La estadística y su ámbito.

Reseña histórica

El vocablo estadística deriva de la voz latina “status” en sus dos sentidos, como estado político y como situación geográfica.
El origen de la estadística se remonta al siglo XVIII en que, por un lado los juegos de azar, y por otro la ciencia política, impulsaron los estudios de probabilidad que dieron lugar a la teoría en la que hoy se sustenta la estadística.

Definición de estadística

- Estadística es todo cúmulo de datos cuantitativos relativos a un conjunto de individuos que presentan algún atributo en común o están afectados por causas comunes.

  (Esta acepción del termino se corresponde con las “series estadísticas”)

- Por Estadística entendemos aquellos métodos especialmente adecuados para dar significado a un conjunto de datos, afectados por el azar, usando instrumentos de la Matemática.

  (Esta acepción equivale a la expresión actual “Estadística Matemática”)

- Estadística es la recopilación, presentación, análisis e interpretación de datos numéricos extraídos de un conjunto de individuos, que nos permiten formular conclusiones válidas, efectuar decisiones lógicas basadas en dicho análisis y extender los resultados desde un grupo pequeño hacia una población.

  (Esta acepción engloba los métodos utilizados por la “Estadística descriptiva” y también por la “Estadística Inductiva o Inferencial”)

Estadística descriptiva e inferencial

Estadística Descriptiva se refiere al conjunto de técnicas que se ocupan de, interpretar los resultados numéricos, elaborar tablas y gráficos explicativos y, posteriormente inferir parámetros estadísticos que caracterizan al conjunto total de datos recolectados.

Estadística Inferencial trata de la generalización hacia las poblaciones de los resultados obtenidos en las muestras y de las condiciones bajo las cuales estas conclusiones son válidas.

Términos frecuentemente utilizados en estadística

- ***Población***: *"Se denomina población al conjunto de todos los elementos que cumplen una determinada característica, que deseamos medir o estudiar."*

  Ej: Si se quiere determinar la vida útil de un lote de lámparas producidas por un turno en una fábrica, la población está constituida por todas las lámparas del lote.

- ***Muestra***: *"Se denomina muestra a cualquier subconjunto de la población."*

  Ej: En el caso de las lámparas, la muestra serán un conjunto de 10 ó 20 luminarias extraídas al azar del lote.

- ***Unidad estadística***: *"En Estadística se considera unidad estadística a cada individuo de una población."*

  Ej: En el ejemplo de las lámparas, cada lámpara es una unidad estadística.

- ***Carácter estadístico***: *"Cada una de las propiedades o aspectos que pueden estudiarse en los individuos de una población recibe el nombre de carácter estadístico."*

  Ej: Para el lote de lámparas, la duración de su vida antes de fallar es el carácter analizado.

Carácter estadístico cualitativo o atributo: es aquel no susceptible de ser medido ni contado.

Ej: estado general de un animal, la nacionalidad.

En el proceso de medición de estas variables, se pueden utilizar dos escalas:

Escalas nominales: ésta es una forma de observar o medir en la que los datos se ajustan por categorías que no mantienen una relación de orden entre sí (color de los ojos, religión, profesión).

Escalas ordinales: en las escalas utilizadas, existe un cierto orden o jerarquía entre las categorías (grados de fatiga, estado de un tumor).

Carácter estadístico cuantitativo: es el que surge de un proceso de medición o conteo.

Se distinguen carácter estadístico *discreto* y *continuo*:

"El carácter estadístico es **discreto** solo si entre dos valores consecutivos del mismo no puede existir un valor intermedio. Esto indica que su recorrido es un conjunto definido en el campo de los números naturales o enteros."

"El carácter estadístico es **continuo** si entre dos valores cualesquiera de su recorrido puede existir siempre uno intermedio. Esto indica que su recorrido es un intervalo incluido en el conjunto de los números reales."

