1. ANÁLISIS DESCRIPTIVO DE LOS DATOS

1. VARIABLES

Una variable es una característica cambiante y medible del objeto de estudio.

ESCALAS DE MEDICIÓN

Precisión con la que se miden las variables. Son las siguientes:

- **Cualitativas**: cuando la variable es una cualidad (letras), puede ser:
  - **Nominal**: sus valores son nombres.
  - **Ordinal**: sus valores son nombres que además tienen un orden.
- **Numéricas**: cuando la variable es un número, puede ser:
  - Según la secuencia de los datos: **discreta** (sólo números enteros) y **continua** (números enteros y decimales).
  - Según el valor del punto cero: **intervalar** (el 0 es convencional por lo que acepta valores negativos, por ejemplo, 0 grados centígrados es el punto de congelación del agua, no la ausencia absoluta de calor) y de **razón** (el 0 es absoluto, por ejemplo, talla, peso, etc.).

2. ANÁLISIS DESCRIPTIVO

MEDIDAS DE TENDENCIA CENTRAL

Estimación del punto medio alrededor del cual se agrupan los datos. Son diferentes según la escala de medición así:

- Para escalas cualitativas:
  - **Porcentajes**: cantidad que corresponde proporcionalmente a una parte de cien.
  - **Moda**: valor que más se repite (y más de 1 vez).
- Para escalas numéricas:
  - **Media**: es el promedio aritmético de las observaciones. Es la que más se utiliza porque: (1) es la más **representativa,** dado que tiene en cuenta los valores todas las observaciones, y (2) en una muestra, es la más **confiable** en su estimación de la tendencia central de la población. No se usa cuando existen datos imprecisos que no se pueden sumar (como "mayor a 10") o datos inusualmente extremos (por ejemplo, un adulto mayor en una población juvenil) con lo cual la media no indica adecuadamente la tendencia central; en estos casos se utiliza la mediana. Convencionalmente, las medidas de una población (parámetros) se escriben con letras griegas, por ejemplo: μ (que se lee "mi" y es la letra "m" minúscula en griego) para la media poblacional; y las de una muestra (estadísticos) se escriben con letras romanas, por ejemplo: $\bar{x}$ (que se lee "x barra") para la media muestral. Las fórmulas son:

    De una muestra: $\bar{x} = \frac{\sum x}{n}$ *De una población:* $\mu = \frac{\sum x}{N}$

    O sea, la media ($\bar{x}$ o μ) es igual a la sumatoria (∑) de todos los datos individuales (x) dividida entre el tamaño de la muestra: (n) o de la población (N).
  - **Mediana**: es el punto a partir del cual la mitad de los datos son mayores y la otra mitad son menores.

    Ejemplos:

Los datos:	*Se ordenan:*	*Se identifica el puesto: (n+1)/2*	*Y se identifica la mediana:*
4, 3, 5, 4, 6 →	*3, 4, 4, 5, 6*	*(5+1)/2 = 3º*	*Primero: 3. Segundo: 4. Tercero: 4. La mediana es: **4**.*
7, 5, 9, 10 →	*5, 7, 9, 10*	*(4+1)/2 = 2 ½*	*Primero: 5. Segundo: 7. Mitad entre segundo y tercero: (7+9)/2 = 8. La mediana es **8**.*

MEDIDAS DE DISPERSIÓN

Estimación el grado de separación de los datos; se utilizan sólo en escalas numéricas. Las principales son las siguientes:

- **Desviación estándar:** es la raíz cuadrada de la **varianza**, esto es, de la "media de las desviaciones elevadas al cuadrado"; se utiliza conjuntamente con la media. No obstante, para obtener la desviación estándar de una muestra se usa el número total de datos menos 1, porque así aumenta la desviación estándar muestral (a más pequeña la muestra, más se incrementa), con lo cual se asemeja más a la desviación estándar de la población, ya que las muestras pequeñas tienden a no incluir los datos extremos de la población por su baja frecuencia (a más pequeñas, más aún), con lo cual tienden a mostrar una dispersión menor que la del universo. En fórmula:

$$s = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$$

O sea, la desviación estándar de una muestra (s) es igual a la raíz cuadrada de: la suma (∑) de las desviaciones (cada dato individual menos la media: x-$\bar{x}$) elevadas al cuadrado, dividida entre el número total de datos (n), menos 1.

