UNIDAD 8: ESTADÍSTICA. APUNTES

1. ESTADÍSTICA UNIDIMENSIONAL. FORMULARIO REPASO

Media: $\bar{x} = \frac{\sum x_i \cdot n_i}{n}$

Varianza: $\sigma^2 = \frac{\sum x_i^2 \cdot n_i}{n} - \bar{x}^2$

Desviación típica: $\sigma = \sqrt{\sigma^2}$

Coeficiente de variación: $CV = \frac{\sigma}{\bar{x}}$

2. ESTADÍSTICA BIDIMENSIONAL. DEFINICIÓN

La **estadística bidimensional** es la rama de la estadística que estudia de manera conjunta **dos variables relacionadas entre sí** en un mismo conjunto de datos. A diferencia de la estadística unidimensional, que analiza una sola característica, **la bidimensional permite observar cómo influye una variable sobre la otra y detectar posibles relaciones entre ambas**.

Para ello, se utilizan herramientas como las **tablas de doble entrada**, **los diagramas de dispersión** o **el cálculo de parámetros como la covarianza y el coeficiente de correlación**. Gracias a estos recursos, es posible identificar si existe una relación entre las variables y determinar su intensidad y sentido (positiva, negativa o nula).

Un ejemplo sencillo de **estadística bidimensional** sería estudiar la relación entre las **horas de estudio** y la **nota obtenida en un examen**.

Imagina que recogemos los datos de varios alumnos:

- Alumno A: 2 horas → 2
- Alumno B: 4 horas → 3
- Alumno C: 6 horas → 5
- Alumno D: 8 horas → 7

Aquí tenemos dos variables:

- Variable 1: horas de estudio
- Variable 2: nota del examen

Al analizar estos datos, podemos observar que **a mayor número de horas de estudio, mayor suele ser la nota**, por lo que existe una **relación positiva** entre ambas variables. Este tipo de análisis es precisamente lo que estudia la estadística bidimensional: cómo se relacionan dos variables y qué tipo de relación existe entre ellas.

Vamos a desarrollar el tema en torno a 2 ejemplos, uno con **variables cuantitativas** y otro con **variables cualitativas**.

Ejemplo 1:

Con el fin de hacer un estudio de aceptación sobre dos modelos de impresoras 3D de reciente fabricación, se consideraron el número de ventas efectuado por un determinado distribuidor durante 25 días.

Modelo A: 0 2 2 2 1 3 3 3 3 4 4 2 3 3 3 3 2 3 2 4 2 2 3 3 3

Modelo B: 2 1 2 2 3 1 1 1 2 0 1 1 1 1 1 2 2 1 1 1 2 2 2 2 1

Y vamos a llamar:

- X: número de impresoras del modelo A vendidas en un día. (primera variable estadística)
- Y: número de impresoras del modelo B vendidas en un día. (segunda variable estadística)
- n: número de pares de observaciones.
- x_i: cada dato diferente observado en la muestra de X.
- k: número de valores distintos de X.
- y_j: cada dato diferente observado en la muestra de Y.
- h: número de valores distintos de Y.

Ejemplo 2:

Se estudiaron 600 enfermos con lesiones de hígado mediante un procedimiento gráfico, y se comprobaron los resultados mediante un procedimiento histológico para conocer si había sido el diagnóstico correcto. Los datos que se obtuvieron fueron**:**

De un total de 230 personas con lesión maligna, habían tenido un diagnóstico correcto 210 mientras que 20 habían sido mal diagnosticadas; y de un total de 370 personas con lesión benigna, habían sido bien diagnosticadas 320, resultando mal diagnosticadas las 50 restantes.

3. TABLAS DE CONTINGENCIA

Las **tablas de contingencia** son tablas que se utilizan para organizar y resumir datos de dos variables, **mostrando las frecuencias con las que aparecen las distintas combinaciones de sus valores**.
Así, la tabla de contingencia del ejemplo 1 sería:

		Variable X					
		0	1	2	3	4	Total
Variable Y	0	0	0	0	0	1	1
	1	0	0	3	8	2	13
	2	1	0	5	4	0	10
	3	0	1	0	0	0	1
	Total	1	1	8	12	3	25

Y en el caso de la tabla de contingencia del ejemplo 2:

	Diagnóstico correcto (C)	Diagnóstico incorrecto (I)	Total
Lesión maligna (M)	210	20	230
Lesión benigna (B)	320	50	370
Total	530	70	600

