TEMA 1 → CONCEPTOS BÁSICOS Y ORGANIZACIÓN DE DATOS

OBJETIVOS DE APRENDIZAJE:

- Diferenciar y manejar los conceptos básicos, la nomenclatura y las definiciones centrales de la estadística, a fin de poder aplicarlos en el estudio formal de la materia.
- Entender la importancia de la medición en el ámbito psicológico, distinguiendo entre las distintas escalas de medida (nominal, ordinal, de intervalo y de razón), y conociendo las relaciones que pueden establecerse en cada una de ellas.
- Manejar con soltura las distintas denominaciones y clasificaciones de las variables.
- Saber elaborar, a partir de un conjunto de datos, una distribución de frecuencias, adquiriendo y desarrollando la capacidad para recopilar, organizar, presentar, e interpretar datos numéricos.

INTRODUCCIÓN:

La Estadística es una ciencia que recoge, ordena y analiza los datos de una muestra, extraída de cierta población, y que, a partir de esa muestra, valiéndose del cálculo de probabilidades, se encarga de hacer inferencias acerca de la población (Amón 1984).

→ **El Análisis de Datos en Psicología** es una herramienta metodológica necesaria, de carácter fundamentalmente estadístico, para la investigación en Psicología.

1.- EL MÉTODO CIENTÍFICO ➔ Se caracteriza por ser **Sistemático** (proceso que dispone de unas etapas claramente definidas) y **Replicable** (Los datos pueden ser refutados o replicados por cualquier investigador que lo desee).

Simplificando, una investigación se afronta mediante las siguientes etapas:

1. Definición del **problema**.
2. Deducción de **hipótesis** contrastables.
3. Establecimiento de un **procedimiento** de recogida de datos.
4. **Análisis** de los resultados obtenidos.
5. **Discusión** de dichos resultados y búsqueda de conclusiones.
6. Elaboración de un **informe** de investigación.

El Análisis de Datos se ocupa de la fase 4.

2.- CONCEPTOS BÁSICOS ➔

➔ **POBLACIÓN (Universo)**: Conjunto de todos los elementos que cumplen una determinada característica objeto de estudio. Puede ser finito (municipios con más de 5000 habitantes, accidentes de tráfico en 1998,…) o infinito (números naturales, lanzamiento reiterado de una moneda, etc.). Los índices que describen la Población se denominan **Parámetros** y se simbolizan con letras griegas:

(Media = μ // Varianza = σ^2 // Proporción = π // etc.).

➔ **MUESTRA**: Subconjunto de elementos representativos de la población. Todos los elementos de la población han de tener la misma probabilidad de ser elegidos (Muestreo probabilístico o aleatorio). Los índices que describen la muestra se denominan **Estadísticos** y se simbolizan mediante letras del alfabeto latino:

(Media $\bar{X}$, $\bar{Y}$,… // Varianza = S^2_x // Proporción = **p**, etc.).

➔ **CARACTERÍSTICA:** Propiedad de los individuos que componen una población.

➔ **MODALIDAD:** Diferentes variedades que presentan las características que se estudian. Cuando la escala es Nominal se denominan **categorías** (pueden tomar dos valores – dicotómicas - o más de dos valores - politómicas). Cuando son escalas superiores se denominan **valor o puntuación**. A nivel ordinal se entienden como rangos u órdenes.
Ejemplo: **Mamífero** = Característica; **Sexo de los mamíferos** = Modalidad **(**al tratarse de dos categorías o modalidades, macho y hembra, se denomina **dicotómica).**

3.- MEDICIÓN Y ESCALAS DE MEDIDA ➔

MEDIR: Proceso mediante el que se asignan números a objetos o características según determinadas reglas.

Las relaciones entre los números son válidas en el mundo aritmético (mundo ideal) y las relaciones entre las modalidades de las características son verificables en el mundo empírico (mundo real). Medir supone relacionar el mundo ideal de los números y el real de los objetos, de forma que las relaciones que se dan en el mundo numérico preserven exactamente las relaciones que se observan en el mundo empírico (sólo son válidas las relaciones numéricas que pueden ser verificables empíricamente, Stevens 1946).

ESCALAS DE MEDIDA → Procedimiento por el cual se relacionan de manera biunívoca un conjunto de modalidades (distintas) con un conjunto de números (distintos). A cada modalidad le corresponde un solo número y a cada número una sola modalidad.

Tipo de escala	Conclusiones a cerca de …	Relaciones Válidas	Ejemplos
Nominal	Relaciones del tipo "igual que" o "distinto que"	"Igual que" / "distinto que"	Sexo, raza, estado civil, diagnóstico clínico
Ordinal	Relaciones del tipo "mayor que", "menor que", "igual que"	Establece un orden "Mayor que" / "Igual que"	Dureza minerales, status socioeconómico, ubicación ideológica
Intervalo	Igualdad de diferencias	Igualdad / desigualdad de diferencias	Calendario, temperatura, inteligencia
Razón	Igualdad o desigualdad de razones	Igualdad / desigualdad de razones	Longitud, masa, tiempo

➔ Una escala tiene las propiedades de la que está por debajo, pero no viceversa.
➔ La escala de Razón es la de mayor nivel, tiene las propiedades de todas las anteriores.

ESCALA DE RAZÓN ↓	➔ Supone utilizar una unidad de medida; el punto cero es absoluto (carencia absoluta de la característica medida). ➔ Pueden compararse dos medidas mediante un cociente.
ESCALA DE INTERVALO ↓	➔ Supone utilizar una unidad de medida; por tanto, permite establecer la igualdad o desigualdad de las diferencias entre las magnitudes de los objetos medidos (cuantifica las diferencias entre objetos o cualidades). ➔ El cero es arbitrario y no significa carencia absoluta de la característica medida.
ESCALA ORDINAL ↓	➔ Ordena una característica según el grado que presentan los elementos (orden de magnitud, creciente o decreciente, en que los objetos presentan determinada característica). ➔ Los valores numéricos carecen de propiedades matemáticas, son meros símbolos (no pueden realizarse operaciones).
ESCALA NOMINAL	➔ Es una manera de Clasificar o Etiquetar las diferentes modalidades que presentan determinadas características (asignar nombres a objetos o cualidades) ➔ Los valores numéricos carecen de propiedades matemáticas, son meros símbolos (no pueden realizarse operaciones). ➔ La única comprobación empírica que puede hacerse es la Igualdad o Desigualdad.

EJEMPLOS DEL NIVEL DE MEDIDA:

Escala nominal o categórica ➔ Extraemos una muestra de 100 enfermos de un centro psiquiátrico:

VARIABLE ➔ Padecer una de las cinco enfermedades (pertenecer a una de las cinco categorías). No tiene sentido realizar operaciones (**variable cualitativa o categórica**).

CATEGORÍAS	ENFERMEDAD	NUMERO DE ENFERMOS
5	DEPRESIÓN	4
4	PSICOSIS	8
3	ESQUIZOFRENIA	17
2	ALZHEIMER	67
1	DEMENCIA SENIL	4
TOTALES		**100**

Escala ordinal ➔ Extraemos una muestra de 1000 europeos y les pedimos su opinión sobre el cierre de los centros psiquiátricos:

VARIABLE ➔ Grado de acuerdo sobre el cierre de los centros psiquiátricos. Variable cualitativa con cinco categorías ordenadas (**variable cuasi - cuantitativa**).

ORDEN	OPINIÓN	FRECUENCIAS	FRECUENCIAS ACUMULADAS	PROPORCIONES ACUMULADAS
5	TOTALMENTE DE ACUERDO	498	1000	1000

4	DE ACUERDO	300	502	0,502
3	POCO DE ACUERDO	95	202	0,202
2	EN DESACUERDO	47	107	0,107
1	TOTALMENTE EN DESACUERDO	60	60	0,060
TOTAL		**1000**		

Escalas de intervalo vs razón ➔Extraemos una muestra de 200 españoles y les hemos pedido valorar en una escala de 0 a 10 su grado de preocupación por la pobreza en el mundo.

VARIABLE ➔ Grado de preocupación por la pobreza en el mundo (**Variable cuantitativa**). 0 indica poca preocupación y 10 una preocupación máxima.

PREOCUPACIÓN	FRECUENCIAS	FRECUENCIAS ACUMULADAS	PROPORCIONES ACUMULADAS
9 – 10	3	200	1,00
6 – 8	80	197	0,985
3 – 5	102	117	0,585
0 - 2	15	15	0,075
TOTALES ➔	200		

4.- VARIABLES: CLASIFICACIÓN Y NOTACIÓN ➔

Variable: representación numérica de una característica que representa más de una modalidad (valor) de un conjunto determinado (refleja el hecho de que los objetos varían respecto a esa característica)
Constante: característica con una sola modalidad.