$$\sigma = \sqrt{\frac{\sum(x-\mu)^2}{N}}$$

*O sea, la desviación estándar de una población (*σ, letra griega "sigma" equivalente a la S mayúscula*) es igual a la raíz cuadrada de: la suma (∑) de las desviaciones (cada dato individual menos la media: x-$\bar{x}$) elevadas al cuadrado, dividida entre el número total de datos (N).*

Ejemplo: los datos 1, 2, 3, 4, 5. La media es: 1 + 2 + 3 + 4 + 5 = 15; 15 ÷ 5 = 3. La desviación estándar se obtiene así:

Se obtiene la desviación de cada dato (se le resta la media):	*Se elevan las desviaciones al cuadrado:*	*Se obtiene la media de las desviaciones elevadas al cuadrado (varianza):*	*Se obtiene la raíz cuadrada de la varianza:*
1 – 3 = -2 *2 – 3 = -1* *3 – 3 = 0* *4 – 3 = 1* *5 – 3 = 2*	*$-2^2 = 4$* *$-1^2 = 1$* *$0^2 = 0$* *$1^2 = 1$* *$2^2 = 2$*	$\frac{4+1+0+1+4}{5-1} = 2.5$	$s^2 = 2.5$ $s = \sqrt{2.5} = 1.58$

- **Coeficiente de variación:** porcentaje de dispersión en relación con la media. Se calcula dividiendo la desviación estándar entre la media y multiplicando por 100. Permite comparar unidades diferentes. En fórmulas:

De una muestra: $V = \frac{s}{\bar{x}} \times 100\%$ *De una población:* $V = \frac{\sigma}{\mu} \times 100\%$ ***Ejemplos:*** $V = \frac{0.1}{20.0} \times 100\% = 0.5\%$ $V = \frac{0.006}{0.787} \times 100\% = 0.8\%$

Ejemplo: con dos adipómetros de diferente marca, uno en milímetros y otro en pulgadas, se midió repetidamente el pliegue tricipital de un voluntario con los siguientes resultados: para un adipómetro μ = 20.0 mm y σ = 0.1 mm; y para el otro μ = 0.787" y σ = 0.006"; los coeficientes fueron: 0.5% y 0.8% respectivamente, por lo que se puede concluir que el primero es más preciso.

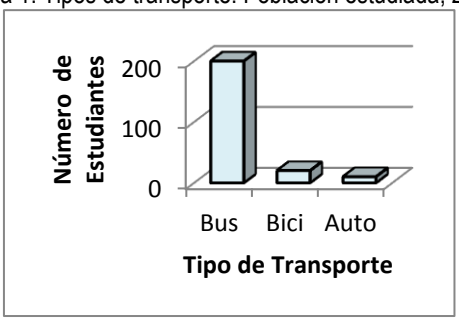
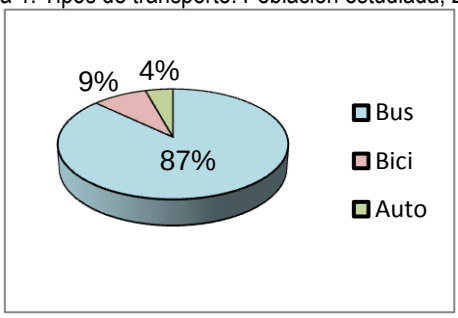
- **Rango intercuartil:** es la distancia entre el percentil 25 (primer cuartil, Q_1) y el 75 (tercer cuartil, Q_3), o sea, los puntos que separa los 25 y 75 valores menores de los mayores, respectivamente, ordenados de menor a mayor. Se utiliza conjuntamente con la mediana (que es igual al percentil 50 o segundo cuartil, Q_2). El cuartil inferior es la mediana de los valores a la izquierda de la mediana general y el cuartil superior, de los valores a la derecha. La fórmula del rango intercuartil es: $Q_3 - Q_1$. **Ejemplos**:

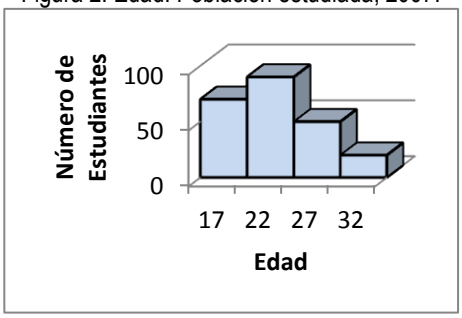
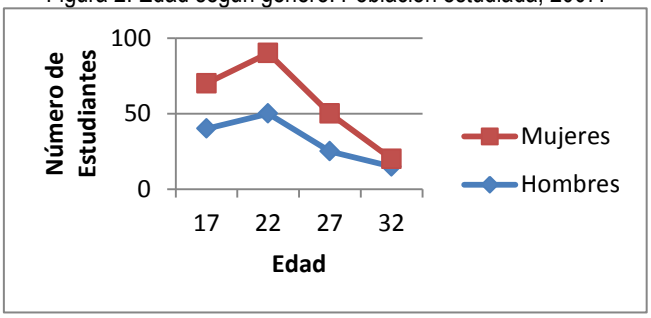
Posición:	*1º*	*2º*	*3º*	*4º*	*5º*	*6º*	*7º*	*8º*	*9º*	*10º*	*11º*	*12º*	*Rango intercuartil:*
Datos:	**72**	**74**	**75**	**77**	**78**	**79**	**82**	**85**	**86**	**90**	**93**	**94**	
Cuartil:			*Q_1 = **76***			*Q_2 = ***80.5***			*Q_3 = **88***				*88 – 76 = **12***

Posición:	*1º*	*2º*	*3º*	*4º*	*5º*	*6º*	*7º*	*8º*	*9º*	*10º*	*11º*	**Rango intercuartil:**
Datos:	**72**	**74**	**75**	**78**	**79**	**82**	**85**	**86**	**90**	**93**	**94**	
Cuartil:			*Q_1 = **75***			*Q_2 = ***82***			*Q_3 = **90***			*90 – 75 = **15***

- **Rango:** es la distancia entre el número mayor y el menor; se utiliza para destacar los valores extremos. En el ejemplo anterior es: 94 – 72 = 22.

FIGURAS

Diagrama de barras	Gráfica circular
Figura 1. Tipos de transporte. Población estudiada, 2007.  	Figura 1. Tipos de transporte. Población estudiada, 2007. 
La altura de cada barra proporcional a la frecuencia de cada opción de respuesta. Se utiliza para: datos cualitativos o numéricos discretos, cuando se quieren comparar las opciones de respuesta entre sí.	El área de cada porción es proporcional al porcentaje de cada opción de respuesta. Se utiliza para: datos cualitativos o numéricos discretos, cuando se quiere comparar cada opción de respuesta con el total.

Histograma	Polígono de frecuencias
Figura 2. Edad. Población estudiada, 2007.  	Figura 2. Edad según género. Población estudiada, 2007. 
El área de cada barra es proporcional a la frecuencia de cada grupo y la etiqueta inferior es el punto central del intervalo. Se utiliza para: datos numéricos continuos agrupados en intervalos.	Se forma uniendo los puntos centrales superiores de cada barra de un histograma. Se utiliza para: comparar el comportamiento de una variable numérica con respecto a otra (cualitativa o numérica), por ejemplo, edad vs. género.