4. DISTRIBUCIÓN DE FRECUENCIAS CONJUNTAS.

Llamamos así a una **tabla de doble entrada** donde se representan en la primera columna los diferentes valores observados para la variable X ordenados de menor a mayor y en la primera fila los diferentes valores observados para la variable Y (también ordenados de menor a mayor), y en el centro de la tabla sus correspondientes **frecuencias conjuntas**, tanto **absolutas** como **relativas**. (Debes repasar el concepto de frecuencia absoluta y relativa de cursos anteriores)
Así, la tabla de frecuencias conjuntas del ejemplo 1 sería:

x_i / y_j	0	1	2	3
0	0/0	0/0	1/0.04	0/0
1	0/0	0/0	0/0	1/0.04
2	0/0	3/0.12	5/0.20	0/0
3	0/0	8/0.32	4/0.16	0/0
4	1/0.04	2/0.08	0/0	0/0

En el caso del ejemplo 2, debes saber que **para las variables bidimensionales cualitativas se le llama tabla de frecuencias conjunta a la propia tabla de contingencias.**

5. DISTRIBUCIÓN DE FRECUENCIAS MARGINALES

Para distinguir las frecuencias de cada variable al estudiarlas aisladamente llamaremos **frecuencias marginales a las de cada variable por separado**. De esta forma tendríamos dos distribuciones unidimensionales a partir de las conjuntas.

5.1- Frecuencia absoluta marginal

Para la X (x_i) sería el **número de veces que se repite el valor x_i sin tener en cuenta los valores de Y**, la representamos por n_i. Para la Y (y_j) sería el **número de veces que se repite el valor y_j sin tener en cuenta los valores de la X**, la representamos por n_j.

5.2- Frecuencia relativa marginal

A partir de las anteriores, y del mismo modo, se construirán estas **frecuencias relativas marginales f_i** y **f_j** .La distribución de frecuencias marginales puede colocarse en una tabla separadamente. Pero si deseamos tener toda la información en una misma tabla lo que se suele hacer es colocar:

- ✔ En la última columna de la tabla conjunta, las frecuencias marginales de X, es decir, n_i, añadiendo tantas columnas como otros tipos de frecuencias marginales se desee añadir.
- ✔ En la última fila de la tabla conjunta, las frecuencias marginales de Y, es decir, n_j añadiendo tantas filas como otros tipos de frecuencias marginales se desee añadir.

Así, en el caso del ejemplo 1, podríamos completar la tabla de frecuencias conjuntas de la siguiente manera:

x_i / y_j	0	1	2	3	n_i	f_i
0	0/0	0/0	1/0.04	0/0	1	0.04
1	0/0	0/0	0/0	1/0.04	1	0.04
2	0/0	3/0.12	5/0.20	0/0	8	0.32
3	0/0	8/0.32	4/0.16	0/0	12	0.48
4	1/0.04	2/0.08	0/0	0/0	3	0.12
n_j	1	13	10	1	25	
f_j	0.04	0.52	0.04	0.04		1

6. DISTRIBUCIÓN DE FRECUENCIAS CONDICIONADAS

La **frecuencias condicionadas** miden valores de una variable condicionados a un cierto valor de la otra variable.

6.1- Frecuencia absoluta condicionada

La **frecuencia absoluta condicionada** es el **número de veces que aparece un valor de una variable para un cierto valor fijado de la otra variable.** Es decir, indica cuántos datos cumplen una condición concreta, considerando que ya estamos fijándonos en una categoría específica de la otra variable.

Para el caso de la variable X, denotaremos la frecuencia absoluta de x_i condicionada al valor y_j como $n_{i(j)}$.

Para el caso de la variable Y, denotaremos la frecuencia absoluta de y_j condicionada al valor x_i como $n_{(i)j}$.

6.2- Frecuencia relativa condicionada

Análogamente a la frecuencia absoluta condicionada, podemos definir las **frecuencias relativas condicionadas** de la siguiente forma:

Frecuencia relativa condicionada para X dado que Y = y_i es: $f_i(j)=\frac{n_i(j)\,j}{n_j}$

Frecuencia relativa condicionada para Y dado que X= x_i es: $f_{(i)}j=\frac{n_{(i)}j}{n_i}$

Así, para el ejemplo 1, hagamos la tabla de frecuencias condicionadas de X para Y= 1

x_i	$n_{i(1)}$	$f_{i(1)}$
0	0	0/13=0
1	0	0/13=0
2	3	3/13=0,23
3	8	8/13=0,62
4	2	2/13=0,15