Tipos de variables en relación con las escalas de medida:

ESCALAS	VARIABLES
NOMINAL	Cualitativa → Dicotómica y Politómica
ORDINAL	Cuasi-cuantitativa
DE INTERVALO / DE RAZÓN	Cuantitativas → Discreta y Continua

Cualitativa → Característica que sólo puede considerarse a nivel Nominal (Variables Nominales: Dicotómicas o Politómicas). Cualquier variable puede convertirse en Nominal (si se divide en dos categorías se dice que se dicotomiza y cuando se hace en varias categorías se politomiza). Ejemplos: el sexo (dicotómica) y la nacionalidad (politómica)
Cuasi-cuantitativa → Característica que puede ser considerada, al menos, a nivel ordinal (Variables Ordinales).
Cuantitativa → Característica que puede ser considerada, al menos, a nivel de Intervalo (Variables De Intervalo y De Razón).
Discreta → Cuando los valores que puede asumir la variable son enteros, se obtienen normalmente por conteo, (Número de hijos...).
Continua → Cuando la variable puede tomar cualquier valor entre dos números dados (valores decimales), son el resultado de medir (metro, balanza, cronómetro,...). Normalmente, y dado que los instrumentos de precisión utilizados para medir no son absolutamente precisos, las variables

continuas pueden considerase a efectos estadísticos como variables discretas. Con este criterio, se consideran dos tipos de valores para las variables continuas:

Valor informado, virtual o aparente (valor que aporta el instrumento)
Límites exactos, valor real (valores que acotan el intervalo → ± 0,5 del valor informado)

Ejemplo → Medimos una distancia y obtenemos 5 cm (valor informado o aparente); sus límites exactos se expresan a través del intervalo → 4,5 y 5,5 (intervalo valor real).

Notación de variables: Para simbolizar las variables → Letras del Alfabeto Latino, el subíndice (i) las distingue de las constantes y n es el número de elementos que componen la muestra: (X_i para i = 5, 6, 7..., n → $X_1 = 5$; $X_2 = 6$...)

RESUMEN VARIABLES	Definición	Es una representación numérica de una característica. Si la característica presenta una sola modalidad, se trata de una constante.		
	Tipos	Cualitativa	Nominal	**Dicotómica**: sólo dos categorías
				Politómica: tres o más categorías
		Cuasi cuantitativa	Ordinal	
		Cuantitativa	De intervalo De razón	**Discreta**: número limitado de valores (el valor se obtiene por conteo y es un numero natural: número de...)
				Continua: puede tomar cualquier valor entre dos números dados, siempre hay un valor intermedio o infinitésimo (para obtener los valores se precisa un instrumento físico de medida y son números reales). Pueden considerarse discretas a efectos prácticos, debido a la imprecisión de los instrumentos de medida: el valor informado indica que el valor real está entre dos límites exactos.
	Notación	Letras mayúsculas del alfabeto latino, afectadas por un subíndice genérico (i): X_i, donde *i* indica el lugar que ocupa un valor concreto que adopta la variable, y *n* simboliza el último valor de la serie: X_i para $i = 1, 2, 3..., n$		

Tipos de variables en relación con la metodología científica:

V. INDEPENDIENTE	Cualquier suceso que sospechamos puede ser la causa de otro y estamos interesados en estudiar.
V. DEPENDIENTE	Efectos de la Var. Independiente
V. EXTRAÑA	Variables que pueden influir sobre la Var. Dependiente y no nos interesa estudiar sus efectos.

Ejemplo → Estudiar cómo influye el tipo de método de enseñanza (**V. Independiente**: Expositivo, Virtual, etc) en el aprendizaje de la asignatura (**V. Dependiente**). No tenemos en cuenta la edad, los estudios previos, etc (**V. Extrañas**)

5.- DISTRIBUCIÓN DE FRECUENCIAS ➔ En cualquier investigación la información recogida (datos) sobre las variables objeto de estudio se organiza en forma de tabla. La forma habitual de organización de datos es la Distribución de Frecuencias (permite agrupar de manera ordenada una gran cantidad de datos).

Distribución de frecuencias: representación de la relación entre un conjunto de medidas exhaustivas y mutuamente excluyentes y la frecuencia de cada una de ellas (Hays, 1988). Sus principales funciones son: facilitar información para obtener los estadísticos muestrales y ofrecer información para realizar representaciones gráficas.

ELEMENTOS FUNDAMENTALES DE LA DISTRIBUCIÓN DE FRECUENCIAS
Frecuencia Absoluta (n_i) → Número de veces que se repite cada uno de los valores de la variable. La suma de todas la frecuencias absolutas representa el total de la muestra **($\sum n_i = n$).**
Frecuencia Relativa o Proporción (p_i) → Cociente entre la frecuencia absoluta de cada clase y el número total de observaciones **($p_i = n_i / n$).**
Porcentaje (P_i) → Frecuencia relativa multiplicada por cien **($P_i = p_i \cdot 100$)**
Frecuencia Absoluta Acumulada (n_a) → Número de veces que se repite cada modalidad o cualquiera de las modalidades inferiores.
Frecuencia Relativa Acumulada o Proporción Acumulada (p_a) → Cociente entre la frecuencia absoluta acumulada de cada clase y el total de las observaciones **($p_a = n_a / n$)**
Porcentaje Acumulado (P_a) → Frecuencia relativa acumulada multiplicada por cien **($P_a = p_a \cdot 100$)**

Distribución de frecuencias (Variables de tipo Cualitativo, Nominal o Categórico):

VARIABLE: Estado Civil (Nominal o Categórica / Politómica)			
Estado Civil: Categorías o Modalidades de la variable	**Frecuencia n_i**	**Proporción ($p_i = n_i / n$)**	**Porcentaje ($P_i = p_i \cdot 100$)**
Casado	794	0,53	53
Viudo	165	0,11	11
Divorciado	213	0,142	14,2
Separado	40	0,027	2,7
Soltero	286	0,191	19,1
Totales →	n = 1.498	1	100

→ Con los datos de la tabla obtenemos información de cómo se distribuye la muestra en los 1.498 casos estudiados, si el muestreo se ha efectuado de manera adecuada (aleatoriamente), la información se podrá generalizar a la Población de Origen. Con las variables de tipo **ordinal** se procede de forma similar, teniendo en cuenta que las **distintas modalidades deben figurar en orden**.

→ Para las variables Nominales no tiene sentido obtener valores acumulados; ya que las únicas relaciones entre las categorías (modalidades) son de tipo igualdad o desigualdad. En las ordinales si es posible averiguarlas; ya que los valores numéricos representan el grado de magnitud de la característica.

Distribución de frecuencias (Variables de tipo Cuantitativo: Intervalo o Razón):

→ Cuando la variable es cuantitativa (continua) conviene agrupar los datos en intervalos, aplicando los conceptos generales que se recogen a continuación:

CONCEPTOS BÁSICOS → ESCALAS DE INTERVALO
Intervalo → Cada uno de los grupos de valores que ocupan una fila en una Distribución de Frecuencias.
Límites Aparentes (virtuales o informados) → Los valores mayor y menor de cada intervalo, teniendo

en cuenta el nivel de precisión del instrumento de medida.
Límites Reales o Exactos → Valores máximo y mínimo que tendría cada intervalo si el instrumento de medida tuviera una precisión perfecta.
Punto Medio del Intervalo → Semisuma de los límites exactos o de los límites aparentes.
Amplitud del Intervalo →Diferencia entre el límite exacto superior y el límite exacto inferior; o entre el límite aparente superior y el límite aparente inferior más la unidad de medida.

PROCEDIMIENTO PARA AGRUPAR DATOS EN INTERVALOS
AMPLITUD TOTAL (RANGO) → R = (Puntuación más alta – Puntuación más baja) + 1
NÚMERO DE INTERVALOS → Nunca debe exceder de $\sqrt{n}$
AMPLITUD DE LOS INTERVALOS (i) → (i) = Rango / nº de intervalos
ESTABLECER LOS LÍMITES DE LOS INTERVALOS → Superior (LES) e Inferior (LEI). Se trata de los límite exactos (± 0,5)
RECUENTO DE FRECUENCIAS → ($\sum n_i = n$) Número total de observaciones.

Ejemplo → Construir la tabla de distribución de frecuencias con los siguientes valores experimentales:

38	47	35	19
32	50	35	64
49	61	45	53
44	42	76	35
47	45	28	56
52	40	50	36
68	26	25	44
73	63	40	57
54	65	48	58
46	46	38	42

Número total de observaciones → n = 40
Rango → [76 – 19] + 1 = 58
Nº de Intervalos → Aprox. $\sqrt{40} = 6'3 \rightarrow 6$ ó 7
Amplitud de los intervalos → 58 / 6 = 9´66 → 9

Construimos los intervalos y hacemos recuento del número de frecuencias absolutas que le corresponde a cada uno de ellos:

LÍMITES VIRTUALES	LÍMITES REALES	PUNTO MEDIO INTERVALO	FRECUENCIA ABSOLUTA (n_I)	FRECUENCIA RELATIVA (p_I)
72-80	71´5-80´5	(72 + 80) / 2 = 76	// 2	2 / 40 = 0´05
63-71	62´5-71´5	67	//// 4	4 / 40 = 0´1
54-62	53´5-62´5	58	///// 5	5 / 40 = 0´125
45-53	44´5-53´5	49	///// ///// // 12	12 / 40 = 0´3
36-44	35´5-43´5	40	///// //// 9	9 / 40 = 0´225
27-35	26´5-35´5	31	///// 5	5 / 40 = 0´125
18-26	17´5-26´5	22	/// 3	3 / 40 = 0´075
		MARCA DE CLASE	n = 40	1

LÍMITES VIRTUALES	PORCENTAJE ($P_i = p_i \cdot 100$)	F. ABSOLUTA ACUMULADA (n_a)	F. RELATIVA ACUMULADA ($p_a = n_a / n$)	PORCENTAJE ACUMULADO ($P_a = p_a \cdot 100$)
72-80	0´05 x 100 = 5	40	1	100
63-71	10	38	0´95	95
54-62	12´5	34	0´85	85

45-53	30	29	0´725	72´5
36-44	22´5	17	0´425	42,5
27-35	12´5	8	0´2	20
18-26	7´5	3	3 / 40 = 0´075	0´075 x 100 = 7´5

Aclaración: La distribución de frecuencias de los valores debe realizarse teniendo en cuenta los límites exactos (reales) de medida, no los valores informados (virtuales). Para obtener los valores exactos, o límites, entre los que se encuentra el valor real de la medida, se aplica la formula:
Límite exacto = (Valor informado +/- 0,5) · I // Siendo I = Unidad del instrumento de medida. "I" está en relación con los decimales de cada valor [ejemplo: 1.55 dos decimales (55) en consecuencia "I" = 0.005 // 1.7 un decimal (7) en consecuencia "I" = 0.05 // 14 no tiene, en consecuencia "I" = 0.5].

