Definición de estadística*:* 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.

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***:* conjunto de todos los elementos que cumplen una determinada característica, que deseamos medir o estudiar.

•**Muestra***:* cualquier subconjunto de la población.

•**Unidad estadística***:* cada individuo de una población.

•**Carácter estadístico***:* cada una de las propiedades o aspectos que pueden estudiarse en los individuos de una población.

•**Carácter estadístico cualitativo***:* es aquel no susceptible de ser medido ni contado.

→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.

•Posee carácter **cuantitativo discreto** si entre dos valores consecutivos del mismo no puede existir un valor intermedio.

•Posee carácter **cuantitativo continuo** si entre dos valores cualesquiera de su recorrido puede existir siempre uno intermedio.

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.

Tipos de frecuencia

•**Frecuencias absolutas (f_i - A)**: número total de observaciones que pertenecen a cada clase o categoría.

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

•**Frecuencias relativas porcentuales ($f_{ir}\%$ - RP)**: expresión porcentual de la frecuencia relativa.

•**Frecuencias acumuladas (F_k - 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} - KR):** es la relación entre las frecuencias acumuladas y el número total de casos. $F_{kr}=\frac{F_k}{n}$

•**Frecuencias acumuladas relativas porcentuales ($F_{kr}\%$ - KRP):** expresión porcentual de la frecuencia relativa acumulada.

Distribución de frecuencias para la variable cualitativa: **A-R-RP**
Distribución de frecuencias para la variable cuantitativa discreta: **A-R-RP-K-KR-KRP**
Distribución de frecuencias para la variable cuantitativa continua:

Rango: diferencia entre la máxima observación y la mínima: $R=x_{max}-x_{min}$

Recorrido*:* el intervalo que va de x_{max} a x_{min}.

Intervalos de clase: se llama de esta manera a cada uno de los sub-intervalos en que queda dividido el recorrido.

Número de intervalos de clase: $K=\sqrt{n}$; $K=1+3{,}3\ln(n)$

Amplitud o tamaño de clase ($c=\frac{R}{K}$): la amplitud se obtiene efectuando el cociente entre el rango y el número de intervalos considerados.

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: el LRS de una clase es igual a la semisuma (sumar dos números y dividirlos por dos) del LS de dicha clase más el LI de la siguiente. El LRI de una clase es la semisuma del LI de una clase y el LS 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:** cualitativo y cuantitativo discreto.
•**Histogramas:** cuantitativo continuo.
•**Polígono de frecuencias:** cuantitativo.
•**Gráfico circular:** cualitativo, cuantitativo.
•**Diagrama de frecuencias acumuladas:** cuantitativo discreto.
•**Ojiva:** cuantitativo continuo.

Parámetros estadísticos

•Número que resume la gran cantidad de datos que pueden derivarse del estudio de una variable estadística.

Medidas de centralización: valores que suelen situarse cerca del centro de la distribución de datos.

•Media aritmética ($\acute{x}$): es el promedio de una serie de datos.

Cuantitativa continua: $\acute{x}=\frac{\sum_{i=1}^{k} M_i x_i}{n}$

Cuantitativa discreta: $\acute{x}=\frac{\sum_{i=1}^{k} x_i f_i}{n}$

•Mediana (*Me*): valor central de los datos cuando éstos se han dispuestos ordenadamente de menor a mayor que divide al conjunto en dos partes iguales.

$$pos=\frac{n}{2}$$

→Cuantitativo discreto:

n impar: se obtiene buscando el x_i correspondiente a la F_k que contiene a la pos.

n par: se obtiene sumando el x_i correspondiente a la F_k que contiene a la pos con el x_i siguiente a este y luego dividiendo esta suma por dos.