Series estadísticas o distribuciones estadísticas

Series cronológicas: el ordenamiento se realiza teniendo en cuenta el tiempo.

Serie espacial o geográfica: en este caso el fenómeno en estudio es estudiado a través del territorio o del espacio.

Serie o distribución de frecuencias: es la correspondencia que hay entre cada valor de la variable (que se denomina solamente variable para la variable discreta y clase para variables continuas) y su respectivo número de observaciones o frecuencias.

Esquema de realización de un trabajo estadístico

Especificación del problema

En esta etapa se deben:

- Establecer con precisión el problema a tratar.
- Definir los objetivos del trabajo.
- Seleccionar el material experimental.

Recolección y ordenación de datos

Para recopilar los datos es necesario proceder con el mayor orden posible y, por lo tanto tener en cuenta los siguientes aspectos:

- Fijar los procedimientos para realizar el experimento.
- Tener a disposición todos los elementos requeridos para recoger dichos datos.
- Examinar el tipo de datos requeridos, es decir si son cuantitativos o cualitativos.
- Disponer los datos en forma creciente o decreciente, según convenga, para que sean de fácil ubicación y análisis.
- Encontrar el rango de variación de los datos recolectados para hallar entre qué valores máximos y mínimos se hallan comprendidos.

Organización de distribuciones de frecuencias

Para facilitar el trabajo de búsqueda y no tener que repetirlo cada vez que analicemos un aspecto del registro, es conveniente organizarlos de alguna manera sistemática; es decir de mayor a menor o viceversa.
Cuando se dispone de una gran cantidad de valores recopilados es conveniente condensar, simplificar o resumir la totalidad de las observaciones.

Distribución de frecuencias para la variable aleatoria cualitativa (atributo)

La distribución de frecuencias para un atributo muestra el número de observaciones para cada una de las clases o categorías del mismo.

Frecuencias absolutas: f_i número total de observaciones que pertenecen a cada clase o categoría.

Frecuencias relativas: f_{ir} relación entre la frecuencia absoluta de cada modalidad y el numero total de observaciones $f_{ir} = \frac{f_i}{n}$; siendo n el número total de casos.

La frecuencia relativa es para las series de frecuencias un concepto análogo al de probabilidad en las variables aleatorias discretas.

Frecuencias relativas porcentuales: $f_{ir}\%$ expresión porcentual de la frecuencia relativa.

Distribución de frecuencias para la variable aleatoria cuantitativa discreta

La distribución de frecuencias para una VAD muestra el número de observaciones para cada uno de los valores de la misma.

Frecuencias absolutas, frecuencias relativas y frecuencias relativas porcentuales: según lo definido anteriormente.

Frecuencias acumuladas: F_k es la suma de las frecuencias absolutas hasta un determinado valor de la variable inclusive. $F_k = \sum_{i=1}^{k} f_i$

Frecuencias acumuladas relativas: F_{kr} es la relación entre las frecuencias acumuladas y el numero total de casos. $F_{kr} = \frac{F_k}{n}$

Frecuencias acumuladas relativas porcentuales: $F_{kr}\%$ expresión porcentual de la frecuencia relativa acumulada.

Distribución de frecuencias para la variable aleatoria cuantitativa continua

Como la VAC puede tomar un sinnúmero de valores entre dos cualesquiera de su recorrido, ésta se presenta en una distribución de frecuencias agrupada en intervalos llamados intervalos de clase; de manera tal que cada valor individual estará contenido en alguno de ellos.

Elementos necesarios para confeccionar una tabla para el caso VAC:

Rango: se denomina así a la diferencia entre el mayor y menor valor observado de la serie de datos.

$$R = X_{max} - x_{min}$$

El intervalo $X_{max} - x_{min}$ se llama recorrido de la VAC.
El rango se divide en varios intervalos conocidos como intervalos de clase.

Intervalos de clase: se llama de esta manera a cada uno de los subintervalos en que queda dividido el recorrido de la VAC.