Valor virtual	Operación	Limites exactos
1.55 – 1.62	1.55-0.005 - 1.62+0.005	1.545 - 1.625
14 - 17	14-0.5 - 17+0.5	13.5 - 17.5
1.7 - 1.9	1.7-0.05 - 1.9+0.05	1.65 - 1.95

Intervalo abierto: Intervalo sin límite superior. En nuestro caso sería: 72 ó más.

6.- REPRESENTACIONES GRÁFICAS:

→ Permiten visualizar un conjunto de datos para formar una primera impresión, de carácter global, sobre la distribución. El tipo de representación gráfica depende de la variable objeto de estudio.

REPRESENTACIÓN GRÁFICA	TIPO DE VARIABLES
	UNA VARIABLE
DIAGRAMA DE BARRAS	NOMINALES, ORDINALES Y CUANTITATIVAS DISCRETAS
D. BARRAS ACUMULATIVO	ORDINALES Y CUANTITATIVAS DISCRETAS
DIAGRAMA DE SECTORES	CUALITATIVAS Y CUASICUANTITATIVAS (ORDINALES)
PICTOGRAMA	CUALITATIVAS
HISTOGRAMA	CUANTITATIVAS CONTINUAS (DATOS EN INTERVALO)
POLÍGONO DE FRECUENCIAS	CUANTITATIVAS (DISCRETAS Y CONTINUAS)
	DOS VARIABLES
DIAGRAMA DE BARRAS CONJUNTO	AL MENOS UNA DE LAS DOS VARIABLES CUALITATIVA
DIAGRAMA DE DISPERSIÓN (NUBE DE PUNTOS)	DOS VARIABLES CUANTITATIVAS

PARA UNA VARIABLE:

Diagrama de barras: Sirve para representar distribuciones de frecuencias (absolutas y acumuladas) de variables **nominales, ordinales y cuantitativas discretas**. Las frecuencias acumuladas representadas gráficamente, pueden ser relativas o absolutas. La altura de la línea de la frecuencia absoluta acumulada del valor máximo es el total de los casos observados. En la representación de frecuencias relativas acumuladas, la altura de la línea del valor máximo es 1 (o 100 en términos de porcentaje). En el eje de abscisas se colocan los valores de la variable o las

distintas categorías, en orden ascendente, y en el de ordenadas las frecuencias, absolutas o relativas.

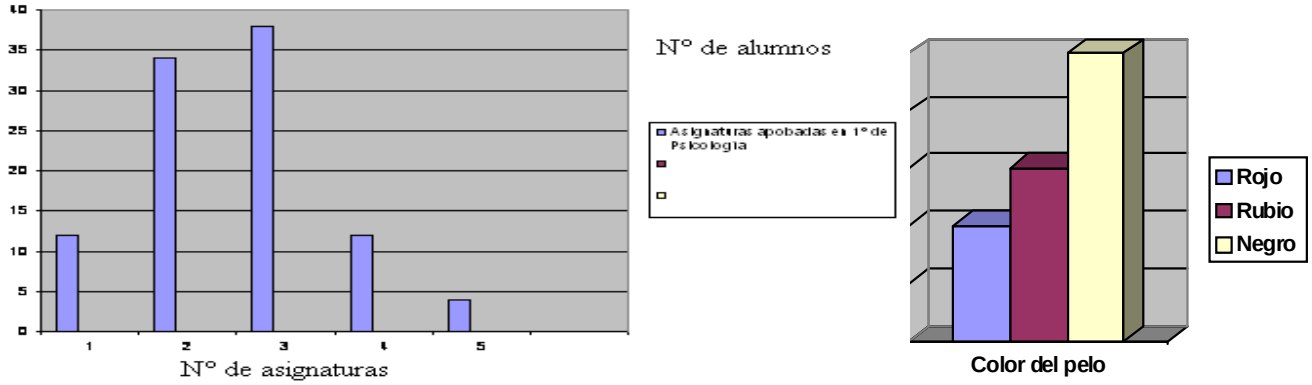
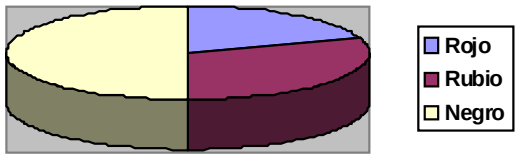


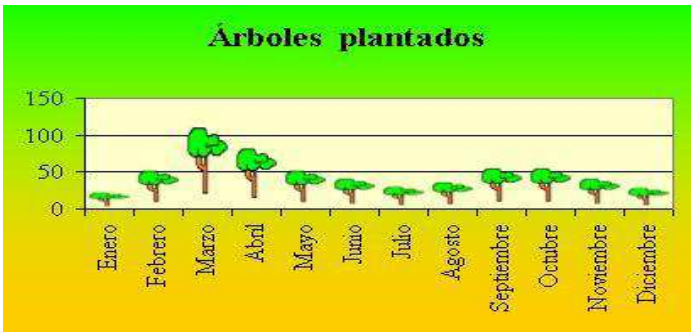
Diagrama de sectores: Se representan de forma circular las variables cualitativas de cada modalidad, ocupando un sector de una superficie proporcional a su frecuencia. Para determinar el ángulo de los sectores de cada modalidad, se multiplica la frecuencia relativa (proporción) por 360.



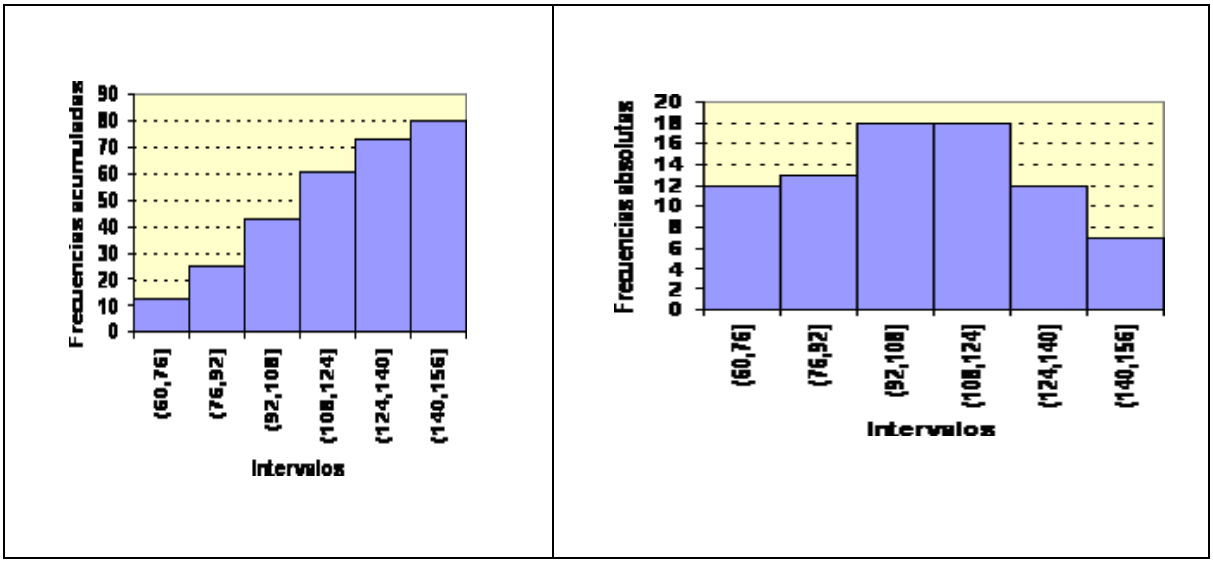
Variable: color del pelo

X	n_i	p_i	Grados diagrama
Rojo	20	0,2	72
Rubio	30	0,3	108
Negro	50	0,5	180
	100	1,00	360

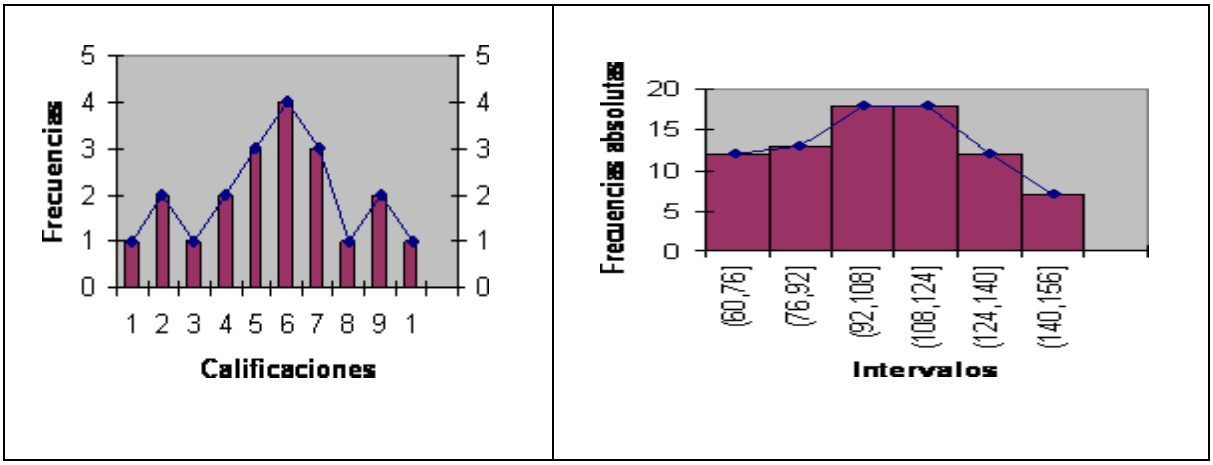
Pictograma: Expresan con dibujos alusivos al objeto de estudio las frecuencias de las modalidades de la variable. Los gráficos son del mismo dibujo a diferentes escalas (el área de cada uno debe ser proporcional a la frecuencia que representa). Se utilizan con variables cualitativas.



