Poblaciones y Muestras

Población:	**Observaciones**	**Ejemplos**
Conjunto de Individuos	personas o no	Los habitantes de Madrid Los coches de Madrid
que comparten una o más características	Lo que comparten es el hecho de tener la característica, no necesariamente sus valores. Si todos los individuos no comparten los valores, a la característica la llamamos **variable**.	Los habitantes de Madrid comparten la característica: Tener un color de ojos, pero cada uno los tiene de un color. El color de Ojos es una variable.
	Si todos los individuos si comparten el valor de la característica, a la característica la llamamos **constante**.	Los coches de Madrid comparten la característica: Tener un número de ruedas. Además todos comparten el valor: 4 ruedas. El número de ruedas de los coches de Madrid es una constante.
objeto de observación y estudio.	Para estudiar la característica necesitamos conocer los distintos valores que puede tomar. A estos valores los llamamos **Modalidades o Clases**. El **proceso de medición** se encarga de definir cómo vamos a apuntar las diferentes modalidades que puede tomar una característica	La variable color de ojos puede presentar las siguientes modalidades: Azul, verde, marrón, …

Cuando lo que estudiamos no es una muestra completa, sino una parte (subconjunto) de la misma, decimos que lo que estamos estudiando es una **Muestra**.

El número de individuos de una muestra o una población lo denominamos: **Tamaño de la muestra o de la población** y lo representaremos por la letra **n**.

Proceso de Medición

Utilizaremos distintas escalas a la hora de anotar los valores (también llamados observaciones y puntuaciones directas) que toman las variables para cada individuo de la población o muestra.

<table>
<tr><th colspan="4">Escalas de Medición</th><th>Variables</th></tr>
<tr><td>Nominal</td><td>Otorga nombres a las diferentes modalidades que puede tomar la variable.</td><td>Permiten ver si dos individuos tienen igual o diferente modalidad.

Permiten clasificar los individuos en función del valor o modalidad.</td><td>Hombre, Mujer

Azul, Marron, Negro, Verde</td><td rowspan="2">Cualitativas

Sexo

Color de ojos

Nivel de satisfacción

Nivel Social</td></tr>
<tr><td>Ordinal</td><td>Otorga nombres a las diferentes modalidades que puede tomar la variable, pero con la peculiaridad de que existe una ordenación entre ellos</td><td>Las modalidades son comparables en grado.

Los valores se pueden ordenar entre si.</td><td>Poco, algo, bastante, mucho

Alto, Medio, Bajo</td></tr>
<tr><td>Intervalo

Tienen un 0 no absoluto, osea, que si la variable toma valor 0 sigue teniendo sentido.</td><td rowspan="2">Otorgan medidas a las diferentes modalidades que puede tomar la variable.

Requieren de una unidad de medida.
Nota:
No tener en cuenta sólo las ud . de medida tradicionales: kg, litro, metro. Por ejemplo:
• La nota de un examen se mide en puntos.
• El número de hijos que tiene una familia se mide en hijos.</td><td rowspan="2">Aportan medidas precisas.

Las medidas se pueden sumar o restar entre sí.</td><td>Temperatura en grados Centígrados

(Una temperatura de 0 Grados centígrados tiene sentido)</td><td rowspan="2">Cuantitativas

• Discretas:
Números enteros
• Continuas:
Números decimales</td></tr>
<tr><td>Razón

Tienen un 0 absoluto, osea, que si la variable toma valor 0 la característica deja de tener sentido.</td><td>Número de alumnos en un curso

(Si hay 0 alumnos en un curso, es que no hay curso… no tiene sentido)</td></tr>
</table>

Estadística

Ciencia dedicada al estudio de **Poblaciones**

- **Descriptiva:**

  Describe una población o una muestra mediante gráficos e índices (como la media, la mediana, la desviación típica…).

- **Inferencial**:

  Trata de inferir (**averiguar**) los índices de una población si se conocen los índices de una muestra de la misma.

Estos índices recibirán nombres diferentes dependiendo de si se obtienen al estudiar una población o una muestra:

- Si los índices se calculan de una **población** se denominan **Parámetros**
- Si los índices se calculan de una **muestra** se denominan **Estadísticos**.

