mayor o menor alejamiento de los valores de una variable respecto a un promedio.Éstas complementan la información sobre la distribución de la variable, indicando si los valores de la variable están muy dispersos o se concentran alrededor de la medida de centralización.

Las medidas de dispersión absoluta más utilizadas son:

- Recorrido
  - Cuando se quieren señalar valores extremos en una distribución de datos, se suele utilizar la amplitud como medida de dispersión. La amplitud es la diferencia entre el valor mayor y el menor de la distribución.$R = x_n – x_1$
- Recorrido Intercuartílico
  - Diferencia entre el tercer y el primer cuartil de una distribución$RI = Q_3 – Q_1$
- Desviación Media
  - Es la suma de los valores en valor absoluto de la diferencia entre cada valor de la variable y la media aritmética por su frecuencia y dividido por el número de datos.

$$D_{\bar{x}} = \frac{\sum_{i=1}^{i=n} \left(x_i - \bar{x}\right) f_i}{N}$$

7.1 Varianza

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

Siempre es positiva (por estar al cuadrado). Como la varianza es siempre positiva, a mayor varianza mayor será la dispersión.

Propiedades:

1. *La varianza siempre es mayor o igual que cero. Tan solo hay un caso en que es cero y es cuando todos los valores de la variable son iguales.*
2. *Si a los valores de la variable le sumo una constante, la varianza de la nueva variable es la misma que la que tenía antes. Es decir si* $x_i{}' = x_i + K$ *entonces* $S_x^2{}' = S_x^2$
3. *Si a los valores de la variable se les multiplica por una constante, la varianza de la nueva variable es la que tenía por el cuadrado de la constante. Es decir si* $x_i' = kx_i$ *entonces* $S_x^2{}' = k^2 S_x^2$
4. *Es consecuencia de las dos anteriores, la varianza de la variable Y=aX+b es la varianza de X multiplicada por el cuadrado de a.*

$$S_y^2 = a^2 S_x^2$$

7.2 *Desviación Típica* s_x

Es la raíz cuadrada positiva de la varianza y es la medida de dispersión más utilizada.