Histograma: Se emplea para representar distribuciones de frecuencias (absolutas y acumuladas) de variables continuas cuando los datos están agrupados en intervalos. En el eje de abscisas se colocan los **límites exactos** de cada uno de los intervalos, levantándose rectángulos cuyas áreas sean proporcionales a la frecuencia correspondiente.



Polígono de frecuencias: En caso de que la variable sea discreta, el polígono de frecuencias es el gráfico que resulta de unir los extremos superiores de las barras o líneas. En el caso de que la variable sea continua, el polígono de frecuencias es el gráfico que resulta de unir los ***puntos medios*** de las barras o rectángulos del histograma. Para el polígono de frecuencias acumuladas se unen mediante ***una línea los puntos de la base superior de cada rectángulo***, trazando una diagonal entre el extremo superior derecho y el inferior izquierdo. A medida que aumenta el número de observaciones (y el de intervalos), el polígono de frecuencias va suavizando su aspecto quebrado hasta que, en el límite, presenta el aspecto (ideal) de una curva.



PARA DOS VARIABLES:

Diagrama de barras conjunto: Se utiliza cuando, al menos una de las variables es cualitativa. Distribución conjunta de las variables calificaciones y grupos.

Cuando se representan dos variables conjuntamente, para utilizar frecuencias absolutas es conveniente que el número de sujetos sea similar en ambas variables; en caso contrario, son preferibles los porcentajes.

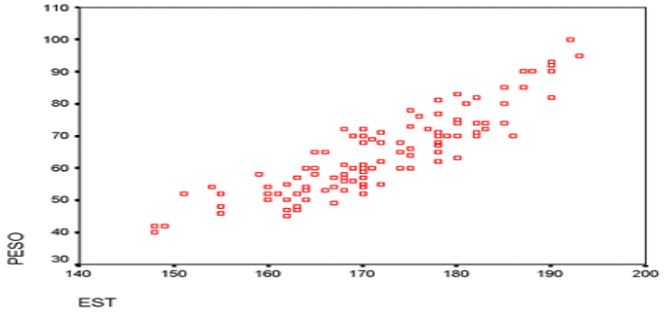


Diagrama de dispersión (nube de puntos): Se utiliza cuando las dos variables son cuantitativas. En el gráfico se muestra la representación conjunta de las variables estatura y peso. Atendiendo al diagrama de dispersión, podemos observar que existe una relación lineal entre ambas.

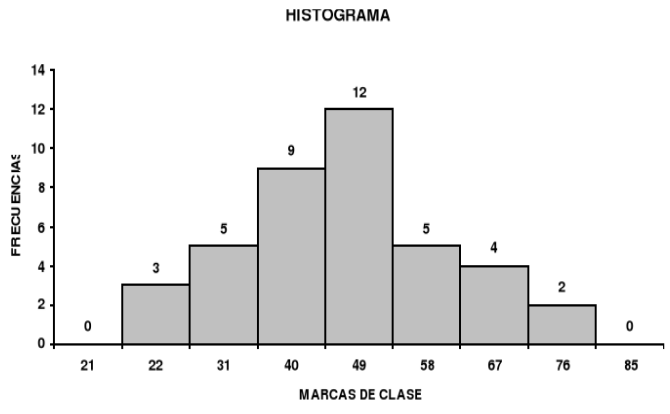
Convenciones respecto a las Representaciones Gráficas:

- ✓ En el eje de abscisas (horizontales) se señalan las diferentes modalidades o los valores y puntuaciones de la variable, y en el de ordenadas (verticales) las frecuencias absolutas, relativas o acumuladas.
- ✓ Los valores de la variable se sitúan en orden ascendente (de izquierda a derecha).
- ✓ Cuando los valores mínimos son muy altos se suelen hacer cortes en los ejes.
- ✓ Se debe incorporar a los gráficos la información necesaria para su correcta comprensión.
- ✓ Cuando se representan conjuntamente dos muestras con notables diferencias en el número de observaciones, las gráficas con frecuencias relativas son las más adecuadas.

3.- PROPIEDADES DE LA DISTRIBUCIÓN DE FRECUENCIAS

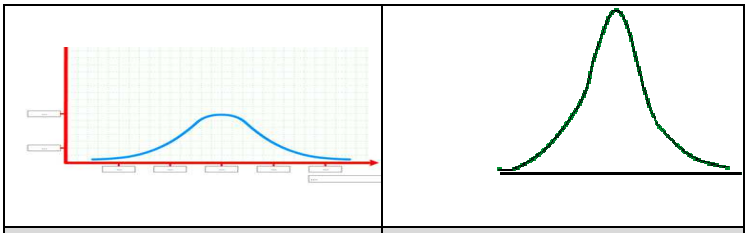
La forma de campana es la más habitual en las distribuciones de frecuencias (variables cuantitativas) que aluden a las características que estudian las Ciencias Sociales. La forma de una distribución de frecuencias teniendo en cuenta sus propiedades básicas son:

Tendencia Central: Promedios (se ubican en el centro y sintetizan los valores de la distribución)



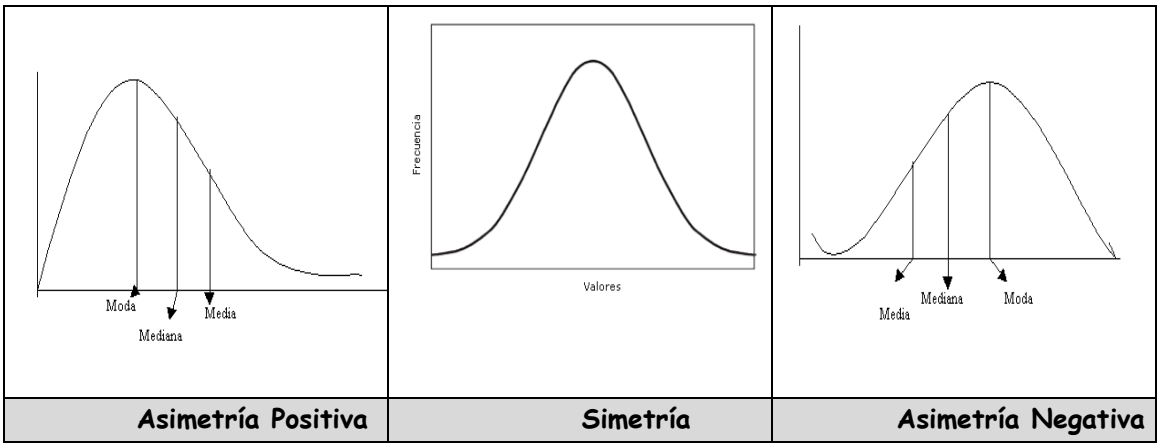
Intervalos X	Punto Medio Marca de clase	Frecuencia Absoluta n_i
18-26	22	3
27-35	31	5
36-44	40	9
45-53	49	12
54-62	58	5
63-71	67	4
72-80	76	2
		N = 40

Variabilidad: Resume el grado de concentración de los valores de una distribución alrededor de su promedio. Cuando los valores están muy cercanos al promedio (distribución homogénea / poca variabilidad); en caso contrario, cuando se alejan del promedio (distribución heterogénea / mucha variabilidad)



Heterogénea	Homogénea

Sesgo o Asimetría: Grado de simetría o asimetría de una distribución de frecuencias respecto al promedio. En caso de observaciones similares por debajo y por encima del promedio (simetría). Cuando se aprecia una mayor frecuencia de valores bajos (asimetría positiva, propia de las tareas difíciles); si es mayor la frecuencia de valores altos (asimetría negativa, propia de las tareas fáciles)



Asimetría Positiva	Simetría	Asimetría Negativa

TEMA 2 → MEDIDAS DE TENDENCIA CENTRAL Y DE POSICIÓN

OBJETIVOS DE APRENDIZAJE:

- Conocer las características de las principales medidas de tendencia central (media aritmética, mediana y moda) y de posición (percentiles, cuartiles y deciles)
- Saber aplicar los índices de tendencia central y de posición.
- Seleccionar los índices de tendencia central y de posición adecuados en cada caso.
- Interpretar correctamente los valores obtenidos mediante los índices de tendencia central y de posición.