Distribuciones de frecuencias

Para estudiar más fácilmente cómo se distribuyen los valores que toma una variable dentro de una muestra o población, utilizamos las tablas de frecuencias. En ellas se indican para cada modalidad o rango de modalidades que admite la variable los siguientes datos:

Frecuencia Absoluta Se representa por la letra n_i	El **número total de veces que se repite la modalidad** entre los individuos de una muestra o población.	Útil **cuando interesa conocer el número total** de casos que presentan una modalidad. La **suma de todas las frecuencias absolutas** debe ser **igual al tamaño de la población**
Frec. Relativa Se representa por la letra f_r $n_r = n_i / N$	El **número total de veces que se repite la modalidad** entre los individuos de una muestra o población, dividido entre el tamaño de la población o de la muestra.	Útil **cuando interesa comparar** datos de dos poblaciones con tamaños diferentes. La **suma de todas las frecuencias relativas** debe ser **igual a 1.**
Porcentaje Se representa por la el símbolo % $\% = n_r \times 100$	El porcentaje de individuos que presentan una determinada modalidad.	Útil **cuando interesa comparar** datos de dos poblaciones con tamaños diferentes, y cuando interesa solamente saber las **diferencias en proporción** entre las diferentes modalidades. La **suma de todos los porcentajes** debe ser **igual a 100.**
Frec. Acumulada Se representa por la letra N_i	El **número total de veces que se repite la modalidad** , sumado al número total de veces que se repiten las modalidades mostradas más arriba en la tabla.	La frecuencia acumulada de la **última modalidad** de la tabla de frecuencias debe ser **igual al tamaño de la población**.
Frec. Acum. Relativa Se representa por la letra N_r $N_r = N_i / N$	La frecuencia acumulada relativa de la **última modalidad** de la tabla de frecuencias debe ser **igual a 1**.	
% Acumulado Se representa por la el símbolo $\%_A$ $\%_A = N_r \times 100$	El porcentaje acumulado de la **última modalidad** de la tabla de frecuencias debe ser **igual a 100**.	

Medidas de tendencia central

Dan una idea de la **magnitud general** de la variable en una muestra o población.

Se **miden en la misma unidad** que la variable.

Moda M	La moda es el **valor de la variable** (modalidad) **que más se repite**. Puede haber **varias modas**.	Para **cualquier tipo de variable**	Corresponde con el valor de la variable que tiene mayor frecuencia.
Mediana Med	La mediana es el **valor que divide a la población** en 2 partes iguales (50%). Para calcularlo: 1. **Ordenamos** los valores de las variables 2. Buscamos el término central y tomamos el valor de la variable: • **Impar**: término (N+1)/2 • **Par**: Hay 2 términos, el N/2 y el siguiente. Hacemos la media de los valores de la variable para ambos términos Sólo usa los **valores de los términos centrales**.	Requiere **ordenar** los valores: • Sólo para variables **cuasicuantitativas (ordinales) y cuantitativas**. Estadístico Resistente / Robusto: • No se ve afectada por valores extremos / atípicos. • Lo usamos para **distribuciones asimétricas**.	Divide a la superficie bajo la línea de frecuencias en 2 partes iguales. Está entre la media y la moda.
Media $\bar{X}$	Es la media aritmética de **todos los valores** de la variable.	**Suma** valores: • Sólo para variables **cuantitativas**. Estadístico no Resistente / no Robusto: • Si se ve afectada por valores extremos / atípicos • Lo usamos para **distribuciones simétricas**.	Se coloca de forma que deje a la mediana entre la media y la moda.

Medidas de Posición:

Percentiles, deciles, cuartiles

Son **los valores de la variable** que dejan por debajo una **determinada porción de la población**. Se miden en las **mismas unidades que la variable**.

Tipos	**Símbolo K**	**Partes en que dividen a la población: P**	**¿Cúantos son?**	**Tamaño de la población entre 2 partes consecutivas**
Mediana	Med	2	1	50%
Cuartiles	Q	4	3	25%
Deciles	D	10	9	10%
Percentiles	P	100	99	1%

Entre 2 percentiles, cuartiles, etc, se deja una determinada parte de la población, pero no podemos decir nada de las medidas que contienen.