Número de intervalos de clase: el número óptimo de intervalos de clase depende de la cantidad de observaciones realizadas.

Fórmulas que dan un indicio del número óptimo a tomar:

" $K = \sqrt{N}$ " ó " $K = \sqrt{n}$ " siendo N el numero total de casos de la población y n el numero total de individuos en una muestra extraída de una población.

Formula de Sturger: " $K = 1 + 3{,}3 . \log N$ " siendo N el numero total de casos de la población.

Amplitud o tamaño de clase: en una distribución de frecuencias los intervalos pueden tener todos la misma amplitud o distinta.

La amplitud se obtiene efectuando el cociente entre el rango y el número de intervalos considerados: $c = \frac{R}{K}$

Límites de clase: son los extremos de un intervalo de clase. La determinación de la cantidad de cifras significativas de los mismos depende de los valores alcanzados por la variable, pudiendo o no coincidir el límite superior de una clase con el inferior del siguiente.

Límites reales de clase: (LRI - LRS) el límite real superior de una clase es igual a la semisuma del límite superior de dicha clase más el inferior de la siguiente. El límite real inferior de una clase es la semisuma del límite inferior de una clase y el superior de la anterior.

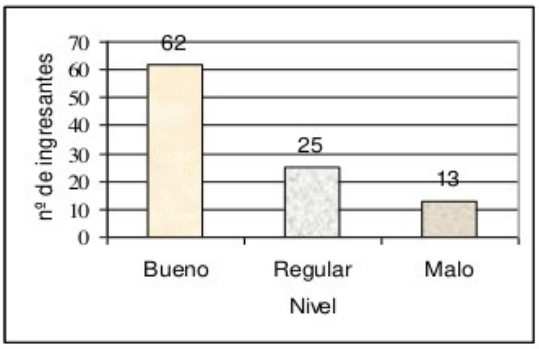
Marca de clase: M_i es el punto medio de cada intervalo de clase y se obtiene efectuando el promedio de los límites de clase o de los límites reales de clase.

Representación de datos por medio de gráficos.

Diagrama de barras: se llama diagrama de barras al gráfico que asocia a cada valor de la variable una barra, generalmente vertical, proporcional a la frecuencia con que se presenta.

Para la VA cualitativa resulta:

Nivel	f_i
Bueno	62
Regular	25
Malo	13
Total	100

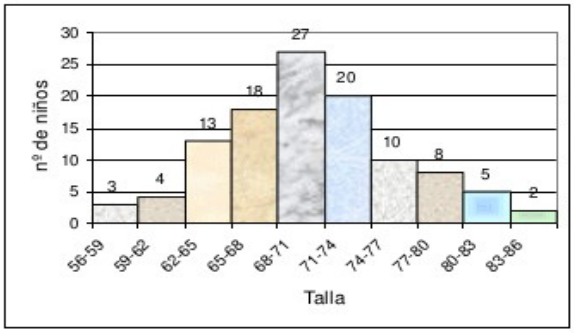


Si se hace diagramas sobre datos geográficos, generalmente se usan barras horizontales:

Histogramas: los histogramas son diagramas de barras para las variables continuas, es decir para las que se agrupan en intervalos de clase.
Un histograma es un conjunto de rectángulos cuyas bases coinciden con el tamaño de clase, sus puntos medios son las marcas de clase y sus extremos son los límites reales de los intervalos. En general las alturas corresponden a la frecuencia absoluta de cada intervalo, aunque en algunas ocasiones la frecuencia absoluta se relaciona con el área de cada rectángulo.

Para el ejemplo es:

X_i	f_i
56-59	3
59-62	4
62-65	13
65-68	18
68-71	27
71-74	20
74-77	10
77-80	8
80-83	5
83-86	2
Σ	110



Polígono de frecuencias: si en un histograma se proyectan las marcas de clase sobre las bases superiores de los rectángulos, y a los puntos obtenidos se los une mediante tramos rectos, se obtiene una poligonal cuyos vértices representan las frecuencias absolutas de cada una de las clases.