INTRODUCCIÓN ➔ Para proceder al análisis descriptivo de una variable se cuantifican las propiedades de su distribución de frecuencias mediante una serie de índices. Entre los que permiten cuantificar la centralidad de la distribución se estudian: **la Moda, la Mediana y la Media**. Los índices de tendencia central son representativos de la distribución, dado que en torno a ellos se concentran la mayor parte de las observaciones registradas.
Por otra parte, los índices de posición permiten determinar la situación relativa de cada puntuación respecto al grupo y se denominan **Percentiles o Centiles** (dividen la distribución en 100 partes iguales).

1.- MEDIDAS DE TENDENCIA CENTRAL: Los índices de centralidad o concentración (Promedios) cuantifican el valor central representativo de toda la distribución (valores alrededor de los cuales se concentran la mayor parte de las observaciones)

A.- LA MODA (**Mo:** valor o categoría de la variable con mayor frecuencia absoluta): Es el índice de tendencia central que menos información aporta ya que sólo señala la categoría (variables nominales y ordinales) o el valor más frecuente (variables cuantitativas). La Moda es el único índice de tendencia central que puede aplicarse con variables cualitativas y cuantitativas.

VARIABLES NOMINALES
(Categoría o modalidad más frecuente)

X_i	n_i	
Solteras	25	
Casadas	***50***	**Mo**
Divorciadas	15	
Viudas	10	
	100	

VARIABLES ORDINALES
(Grado o magnitud de la característica más frecuente)

Es preciso ordenar previamente los valores para facilitar la lectura de las frecuencias.

X_i	n_i	
Pésimo	5	
Regular	15	
Bueno	25	
Muy bueno	***45***	**Mo**
Excelente	10	
	100	

VARIABLES CUANTITATIVAS (DE INTERVALO O RAZÓN)

- Punto medio del intervalo con mayor frecuencia.
- Cuando se trata de intervalos abiertos (intervalo superior o inferior sin límite (o más / o menos) la Moda no puede caer en uno de ellos.

$M_0 = 10$

X_i	n_i	
3 - 5	6	
6 - 8	9	
9 - 11	18	**Punto Medio**
12 - 14	12	
15 - 17	7	

Propiedades de la Moda (M_o):

- ➔ Una distribución de frecuencias puede ser **Amodal / Bimodal / Multimodal** (igual frecuencia en todos los valores o categorías; dos categorías o valores con igual frecuencia y más de dos categorías o valores con igual frecuencia).
- ➔ En una distribución de frecuencias de variables agrupadas en intervalos, la moda es función de los intervalos elegidos (amplitud de intervalo, número de intervalos y límites de los mismos).
- ➔ En distribuciones abiertas de variables agrupadas en intervalos (sin límite inferior y superior en el intervalo inferior y superior respectivamente), la moda puede calcularse siempre que la frecuencia máxima no pertenezca a alguno de los intervalos abiertos.

B.- LA MEDIANA (**Md**: deja por debajo y por encima de sí el 50% de los casos). Aporta más información que la Moda ya que indica la puntuación que divide la distribución en dos grupos con el mismo número de casos. No nos dice nada de los valores extremos de la distribución porque no está afectado por ellos (índice apropiado para distribuciones asimétricas). Para proceder a su cálculo los datos siempre deben ordenarse (habitualmente se sigue un orden ascendente, de menor a mayor).

Propiedades de la Mediana:

➔ La mediana puede ser calculada con **intervalos abiertos** (intervalo superior o inferior sin límite - o más / o menos) siempre que éstos no coincidan con el intervalo crítico (se desconocería su amplitud).
➔ La Mediana divide el área total del Histograma de frecuencias en dos áreas con idéntica superficie.

Cálculo para datos agrupados en intervalos:

Método de INTERPOLACIÓN ➔ $Md = Li + [(n /2) – n_d / n_c] \cdot I$

Li = Límite exacto Inferior del Intervalo Crítico.
n = Número total de observaciones.
n_d = Número de observaciones acumuladas por debajo del Intervalo Crítico.
n_c = Número de observaciones o frecuencia absoluta del Intervalo Crítico.
I = Amplitud del Intervalo Crítico.

Ejemplo: Dada la siguiente distribución de frecuencias, calcular la Mediana.

X	n_i = Frecuencias Absolutas	n_a = Frecuencias Acumuladas	p_i = frecuencia relativa	p_a = frecuencia relativa acumulada
10 - 12	10	200	0,05	1
7 – 9	**100**	**190**	**0,5**	**0,95**
4 – 6	60	90	0,3	0,45
1 - 3	30	30	0,15	0,15
	N = 200		1	

Intervalo Crítico (Intervalo en el que se encuentra la Mediana). Intervalo cuya frecuencia acumulada sea mayor o igual al 50% del número de observaciones (100); en nuestro caso, el intervalo (7 – 9).
Aplicamos la fórmula (Interpolación) ➔ $Md = Li + [(n /2) – n_d / n_c] \cdot I$
Mediana = $Md = 6,5 + [(200 / 2) – 90 / 100] \cdot 3 =$ **6,8**
La Mediana (6,8) deja por encima y por debajo el 50% de las observaciones

Con las proporciones: **$Md = L_i + [(0,5 – p_d) / p_c] \cdot I$**

P_d = Proporción acumulada bajo el intervalo crítico // p_c = Proporción dentro del intervalo crítico

$Md = 6,5 + [(0,5 – 0,45) / 0,5] \cdot 3 =$ **6,8**

Cálculo para datos no agrupados en intervalos:

→ Nº impar de observaciones: 2, 4, 6, **8**, 10, 60, 99 (**Md =** puntuación central **8**)
→ Nº par de observaciones: 5, 7, 8, 10, 12, 78 (**Md =** media puntuaciones centrales **8 + 10 / 2 = 9**)
→ Observaciones repetidas (independientemente de su número): (12, 13, 15, 19, 19, 19, 19, 19, 20, 21, 22). Se agrupan en forma de intervalo de amplitud 1 y se recogen las frecuencias absolutas (n_i) y acumuladas (n_a):

X_i	n_i	n_a
22	1	11
21	1	10
20	1	9
19	5	8
15	1	3
13	1	2
12	1	1

La puntuación 19 es el punto medio del intervalo crítico 18,5 y 19,5
Aplicando el método de Interpolación:

Md = 18,5 + [(11/2 – 3) / 5] · 1 = 19

C.- LA MEDIA ARITMÉTICA (**$\bar{X}$:** suma de las puntuaciones dividida por el número de casos): Índice que toma en consideración todas las puntuaciones registradas. Sólo puede calcularse con variables cuantitativas. Para proceder a su cálculo se toman en consideración dos posibilidades:

Para valores no agrupados en intervalos → $\bar{X} = \Sigma x_i / n$

Ejemplo → Obtener la Media Aritmética de los siguientes valores: 4, 7, 8, 12, 6

$\bar{X} = \Sigma x_i / n$ → (4 + 7 + 8 + 10 + 6) / 5 = **35 / 5 = 7**

Donde: $\Sigma x_i = n \cdot \bar{X}$ → 35 = (5 · 7)

Para valores agrupados en intervalos → $\bar{X} = \Sigma (n_i \cdot X_i) / n$

Ejemplo → Dada la siguiente distribución de frecuencias absolutas (n_i) calcular la Media Aritmética:

X	n_i	X_i	$n_i \cdot X_i$
10 - 12	10	11	110
7 – 9	100	8	800
4 – 6	60	5	300
1 - 3	30	2	60
Intervalos	**N = 200**	**Punto medio Intervalos**	**$\Sigma (n_i \cdot X_i) = 1.270$**

$\bar{X} = \Sigma (n_i \cdot X_i) / n$ →

$\bar{X} = 1.270 / 200 = 6,35$

Ejemplo → Dada la siguiente distribución de frecuencias relativas o proporciones (p_i) calcular la Media:

X	X_i	p_i	$X_i \cdot p_i$
10 - 12	11	10 / 200 = **0, 05**	0, 55
7 – 9	8	100 / 200 = **0, 5**	4
4 – 6	5	60 / 200 = **0, 3**	1, 5
1 - 3	2	30 / 200 = **0, 15**	0, 3
Intervalos	**Punto medio Intervalos**	**Proporciones**	**$\Sigma (X_i \cdot p_i) = 6,35$**

$\bar{X} = \Sigma (X_i \cdot p_i)$ →

$\bar{X} = 6, 35$ (resultado similar al anterior).

Propiedades de la Media Aritmética:

→ **$\Sigma (X_i - \bar{X}) = 0$.** La suma de las diferencias de n puntuaciones respecto a su media vale cero (índice de tendencia central).
→ Para intervalos abiertos (ó más / ó menos) no es posible calcular la Media (no se puede obtener el punto medio).
→ La media es un índice apropiado para una distribución simétrica.