→Cuantitativo continuo: $Me=LRI+c\frac{pos-\sum f_{ANT}}{f_{iMe}}$

•Moda**:** valor de un conjunto de datos que se repite con mayor frecuencia**.**

→Cuantitativo continuo: $Mo=LRI+c\frac{\Delta_1}{\Delta_1+\Delta_2}$

Donde: $\Delta_1=f_{iMo}-f_{iANT}$

$\Delta_2=f_{iMo}-f_{iPOST}$

Medidas de dispersión: Las medidas de dispersión completan el análisis de una serie de datos pues determinan la mayor o menor separación de los datos con respecto a su valor central. Es decir que 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: $R = X_{max} - x_{min}$

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

$d_i = x_i - \acute{x}$ (variable discreta)
$d_i = M_i - \acute{x}$ (variable continua)

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

Datos no tabulados: $Dm = \frac{\sum |d_i|}{n}$

Datos tabulados: $Dm = \frac{\sum |d_i| f_i}{n}$

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

Datos no tabulados: $Var(x) = \sigma^2 = s^2 = \frac{\sum d_i^{\,2}}{n}$

Datos tabulados: $Var(x) = \sigma^2 = s^2 = \frac{\sum d_i^{\,2} f_i}{n}$

•Desviación estándar: esta es la medida de uso más frecuente y se calcula mediante:

$\sqrt{Var(x)} = \sigma = s$

Medidas de dispersión relativas:

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

$CV = 100 \frac{\sigma}{\acute{x}}$

Medidas de concentración:

•Cuartiles ($Q_i : v=4$) , Deciles ($D_i : v=10$) y Percentiles ($P_i : v=100$):

•Generaliles (G_i):

→Cuantitativo discreto:

$$pos=\frac{i*n}{v}$$

El 'generalil' se encuentra buscando el x_i correspondiente a la F_k que contiene a la pos.

→Cuantitativo continuo: $G_i=LRI+c\frac{pos-\sum f_{ANT}}{f_{iGi}}$

Medidas de forma:

•Distribución simétrica: se cumple que $\acute{x}=Me=Mo$.

•Distribución asimétrica:

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.

→A derecha:

•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: $\acute{x}>Me>Mo$

•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: $Mo<Me<\acute{x}$

•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:

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

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

Regresión y correlación lineal

Variables que intervienen

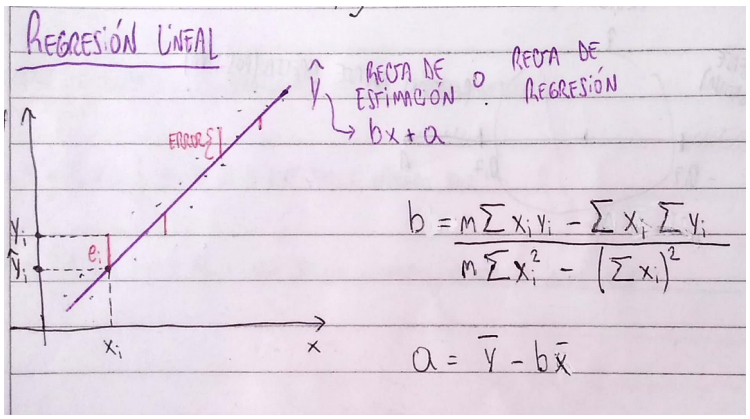
•x e y determinísticas: Conocido el valor de x, el valor de y queda perfectamente establecido.
•x: determinística e y: probabilística (análisis de regresión).
•x e y probabilísticas (análisis de correlación-regresión): conocido el valor de x, el valor de y no queda perfectamente establecido.

Gráfico de dispersión: Es una gráfica que se representa en el sistema de ejes cartesianos los pares ordenados, correspondientes a los datos apareados que resultan de las mediciones.

Tipos de relaciones

Perfectas Imperfectas Positivas Negativas Sin Relación

Regresión: La regresión mide en forma funcional, a través de una ecuación, la posible relación entre las variables con el objetivo de predecir una de ellas en función de la/s otra/s.



Correlación: La correlación se dirige sobre todo a medir la intensidad de la asociación entre variables numéricas.

Coeficiente de correlación (r):

$$r = \frac{Cov(x,y)}{\sigma x * \sigma y}, \text{ donde } Cov(x,y) = \frac{\sum (x_i - \acute{x})(y_i - \acute{y})}{n}$$

•Campo de variación de r: $-1 \leq r \leq 1$

•Si $r \leq -0,7$: relación fuerte negativa.

•Si $r \geq 0,7$: relación fuerte positiva.

Coeficiente de determinación ($D = r^2$):

•Campo de variación de D: $0 \leq D \leq 1$

•Si $D \geq 0,5$ entonces existe una fuerte vinculación entre x e y