Para nuestro ejemplo sobre la talla de los recién nacidos será:

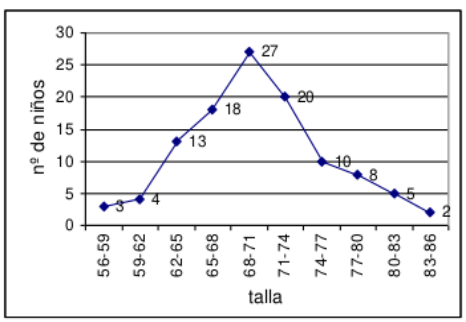


Diagrama de sectores o diagramas de torta: En los diagramas de sectores cada suceso está representado por un sector circular de una amplitud proporcional a su frecuencia. La amplitud de cada sector circular se obtiene mediante una simple regla de tres.

Para la siguiente tabla:

ARGENTINA					
Idiomas	%	Grupos étnicos	%	Religiones	%
Español	96	Europeo	85	Católica	93
Italiano	2	Amerindio/Mestizo	15	Protestante	2
Amerindio	1			Judía	1
Otro	1			Otra	4

Fuente: Programa PCGlobe-1992

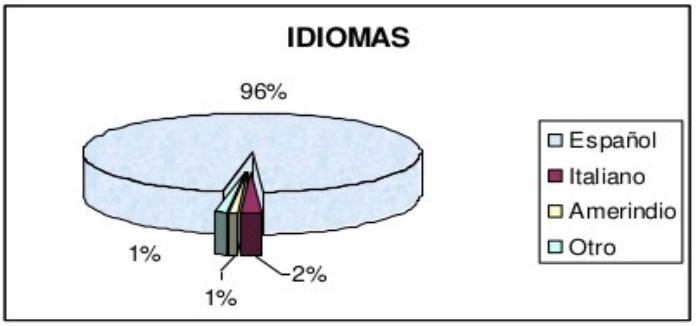
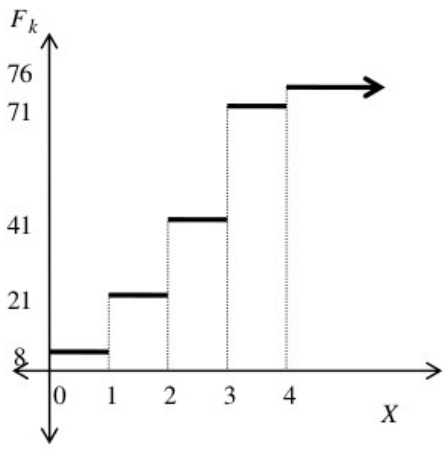


Diagrama de frecuencias acumuladas: esta representación gráfica se utilizar para mostrar frecuencias acumuladas de las variable aleatoria discretas. El modelo es una función “escalonada”, es decir con tramos constantes entre dos valores consecutivos de la variable, cuyos valores se corresponden con las frecuencias absolutas acumuladas o relativas acumuladas hasta el valor considerado.

Para el ejemplo que muestra el número de hijos en edad escolar de los empleados de una fábrica es:

X	f_i	F_k	F_{kr}	$F_{kr}\%$
0	8	8	0,105	11
1	13	21	0,276	28
2	20	41	0,539	54
3	30	71	0,934	93
4	5	76	1	100
	$\Sigma = 76$			



Obtención de parámetros característicos

Se llaman parámetros característicos a todos aquellos valores que describen de manera precisa a un conjunto de datos. Existen distintos tipos de parámetros:

Medidas de Centralización: En la mayoría de los casos los datos de una serie de frecuencias tienden a agruparse alrededor de un punto central. Estos valores centrales son útiles pues representan a todos los valores de dicha serie y por esto se los conoce también como medidas de posición.