→ Si cada una de las puntuaciones de una distribución **(X)** se multiplica por una constante **(b)** y se le suma otra constante **(a)**, la media de las nuevas puntuaciones **(Y)** es igual a la media de las puntuaciones originales **(X)** multiplicada por la constante **b** más la constante **a**:

$$\bar{Y} = (b \cdot \bar{X}) + a \qquad \textbf{(Transformación Lineal)}$$

D.- CRITERIOS PARA ELEGIR EL ÍNDICE DE TENDENCIA CENTRAL

En función del tipo de variable que se esté analizando y de las características de la distribución de frecuencias (Cuando la distribución es simétrica coinciden los tres índices Moda = Mediana = Media). Para la elección del índice de tendencia central más adecuado (moda, mediana o media) podemos utilizar los siguientes criterios:

Tendencia central	*Tipo de variable*
Moda	Cualitativo (nominal) *(sexo, estado civil)*
Mediana	Ordinal (cuasi-cuantitativa)
Media	Intervalo o razón
	Cuantitativa continua

No es posible calcular la **media** cuando el nivel de medida de la variable es nominal u ordinal y cuando hay algún intervalo abierto (no es aconsejable en distribuciones asimétricas)
No es posible calcular la **mediana** cuando el nivel de medida de la variable es nominal y cuando la mediana se encuentra en el intervalo abierto (es aconsejable en distribuciones asimétricas)
No es posible calcular la **moda** cuando la distribución es amodal o la moda se ubica en un intervalo abierto.

2.- MEDIDAS DE POSICIÓN: Permiten determinar la posición relativa de cada puntuación respecto al grupo. Genéricamente se denominan **cuantiles** (percentiles, cuartiles y deciles). La Mediana, ya estudiada, se considera un índice de posición fundamental; ya que, por encima y por debajo, deja el 50% de las observaciones. El cálculo de los cuantiles es una generalización del cálculo de la mediana.

Cálculo para datos no agrupados en intervalos: Se ordenan previamente los valores de forma ascendente y se determina el percentil.

Ejemplo: Sean las siguientes datos → 12, 13, 15, 19, 20, 21, 22, 23, 30, 32, 34, 35, 36, 37, 39. El percentil 20 **(P_{20})** sería aquel que deje 3 observaciones por debajo (3 es el 20% de 15 puntuaciones). Así, será un valor comprendido entre las posiciones tercera y cuarta (15 + 19 / 2 = 17) → **$(P_{20}) = 17$** (la puntuación 17 deja por debajo el 20% de las observaciones).

Cálculo para datos agrupados en intervalos:

Método de INTERPOLACIÓN → $P_k = Li + \{ (k \cdot n / 100) - n_d / n_c \} \cdot I$

P_k = Puntuación correspondiente al Percentil k
K = Porcentaje de observaciones inferiores a **P_k**
Li = Límite exacto Inferior del Intervalo Crítico.
n = Número total de observaciones.
n_d = Número de observaciones acumuladas por debajo del Intervalo Crítico.
n_c = Número de observaciones o frecuencia absoluta del Intervalo Crítico.
I = Amplitud del Intervalo Crítico.

Ejemplo: Dada la siguiente distribución de frecuencias, calcular los Percentiles 36, 50, 25 y 75.

Límites Virtuales de X	Punto Medio Intervalo	n_i = Frecuencias Absolutas	n_a = Frecuencias Acumuladas
103 – 108	105, 5	8	50
97 – 102	99, 5	11	42
91 – 96	93, 5	16	31
85 – 90	87, 5	10	15
79 – 84	81, 5	5	5
		N = 50	

→ Localizamos el **Intervalo Crítico** en el que se encuentra cada Percentil:

P_{25} = (El 25% de 50 observaciones es 12,5); **(k · n / 100) →** (25 · 50 / 100) = **12, 5**

P_{36} = (El 36% de 50 observaciones es 18); **(k · n / 100) →** (36 · 50 / 100) = **18**

P_{50} = (El 50% de 50 observaciones es 25); **(k · n / 100) →** (50 · 50 / 100) = **25**

P_{75} = (El 75% de 50 observaciones es 37, 5); **(k · n / 100) →** (75 · 50 / 100) **= 37, 5**

→ Aplicamos la fórmula (Interpolación): $P_k = Li + \{(k \cdot n / 100) - n_d / n_c\} \cdot I$

P_{25} = Intervalo (85 – 90) → P_{25} = 84, 5 + [(25 · 50 / 100) – 5 / 10] · 6 **= 89**

P_{36} = Intervalo (91 – 96) → P_{36} = 90, 5 + [(36 · 50 / 100) – 15 / 16] · 6 **= 91, 625**

P_{50} = Intervalo (91 – 96) → P_{50} = 90, 5 + [(50 · 50 / 100) – 15 / 16] · 6 **= 94, 25**

P_{75} = Intervalo (97 – 102) → P_{25} = 96, 5 + [(75 · 50 / 100) – 31 / 11] · 6 **= 100**

Las puntuaciones obtenidas se corresponden con el Percentil buscado y se interpretan señalando el porcentaje de puntuaciones que "dejan por debajo". Así, el **P_{25} = 89** (la puntuación 89 deja por debajo de sí el 25% de las observaciones de la distribución). Utilizando el mismo procedimiento se puede averiguar **el Percentil que corresponde a una puntuación dada:**

→ Utilizando la Interpolación comprobamos el Percentil que corresponde a la puntuación X = 100.

$K = [\{(P_k - Li) \cdot n_c / I\} + n_d / n] \cdot 100$

X = 100; Intervalo (97 – 102) → K = [{(100 - 96,5) · 11 / 6} + 31 / 50] · 100 = **75**

Equivalencias de las medidas de posición	
Deciles - Percentiles	**Cuartiles - Percentiles**
$D_1 - P_{10}$	
$D_2 - P_{20}$	
	→ $Q_1 - P_{25}$
$D_3 - P_{30}$	
$D_4 - P_{40}$	
$D_5 - P_{50}$ →	$Q_2 - P_{50}$
$D_6 - P_{60}$	
$D_7 - P_{70}$	
→	$Q_3 - P_{75}$
$D_8 - P_{80}$	
$D_9 - P_{90}$	

TEMA Nº 3 → MEDIDAS DE VARIABILIDAD Y ASIMETRÍA

OBJETIVOS DE APRENDIZAJE:

- Conocer las características de los principales índices para medir la variabilidad en una distribución de frecuencias, con especial énfasis en la varianza y la desviación típica.
- Saber aplicar los índices de variabilidad o dispersión a una determinada distribución.
- Conocer y saber aplicar el índice de Pearson para analizar el grado y el tipo de asimetría de una distribución.
- Distinguir entre los distintos tipos de puntuaciones: directas, diferenciales y típicas, la información que proporcionan y sus propiedades fundamentales.

INTRODUCCIÓN → Los índices de variabilidad y asimetría permiten analizar descriptivamente dos características fundamentales de cualquier distribución de frecuencias: su variabilidad y su forma.

→ Índices de variabilidad: **Amplitud Total y Amplitud Semi–Intercuartil** (miden el grado en el que las puntuaciones se asemejan o diferencian entre sí). **La Varianza y la Desviación Típica** (miden la dispersión con relación a alguna medida de tendencia central como la media) y el **Coeficiente de Variación** (permite comparar la dispersión de dos o más distribuciones)
→ Índice de sesgo o asimetría de **Pearson.**

1.- MEDIDAS DE VARIABILIDAD → Cuantifican en qué medida las puntuaciones en una determinada variable están próximas o alejadas del índice de tendencia central que representa a la distribución. Los más utilizados son:

Amplitud Total (Rango o Recorrido) → $A_T = X_{máx} - X_{min}$
Distancia en la escala numérica entre la puntuación máxima y la mínima. Se denomina **incluyente** cuando se toman los valores reales (± 0,5) y **excluyente** los informados. **Problema:** sólo toma en consideración dos puntuaciones, si permanecen constantes, es insensible a las variaciones de la distribución(se utiliza como información complementaria de otras medidas de dispersión)

Amplitud o Rango Semi–Intercuartil → $Q = (Q_3 - Q_1) / 2 = (P_{75} - P_{25}) / 2$
Distancia media entre el tercer y el primer cuartil. Toma en consideración el 50% de las puntuaciones centrales (distancia entre el tercer y el primer cuartil). Se utiliza cuando el índice de tendencia central más adecuado es la **Mediana**. También se conoce como índice de ambigüedad (escalas de actitudes). **Problema:** no es útil cuando las distribuciones presentan valores extremos.

Varianza (S^2_X), Desviación Típica (S_X) y Cuasivarianza (S^2_{n-1}) →

Datos sin agrupar →	Datos agrupados en intervalos →
$S^2_X = \Sigma (X_i - \bar{X})^2 / n$ $S^2_X = \Sigma p_i \cdot X^2 - \bar{X}^2$ $S^2_X = (\Sigma X^2 / n) - \bar{X}^2$	$S^2_X = \Sigma n_i \cdot (X_i - \bar{X})^2 / n$ $S^2_X = \Sigma p_i \cdot X^2 - \bar{X}^2$ $S^2_X = (\Sigma n_i \cdot X^2 / n) - \bar{X}^2$

Desviación Típica → $(S_X) = \sqrt{S^2_X} = \sqrt{(\Sigma n_i \cdot X^2 / n) - \bar{X}^2}$
En todos los casos se obtiene extrayendo la raíz cuadrada de la VARIANZA.

Cuasivarianza → $S^2_{n-1} = \Sigma (X_i - \bar{X})^2 / n - 1$
La raíz cuadrada de la Cuasivarianza da lugar a la Cuasidesviación típica → S_{n-1}

La **varianza** se define como el promedio de las diferencias de cada puntuación respecto a su media (las diferencias se elevan al cuadrado para evitar que sean nulas). La **desviación típica** es la raíz cuadrada de la varianza (se suele utilizar más que la varianza debido a que se expresa en las mismas

unidades de medida que la variable objeto de estudio). La **Cuasivarianza** es el mejor estimador de la varianza poblacional (siempre es mayor que la varianza, se van igualando cuando las muestras son grandes).