A priori no hay ninguna relación entre las medidas de ellos. Sólo habría relación si la distribución es simétrica. Entonces ocurre por ejemplo que:

$Q_1 = Q_3$

$D_1 = D_q$; $D_3 = D_7$

$P_{20} = P_{80}$; $P_{30} = P_{70}$

Calculo de las medidas de posición:

1. Se ordenan los términos
2. Se calcula el término central:
    a. Si el tamaño de la población es impar, tomamos el valor de la variable para el término central. El término central lo calculamos con la siguiente formula:

$$Término\ para\ \mathrm{K}_j : \frac{j}{P} \cdot\ (N+1)$$

    b. Si el tamaño de la población es par, tomamos el valor de la variable para cada uno de los 2 términos centrales y hacemos su media. El segundo término es el que hay a continuación del primero. El primero lo calculamos con la siguiente formula:

$$Término\ para\ \mathrm{K}_j : \frac{j}{P} \cdot\ (N)$$

Medidas de Variabilidad

Miden como de dispersos están los valores de una distribución: Si difieren mucho o no de la magnitud general de la distribución.

- Valores **altos** de la variabilidad implican:

  Valores de la variable **muy dispersos**: Distribución **heterogénea**.

- Valores **bajos** de la variabilidad implican:

  Valores de la variable **poco dispersos**: Distribución **homogénea**.

Son **siempre iguales o mayores a 0**

No tienen nada que ver con la simetría o asimetría

No robustos / No resistentes	**Robustos / Resistentes**
• Muy afectados por valores extremos / atípicos • Los usamos para distribuciones simétricas. • Se basan en la media. • Sólo variables cuantitativas	• Poco afectados por valores extremos / atípicos • Los usamos para distribuciones asimétricas. • Se basan en la mediana y los cuartiles. • Variables ordinales y cuantitativas
Rango: $R = X_{Max} - X_{Min}$	**Rango Intercuartil:** $R_{IC} = Q_3 - Q_1$
Varianza: $S^2 = \frac{\Sigma(X_i - \bar{X})^2}{n}$	**Rango Semi-Intercuartil:** $R_{SIC} = \frac{Q3 - Q1}{2}$
Desviación Típica: $S = \sqrt{S^2} = \sqrt{\frac{\Sigma(X_i - \bar{X})^2}{n}}$	
Coef. de variación: $C_V = \frac{s}{\bar{X}}$ • Permite comparar la variabilidad de 2 distribuciones con medias y unidades diferentes. • Sólo se puede calcular para distribuciones con media positiva	**Coef. de variación Intercuartílico:** C_{VI} • Permite comparar la variabilidad de 2 distribuciones con medias y unidades diferentes. • Sólo se puede calcular para distribuciones con media positiva

Puntuaciones

Llamamos puntuaciones a cada una de las mediciones que hacemos de una variable para cada uno de los individuos de una población o una muestra.

<table>
<tr><td>Directa
X_i</td><td>La medida que anotamos de una observación.</td><td>Da el valor absoluto de la variable medido para un individuo concreto.</td><td>Puede ser positiva o negativa en función de la variable que estemos midiendo</td></tr>
<tr><td>Diferencial
$P_D = X_i - \bar{X}$</td><td>Su diferencia a la medida (por encima o por debajo).</td><td>Permite saber si la medición de una variable para un individuo está por encima o por debajo de la media.</td><td rowspan="2">Puede ser positiva o negativa en función de si la medida se encuentra por encima o por debajo de la media.

Si alguna de estas puntuaciones vale 0, significa que la medida directa se encuentra es igual a la media.</td></tr>
<tr><td>Típica
$P_T = \frac{Xi - \bar{X}}{S}$</td><td>Indica el número de desviaciones típicas que una medida se acerca o aleja de la media.</td><td>Permite comparar las posiciones relativas de las medidas de:
1 variables en 2 grupos
2 variables en 1 grupo
2 variables en 2 grupos
Cuanto mayor sea, mejor es la puntuación relativa de un individuo dentro de un grupo.</td></tr>
</table>

Cada Puntuación directa tiene una puntuación típica y una diferencial.

2 puntuaciones directas iguales (2 X_i iguales) no implica que sus puntuaciones típicas ni diferenciales sean iguales. Dependerá siempre de la media de su distribución, y en el caso de la puntuación típica también depende de la desviación típica.

Correlaciones

Dan una idea de la **relación que presentan 2 variables entre sí.** Es muy importante tener en cuenta que **correlación no implica causalidad**.

Que dos variables estén correlacionadas significa que existe una **tendencia general** que se manifiesta entre ambas variables, pero que no tiene que cumplirse para la totalidad de los casos (solamente se mostraría para la totalidad de los datos si estuviéramos hablando de una relación perfecta).