7. INDEPENDENCIA ESTADÍSTICA

Dos variables **X e Y** se dice que son **independientes estadísticamente cuando la frecuencia relativa conjunta es igual al producto de las frecuencias relativas marginales, es decir, para todo i, j:**

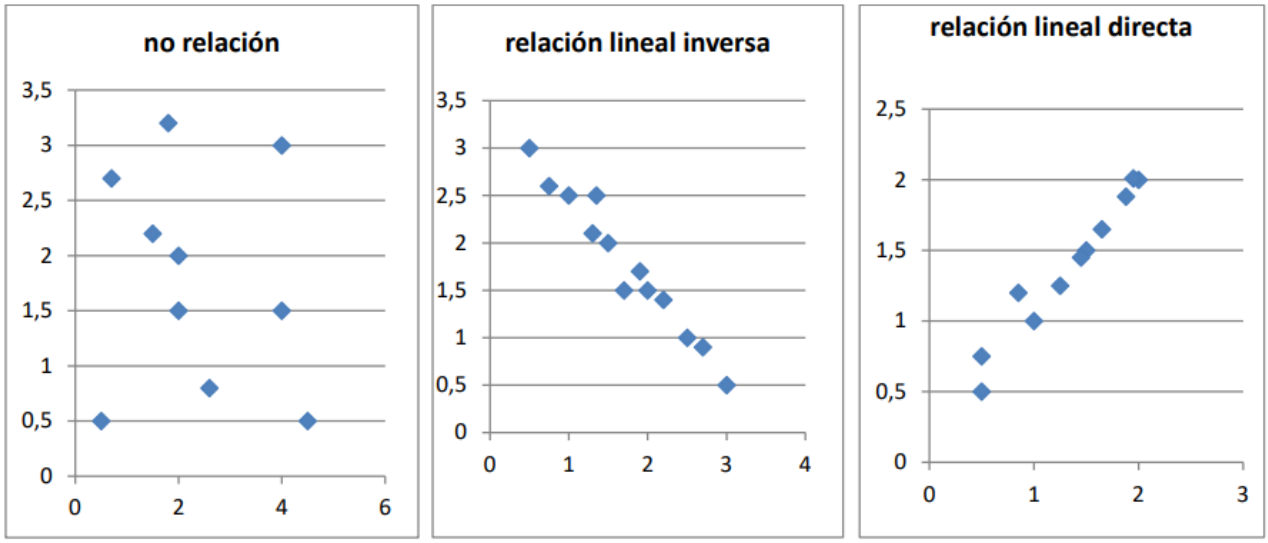
$$f_{ij}=\frac{n_{ij}}{n}=f_i\cdot f_j=\frac{n_i}{n}\cdot\frac{n_j}{n}$$

Dos variables **X e Y** se dicen que son **independientes estadísticamente cuando todas las frecuencias relativas condicionadas son iguales a sus correspondientes frecuencias marginales, es decir:**

$f_{i(j)} = f_i$ para todo j y $f_{(i)j} = f_j$ para todo i.

8. DIAGRAMA DE DISPERSIÓN. NUBE DE PUNTOS

Se obtiene representando cada par observado (x_i, y_j), como un punto del plano cartesiano. Se utiliza con los datos sin agrupar y sobre todo para variables continuas. Si los datos están agrupados se toman las marcas de clase. Es muy útil porque nos permite ver visualmente la relación entre las dos variables.



9. IDEA DE CORRELACIÓN. COVARIANZA

Al analizar dos variables cuantitativas de forma conjunta, el objetivo que se pretende es, por lo general, determinar si existe o no algún tipo de **variación conjunta** o **covarianza** entre ellas: si una variable aumenta, la otra también o lo contrario.
La cantidad se denomina **covarianza S_{xy}** y tiene la siguiente expresión:

$$S_{xy} = \frac{\sum_i \sum_j (x_i - \bar{x}) \cdot (y_i - \bar{y}) \cdot n_{ij}}{n} = \frac{\sum_i \sum_j x_i \cdot y_i \cdot n_{ij}}{n} - \bar{x} \cdot \bar{y}$$

✔ Cuando el r**esultado es positivo**, hay una tendencia a que a mayores observaciones de X correspondan mayores observaciones de Y. Por ejemplo: A mayor cantidad de agua de lluvia en un año, suele corresponder una mejor cosecha.

✔ Cuando el **resultado es negativo**, la tendencia resulta contraria; es decir a mayores valores de la variable X solemos encontrar menores valores de la variable Y. Por ejemplo: A mayor renta per cápita en los países suele encontrarse una menor mortalidad infantil.