Propiedades de la Varianza y la Desviación Típica ➔ ($S^2_y = b^2 \cdot S^2_X$)

- ➔ Sólo pueden tomar **valores positivos** y son **sensibles a la variación** de cada una de las puntuaciones de la distribución.
- ➔ Miden la variabilidad de los datos cuando utilizamos como índice de tendencia central la media aritmética.
- ➔ Si a un conjunto de puntuaciones se le suma una constante **(a)**, la varianza y la desviación típica de las nuevas puntuaciones son idénticas a las de las puntuaciones originales.
- ➔ Si un conjunto de puntuaciones se multiplica por una constante **(b)**, la varianza de las nuevas puntuaciones es igual a la varianza de las puntuaciones originales multiplicada por el cuadrado de **(b^2)** y la desviación típica por el valor absoluto de b.

Coeficiente de Variación → $CV = (S_X / \overline{X}) \cdot 100$
Se utiliza para **comparar la variabilidad** de las muestras de distintos grupos (con similares o distintas características). Al multiplicarlo por 100, se obtiene un porcentaje.

Problemas Ejemplo: A partir de la distribución que se indica, averiguamos de entre los distintos índices expuestos, los más utilizados:

X	n_i	n_a	X_i	$n_i \cdot X_i$	$n_i \cdot X^2_i$	$(X - \overline{X})^2$	$n_i \cdot (X - \overline{X})^2$	$n_i \cdot (X - \overline{X})^3$
14 – 16	3	70	15	45	675	51,26	153,78	1101,18
11 – 13	12	67	12	144	1728	17,30	207,6	863,89
8 – 10	20	55	9	180	1620	1,34	26,8	31.,218
5 – 7	25	35	6	150	900	3,38	84,5	155,738
2 – 4	10	10	3	30	90	23,42	234,2	1133,80
Sumatorios Σ	**70**			**549**	**5. 013**	**96,7**	**706,88**	**3285,83**

Amplitud Total (Rango o Recorrido) → $A_T = X_{máx} - X_{min}$

A_T = 16 – 2 = 14 (excluyente) → A_T = 16,5 – 1,5 = 15 (incluyente)

Amplitud o Rango Semi–Intercuartil → $Q = (Q_3 - Q_1) / 2 = (P_{75} - P_{25}) / 2$

$Q_3 = P_{75} = 7,5 + \{(52,5 - 35) / 20) \cdot 3\} = 10, 125$ // $Q_1 = P_{25} = 4,5 + \{(17,5 - 10) / 25) \cdot 3\} = 5, 4$

$Q = (Q_3 - Q_1) / 2 = (P_{75} - P_{25}) / 2 = (10,125 - 5,4) / 2 = 2,36$

Varianza (S^2_X) → $S^2_X = \Sigma n_i \cdot (X_i - \overline{X})^2 / n = (\Sigma n_i \cdot X^2 / n) - \overline{X}^2$

$\overline{X} = (\Sigma n_i \cdot X_i / n) = 549 / 70 = 7,84$ // $S^2_X = (\Sigma n_i \cdot X^2 / n) - \overline{X}^2 = (5. 013 / 70) - 7,84^2 = 10, 10$

Desviación Típica (S_X) → $S_X = \sqrt{S^2_X} = \sqrt{10,10} = 3,18$

Cuasivarianza (S^2_{n-1}) → $S^2_{n-1} = \Sigma n_i \cdot (X_i - \overline{X})^2 / n - 1 =$ 706, 88 / 69 = 10,24

Coeficiente de Variación: $CV = (S_X / \overline{X}) \cdot 100$

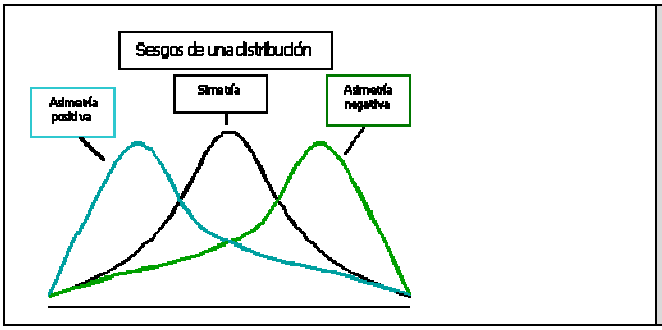
Notas alumnos de las clases A y B → **A =** (X = 6´12 y S_X = 1´27) y **B =** (X = 102 y S_X = 4)

CV_A = (1´27 / 6´12) · 100 = 20´75 CV_B = (4 / 102) · 100 = 3´92

El coeficiente de variación es mayor en el grupo A (mayor dispersión de los datos en el grupo A). El grupo B es más homogéneo (menor dispersión de los datos).

2.- <u>MEDIDA DE SESGO O ASIMETRÍA</u> ➔ Cuando una distribución es simétrica: Media = Mediana = Moda. Asimismo coinciden las distancias entre los cuartiles. Cuando una distribución es asimétrica se cuantifica mediante el índice de asimetría de Pearson. Sólo es útil para distribuciones unimodales. Es un índice adimensional (no tiene unidades de medida)

<u>Índice de Asimetría de Pearson</u> →



$A_S = (\bar{X} - Mo) / S_X$

RESULTADOS
$A_S < 0$ → Asimetría -
$A_S > 0$ → Asimetría +
$A_S = 0$ → Simetría

Índice de Asimetría de Pearson → $A_S = (\bar{X} - Mo) / S_X$ (Con los datos del problema ejemplo)

$\bar{X} = (\Sigma n_i \cdot X_i / n) = 549 / 70 = \mathbf{7,84}$ $\quad S_X = \sqrt{S^2_X} = \sqrt{10,10} = \mathbf{3,18}$ $\quad Mo = 6$

$A_S = (\bar{X} - Mo) / S_X$ → $(7,84 - 6) / 3,18 = \mathbf{0,58}$ (Asimetría Positiva)

3.- <u>PUNTUACIONES TÍPICAS</u>

Puntuaciones Directas (X) → Puntuación obtenida / atribuida a un sujeto / objeto sometido a cualquier prueba. No aporta mucha información si no la relacionamos con el promedio del grupo.

Puntuaciones Diferenciales (de desviación o de diferencia) → $(x = X - \bar{X})$
Puntuación en relación con la media (índice que representa al grupo). No sirven para comparar puntuaciones de dos grupos con diferente unidad de medida. Su media vale cero y su varianza es igual que la de las puntuaciones directas.

<u>Puntuaciones Típicas o tipificadas</u> (Z) → Indican el número de desviaciones típicas que se aparta de la media una determinada puntuación.

$Z = (X - \bar{X}) / S_X$ → P. Diferencial / D. Típica del grupo de referencia.

<u>Propiedades:</u> Su media vale cero ($\bar{Z} = 0$), su varianza vale uno ($S^2_X = 1$). Permiten relacionar las puntuaciones con independencia de la unidad de medida. Cuando la Puntuaciones Típicas son iguales, las Puntuaciones Directas correspondientes son equivalentes.

<u>Problema ejemplo</u>: Dadas dos series de puntuaciones directas, determinar si son equivalentes.

X = 2, 3, 5, 2 Y = 3, 5, 9, 3

X	Y	$(X - \bar{X})^2$	$(Y - \bar{Y})^2$	$Z = (X - \bar{X}) / S_X$	$Z = (Y - \bar{Y}) / S_Y$
2	3	1	4	- 0,81	- 0,81
3	5	0	0	0,0	0,0
5	9	4	16	1,63	1,63
2	3	1	4	- 0,81	- 0,81
12	**20**	**6**	**24**		

$\bar{X} = 12 / 4 = 3$ $\quad \bar{Y} = 20 / 4 = 5$ $\quad S_X = \sqrt{\Sigma (X - \bar{X})^2 / n}$ → $S_X = \sqrt{6 / 4} = 1,23$

$S_Y = \sqrt{\Sigma (Y - \bar{Y})^2 / n}$ → $S_Y = \sqrt{24 / 4} = 2,45$

→ Transformadas las Puntuaciones Directas (X e Y) en Puntuaciones Típicas (Z_X y Z_Y), se comprueba que tienen idénticos valores, luego ambas series son equivalentes.

TEMA Nº 4 → ANÁLISIS CONJUNTO DE DOS VARIABLES

OBJETIVOS DE APRENDIZAJE:

- Distinguir entre variables cualitativas y cuantitativas, y saber elegir los métodos en cada caso.
- Conocer métodos gráficos y cuantitativos para analizar la relación existente entre dos variables.
- Adquirir la capacidad para saber si dos variables están más o menos relacionadas entre sí, la forma de esa relación, y el significado de que dos variables estén relacionadas.
- En el caso de dos variables cuantitativas, entre las que hay relación lineal, aprenderemos a hacer predicciones de los valores de la variable Y, correspondientes a cada valor de la variable X, mediante la recta de regresión.

INTRODUCCIÓN → Cuando la información caso-sujeto contiene dos variables, se puede analizar la relación o asociación que hay entre ellas.