- Media aritmética: la media aritmética o simplemente media es el promedio de una serie de datos.

  La MA se calcula sumando el valor de todos los datos y dividiendo el resultado por el número total de ellos. Si la variable toma los valores $X_1; X_2; X_3; \ldots; X_i$

  En símbolos: $\bar{x} = \sum_{i=1}^{n} x_i$

- La mediana: la mediana es el valor que divide al conjunto en dos partes iguales. Por esto, se denomina mediana al valor central de los datos cuando éstos se han dispuestos ordenadamente de menor a mayor, sin importar qué valores toma dicha variable.

  Si el número de datos es impar su cálculo es directo, pero si la cantidad de datos es par su valor se obtiene haciendo la semisuma de los dos centrales.

  Ejemplo:

  Para los datos dados, previo ordenamiento de los mismos resulta:

  12 15 17 23 25 28 $\qquad$ $\text{Me} = \frac{17 + 23}{2} = 20$

  Si en cambio tenemos:

  12 15 17 23 25 $\qquad$ Me = 17

  Si los datos se presentan en intervalos de clase la fórmula para su cálculo es:

  $$Me = LRI + \frac{c.\ \frac{n}{2} - \sum f\ \ ant}{f_{Me}}$$ siendo:

  LRI: límite real inferior de la clase que contiene a la mediana
  c: tamaño de clase
  $\sum f\ ant$: sumatoria de las frecuencias anteriores a la clase que contiene a la Me
  f_{Me}: frecuencia de la clase mediana

- Moda: la moda es el valor de un conjunto de datos que se repite con mayor frecuencia. Esta medida es la más adecuada si se trabaja con datos cualitativos. En los casos en que dos valores se repiten mayoritariamente se dice que la distribución es bimodal y si existen más de dos valores se llama multimodal.

  Si los datos se presentan en intervalos de clase la fórmula para su cálculo es:

  $$M_o = LRI + \frac{c.\Delta_1}{\Delta_1 + \Delta_2}$$

  LRI: límite real inferior de la clase modal
  c: tamaño de clase
  Δ_1 = diferencia entre la frecuencia absoluta de la clase modal y la frecuencia de la clase inmediata anterior
  Δ_2 = diferencia entre la frecuencia absoluta de la clase modal y la frecuencia de la clase inmediata posterior

Medidas de dispersión: indican el alejamiento de los valores de la variable con respecto a sus medidas de centralización; siendo por lo tanto una forma de evaluar la heterogeneidad de los datos.

Medidas de dispersión absolutas:

- Rango: es la diferencia entre el mayor y el menor de los valores que toma nuestra variable.

$$R = X_{max} - x_{min}$$

- Desvíos: el desvío de cada valor de la VA con respecto a la media aritmética es igual a la diferencia entre dicho valor y la media de un conjunto de datos.

$$d_i = x_i - \bar{x}$$

- Desviación media: es el promedio de los valores absolutos de los desvíos antes definidos.

$$DM = \frac{\sum_{i=1}^{n} |d_i|}{n}$$

Para datos agrupados en series de frecuencias $DM = \frac{\sum_{i=1}^{n} |d_i| \cdot f_i}{n}$

- Varianza: se define como el promedio de los cuadrados de los desvíos con respecto a la media aritmética.

Podemos simbolizarla como: $Var(X) = \frac{\sum_{i=1}^{n} d_i^{\,2}}{n} = \frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}{n}$

Si los datos se repiten con una determinada frecuencia absoluta f_i cada uno de ellos resulta: $Var(X) = \frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2 \cdot f_i}{n}$

Si se trabaja con una población la varianza se denota con σ_X^2, pero si se analizan datos muestrales su símbolo es s^2.