10- COEFICIENTE DE CORRELACIÓN LINEAL

El valor de la covarianza dependerá de los valores de las variables, por tanto, de sus unidades. Para poder eliminar las unidades y tener **una medida adimensional** utilizamos el **coeficiente de correlación, r_{xy}** :

$$r_{xy} = \frac{S_{xy}}{s_x \cdot s_y}$$

Citamos las siguientes propiedades:

- Es un coeficiente adimensional.
- Toma valores entre −1 y 1.
- Si hay relación lineal positiva el valor será positivo y próximo a 1.
- Si hay relación lineal negativa el valor será negativo y próximo a −1.
- Si no hay relación el valor se aproxima a cero.
- Si X e Y son independiente el valor del coeficiente es cero. Pero no al contrario. Puede ocurrir que el coeficiente de correlación valga cero y las variables sean dependientes.

11- RECTA DE REGRESIÓN LINEAL. REGRESIÓN CUADRÁTICA

El **diagrama de dispersión o nube de puntos** nos permitía visualizar la relación entre dos variables X e Y. Al representar el diagrama de dispersión podemos encontrar las siguientes situaciones:

- Distribuciones estadísticas para las que **la nube de puntos se dispone de tal forma que existe una función matemática cuyos puntos son una parte de su representación gráfica**.
- Sin coincidir sus puntos con los de una gráfica de una función matemática, **se aproximan a ella con mayor o menor intensidad**.
- La nube de puntos presenta un aspecto tal que **no existe concentración de puntos hacia ninguna grafica matemática**, distribuyéndose de una forma uniforme en una región del plano.

En el primer caso se dice que existe una **dependencia funcional o exacta** entre las variables X e Y, es decir existe una función matemática tal que y = f(x). En el segundo caso se dice que existe una **dependencia estadística o aproximad**a entre las dos variables. Y en el último caso decimos que **las variables son independiente**s.

Es el segundo caso del que se ocupa la **teoría de regresión**.

Las técnicas de regresión tienen por objeto modelar, es decir, encontrar una función que **aproxime lo máximo posible la relación de dependencia estadística entre variables y predecir los valores de una de ellas**: Y (variable dependiente o explicada) **a partir de los valores de la otra** (u otras): X (variable independiente o explicativa).

Llamamos **regresión Y sobre X** a la función que explica la variable Y (dependiente) para cada valor de la X (independiente). Los dos tipos de funciones más usados son **la regresión lineal**, si la función de regresión es una recta, y **la regresión cuadrática**, si la función es una parábola.

La **recta de regresión es una función lineal** porque el modelo de función de regresión seleccionado es una recta.

Recta de regresión Y sobre X es $y = a + bx$ donde $a = \bar{y} - b\bar{x}$ y $b = \frac{S_{xy}}{{s_x}^2}$.

Recta de regresión de X sobre Y es $x = a' + b'y$ donde $a' = \bar{x} - b'\bar{y}$ y $b' = \frac{S_{xy}}{{s_y}^2}$.

Los valores de b y b' son los correspondientes **coeficientes de regresión** para cada una de las rectas. Hay que tener en cuenta que **la recta de regresión de x sobre y no se obtiene despejando x de la recta de regresión de y sobre x**.

Si usamos como función para el modelo una parábola tenemos una **regresión cuadrática**, buscando los valores de los coeficientes de una parábola que mejor se ajusten a unos datos.

En ambos casos se usa el "**Método de Mínimos cuadrados**".

Pero las herramientas informáticas, como las hojas de cálculo, te pueden ayudar. Basta con que escribas los datos en una hoja de cálculo, le pidas a la hoja de cálculo que dibuje una nube de puntos, y luego, puedes ajustar esa nube con una recta, una parábola….

12- CORRELACIÓN Y CAUSALIDAD

La **correlación** significa que dos variables están relacionadas: cuando una cambia, la otra también tiende a cambiar mientras que la **causalidad** implica algo más fuerte: que una variable **provoca** directamente el cambio en la otra.

⚠ Importante: que haya correlación **no significa necesariamente** que haya causalidad.

Un ejemplo sencillo:

En verano aumenta el consumo de helados y también los casos de insolación.
👉 Ambas cosas están **correlacionadas** (suben a la vez).

Pero comer helados **no causa** la insolación.
👉 La **causa real** es el calor: hace que la gente coma más helados y también que haya más insolaciones.