Al hacer un estudio de correlación tenemos en cuenta 2 factores:

1. El **sentido** de la correlación:
    - **Directo:**
        - Individuos con valores altos en una variable tienden a tener valores altos en la otra.
        - Individuos con valores bajos en una variable tienden a tener valores bajos en la otra.
    - **Inversa:** Individuos con valores altos en una variable tienden a tener valores bajos en la otra.
2. La **intensidad** de la correlación: Ósea, si la tendencia general se presenta de una forma muy marcada o poco marcada.

Los **diagramas de dispersión** representan los valores (puntuaciones directas) que toman 2 variables medidas en el mismo conjunto de individuos, y son muy útiles para ayudarnos a identificar qué tipo de relación presentan dos variables entre sí (en caso de presentar alguna).

Dos variables pueden presentar una **relación lineal** (donde al dibujar un diagrama de dispersión de ambas variables los datos se distribuyen de forma similar a una línea recta); pero también pueden presentar otro tipo de relación (donde los datos se distribuirían de acuerdo a otro tipo de patrones que no sean una línea recta).

Existen ciertos **índices** que nos ayudan a identificar **relaciones lineales** entre variables. Si éstos índices valen 0, significará que **no hay relación lineal**, pero podría existir otro tipo de relación entre las mismas.

Lo que miden cualquiera de estos índices es lo cerca de una línea recta que están de **media** todos los puntos dibujados en un diagrama de dispersión. No hay que olvidar que hablados de la tendencia general (media), es decir, en unos tramos los puntos pueden estar más cerca y en otros más alejados de la línea recta, incluso podría haber puntos muy alejados, también llamados atípicos. Por eso, **el valor de un índice de correlación medido en un tramo no tiene nada que ver con el valor del índice medido en otro tramo**.

Cualitativas (Nominales):

En cuanto una de las variables es nominal se usan estos índices:

$$x^2 = \sum \frac{(f_e \cdot f_t)}{f_t}$$

$$x^2 \geq 0$$

- f_e es la frecuencia empírica (la que medimos)
- f_t es la frecuencia teórica (calculada como el producto de las frecuencias marginales entre el total de casos)

Este índice no tiene límite superior. No da información ni de intensidad ni de sentido. Usamos mejor el **índice de Contingencia**:

$$C = \sqrt{\frac{x^2}{n+x^2}} \qquad 0 \leq C < 1$$

Su valor depende del número de filas y columnas de la tabla de contingencia. No es comparable entre variables que presenten diferente número de modalidades.

Para tablas cuadradas (con k filas y columnas), su valor máximo está acotado por la expresión:

$$C_{max} = \sqrt{\frac{k-1}{k}}$$

Sólo da información de intensidad. Para obtener el sentido de la correlación hay que comparar f_e *con* f_t.

Cuasi-Cuantitativas (Ordinales):

Se usa cuando analizamos:

- 2 variables ordinales
- 1 ordinal y otra cuantitativa:

Índice de Spearman:

$$R_s = 1 - \frac{6 \leq d_i^2}{\mathrm{n}\,(n^2 1)}$$

$$-1 \leq R_S \leq 1$$

donde: d_i es la distancia (resta) entre las posiciones (orden) que ocupa cada individuo dentro de su grupo para cada variable a correlacionar.

Cuantitativas (Pearson):

Se usa cuando analizamos 2 variables cuantitativas.

Índice de Pearson:

$$R_{xy} = \frac{\sum(\frac{x_i - \bar{x}}{S_x})\,(\frac{y_i - \bar{y}}{S_y})}{n} = \frac{\Sigma(\,x_i - \bar{x}_i)\,(y_i - \bar{y}_i)}{\frac{n}{S\,x\,S_y}} = \frac{S_{xy}}{S_x \cdot S_y}$$

$$-1 \leq R_{xy} \leq 1$$

Para su cálculo se usan las desviaciones típicas y la covarianza:

$$S_{xy} = \frac{\sum(\,x_i - \bar{x})\,(y_i - \bar{y})}{\mathrm{n}} = \frac{\Sigma\,x_i\,y_i}{\mathrm{n}} - \bar{x}\,\bar{y}$$

El signo del índice (tanto del de Spearman, como del de Pearson) da información del sentido de la correlación:

(+) Relación Directa entre las variables
(-) Relación Inversa entre las variables

El valor absoluto del índice (sin signo) da información de la intensidad de la relación:

Valor 0: No hay relación lineal entre las variables
Valor 1: Hay una relación lineal perfecta