- Desviación típica o estándar: esta es la medida de uso más frecuente y se calcula mediante:

$\sigma = \sqrt{Var(X)}$ para población $\qquad s = \sqrt{Var(x)}$ para muestras

Medidas de dispersión relativas:

- Coeficiente de variación: es la desviación típica expresada en porcentaje de la media aritmética.

$$\text{En símbolos: } C.V. = \frac{\sigma}{\overline{X}}.100$$

Medidas de concentración:

- Cuartiles: son los valores de la variable aleatoria que dividen al conjunto en cuatro grupos iguales.

  Si los datos están agrupados en intervalos de clase, se obtienen mediante:

$$Q_i = LRI + \frac{\left(\frac{i.n}{4} - \sum f \quad ant\right).c}{f_{Q_i}}$$ donde:

LRI: límite real inferior de la clase que contiene al cuartil
$f_{Q\,i}$: frecuencia de la clase del cuartil ***i***.
c: tamaño de clase

- Deciles: son los valores de la variable aleatoria que dividen al conjunto en diez grupos iguales.

  Si los datos están agrupados se obtienen mediante:

$$D_i = LRI + \frac{\left(\frac{i.n}{10} - \sum f \quad ant\right).c}{f_{D\,i}}$$ donde:

LRI: límite real inferior de la clase que contiene al decil
$f_{D\,i}$: frecuencia de la clase del decil ***i***.
c: tamaño de clase

- Percentiles: son los valores de la variable aleatoria que dividen al conjunto en cien grupos iguales.

  Si los datos están agrupados se obtienen mediante:

$$P_i = LRI + \frac{\left(\frac{i.n}{100} - \sum f \quad ant\right).c}{f_{P\,i}}$$ donde:

LRI: límite real inferior de la clase que contiene al percentil
$f_{P\,i}$: frecuencia de la clase del percentil ***i***.
c: tamaño de clase

Forma de una distribución

La desviación típica y las demás medidas de dispersión miden la disgregación de una distribución de frecuencias.

- Sesgo: mide el grado de concentración de los datos de una distribución a un lado y otro de la media; expresando, por lo tanto la asimetría de la misma.

  Si una distribución es unimodal y simétrica la media aritmética, la mediana y la moda coinciden, $\bar{x} = Me = Mo$.

  En una distribución con **sesgo a la derecha**, la media aritmética es la mayor de las tres, pues en ella influyen los valores muy altos de la variable y queda la mediana entre la moda y la media
  $Mo < Me < \bar{x}$

  En una distribución **sesgada a izquierda**, la media es el menor valor, ubicándose la mediana entre ella y la moda, que resulta ser el mayor valor.
  $\bar{x} > Me > Mo$

- Coeficiente de Pearson:

  Se define como:

  $$A_s = \frac{\bar{X} - M_o}{s}$$

  $\bar{X}$: media aritmética
  M_o : moda
  s : desviación estándar

  Siendo cero cuando la distribución es simétrica, positivo cuando existe asimetría a la derecha y negativo cuando existe asimetría a la izquierda.

- Coeficiente de asimetría de Fisher: se basa en las diferencias $x_i - \bar{x}$. Como analizamos al estudiar los desvíos, su media es siempre nula; si las elevamos al cuadrado, serían siempre positivas, por lo que tampoco servirían, por lo tanto se elevan esas diferencias al cubo. Para evitar el problema de la unidad, y hacer que sea una medida adimensional y por lo tanto relativa, dividimos por el cubo de su desviación típica. Con lo que resulta la siguiente expresión:

  $$A_s = \frac{\sum_{i=1}^{n} \frac{X_i - \bar{X}^{\,3} \cdot f_i}{n}}{s^3}$$

  X_i : valores de la variable
  $\bar{X}$: media aritmética
  s : desviación estándar
  f_i : frecuencia absoluta de cada valor de la variable

  n : cantidad total de datos

  La interpretación del coeficiente de Fisher es la misma que la del coeficiente de Pearson, si la distribución es simétrica vale cero, siendo positivo o negativo cuando exista asimetría a la derecha o izquierda respectivamente.