Asociación y/o relación: dos variables están relacionadas entre sí, cuando ciertos valores, de una de las variables, se asocian con ciertos valores de la otra variable. Los índices que cuantifican esta relación tienen en cuenta el nivel de medida de las variables; en nuestro caso, consideramos el estudio de las relaciones entre:
Dos variables cualitativas (categóricas / medidas en una escala nominal): Chi-Cuadrado (χ^2) y Coeficiente de Contingencia (**C**).
Dos variables cuantitativas (medidas en una escala de intervalo o de razón): Covarianza **(S_{XY})** y Coeficiente de correlación lineal de Pearson **(r_{XY})**
Cuando las variables están relacionadas linealmente, podemos utilizar una de ellas para realizar predicciones o pronósticos sobre la otra **(análisis de regresión)**

1.- RELACIÓN ENTRE DOS VARIABLES CUALITATIVAS →

LA PRUEBA CHI – CUADRADO (χ^2) → Permite determinar si dos variables están o no relacionadas. Se fundamenta en la comparación de las frecuencias conjuntas empíricas u observadas **(n_e)** y las frecuencias conjuntas teóricas o esperadas **(n_t)**, en el caso de que ambas variables sean independientes (no relacionadas). Uno de los problemas que plantea este índice es que es difícil de interpretar.

Problema ejemplo: Determinar si existe relación entre situación laboral (se trabaja / no se trabaja) y tipo de enseñanza elegida (tradicional / a distancia) por estudiantes universitarios. En una muestra de 2000 alumnos se observó que el 20% estaban matriculados en enseñanza a distancia, y de ellos trabajaban 300, mientras que de los alumnos matriculados en la enseñanza tradicional sólo trabajaba un 10%.

X = TIPO DE ENSEÑANZA ↓	Y = SITUACIÓN LABORAL →		
	TRABAJAN (0)	NO TRABAJAN (1)	MARGINAL DE X ↓
A DISTANCIA (0)	**300**	100	**400**
TRADICIONAL (1)	**160**	1440	2000-400=1600
MARGINAL DE Y →	460	1540	**2000**

Datos problema: 20% (a distancia) de 2000 → 400 y 10% de 1600 (tradicional) → 160. Para calcular **la tabla de frecuencias teóricas (n_t) → (**Σ Sumatorio de la fila x Σ Sumatorio de la columna) / Nº de casos

X = TIPO DE ENSEÑANZA ↓	Y = SITUACIÓN LABORAL →		
	TRABAJAN (0)	NO TRABAJAN (1)	MARGINAL DE X ↓
A DISTANCIA (0)	**(400·460)/2000 92**	**(400·1540/2000 308**	**400**
TRADICIONAL (1)	**(1600·460)/2000 368**	**(1600·1540)/2000 1232**	**1600**
MARGINAL DE Y →	**460**	**1540**	**2000**

$\chi^2 = \Sigma\Sigma (n_e - n_t)^2 / n_t$	n_e → frecuencia empírica conjunta (datos problema) n_t → frecuencia teórica conjunta (datos esperados)
$\chi^2 =$ {(300 - 92)^2 / 92 + (100 - 308)^2 / 308 + (160 - 368)^2 / 368 + (1440 - 1232)^2 / 1232} **= 763´41**	

Interpretación: Se estudian las diferencias entre las frecuencias empíricas (n_e) y las frecuencias teóricas (n_t). La suma de filas y la suma de columnas en esta tabla debe ser igual a cero.

X = TIPO DE ENSEÑANZA ↓	**Y = SITUACIÓN LABORAL →**	
	TRABAJAN (0)	**NO TRABAJAN (1)**
A DISTANCIA (0)	300 – 92 = 208	100 – 308 = - 208
TRADICIONAL (1)	160 – 368 = - 208	1440 – 1232 = 208

Los valores positivos de las diferencias (208) indican relación positiva entre A distancia y Trabajan (0,0) / Tradicional y No trabajan (1,1). La dificultad de interpretación radica en que no conocemos el límite superior de Chi-Cuadrado (χ^2 = 763´41 → no conocemos el valor máximo) y por tanto, desconocemos la intensidad de la relación.

Para ello se define el **COEFICIENTE DE CONTINGENCIA (C)** → $\mathbf{C = \sqrt{\chi^2 / (\chi^2 + n)}}$
Sus valores están comprendidos entre 0 y 1. Cuando el valor está próximo a 0 significa que las variables son independientes y cuando está próximo a 1 que están relacionadas. Con los datos de nuestro problema ejemplo:

$$\mathbf{C = \sqrt{763{,}41 / (763{,}41 + 2000)} = 0{,}526}$$

El valor C obtenido **se puede comparar** con un C máximo (dado que la tabla tiene igual número de filas que de columnas → k = 2)

$$\mathbf{C_{máx.} = \sqrt{(k - 1) / k} \rightarrow C_{máx.} = \sqrt{(2 - 1) / 2} = 0{,}707}$$

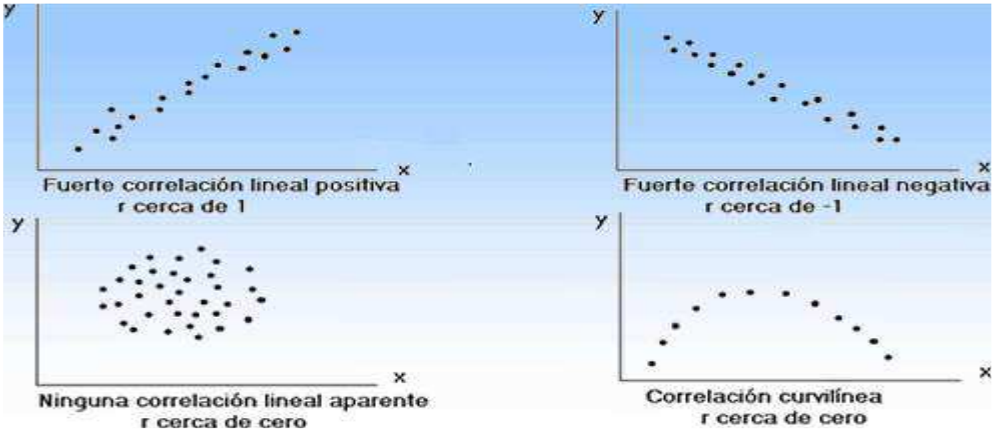
El valor obtenido **(C = 0´526)** comparado con **($C_{máx.}$ = 0´707)** es aproximadamente el 75%. La relación, por tanto, es alta.

En el caso de utilizar tablas de contingencia de más de dos filas o columnas, se procede de modo similar y el coeficiente de contingencia se interpreta de forma directa (se observa si se aproxima más o menos a 1 (no es posible averiguar el $C_{máx.}$ cuando el número de filas y columnas no es igual.

Propiedades del Coeficiente de Contingencia:
C sólo asume **valores entre 0 y 1** (0 = no hay relación entre las variables, las frecuencias empíricas coinciden con las teóricas y 1 = cuando no hay frecuencias empíricas, suceso imposible)
A **mayor** valor de **C, mayor relación** entre las variables y viceversa (a menor C, menor relación)
Nunca se puede afirmar que una **variable** es **causa** de la otra (muchas variables se relacionan entre sí porque existen variables ajenas que se relacionan con ambas)

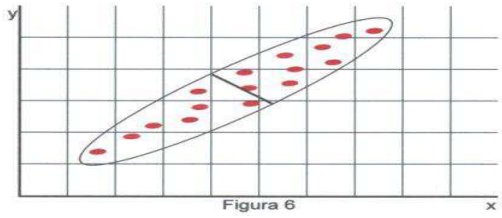
2.- <u>RELACIÓN ENTRE DOS VARIABLES CUANTITATIVAS</u> ➔ Los índices que permiten cuantificar la relación lineal entre dos variables cuantitativas son: La Covarianza y la Correlación de Pearson.

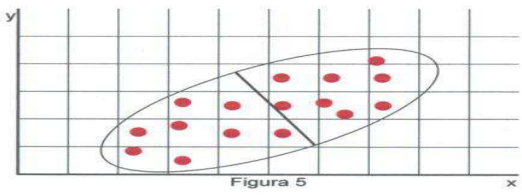
Representación gráfica de las variables cuantitativas → **Diagrama de dispersión** (nube de puntos)

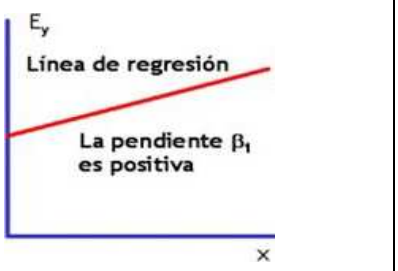


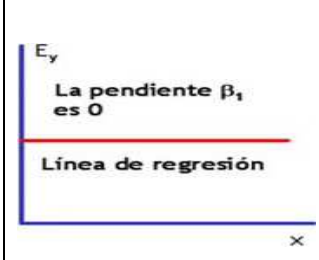
La relación entre variables puede ser no lineal (el coeficiente de correlación lineal sólo detecta relaciones lineales (una correlación cercana a cero descarta la relación lineal pero no descarta relaciones de otro tipo)

Cuanto menor es el diámetro de la elipse que encierra los puntos de un diagrama de dispersión, mayor es la relación de las variables que contiene. Cuando esto ocurre decimos que la correlación es fuerte. La figura 6 muestra una correlación positiva fuerte.









<u>LA COVARIANZA</u> ➔ Cov (X, Y) = S_{XY}

Es una medida de la relación entre dos variables (índice que cuantifica la variabilidad conjunta de dos variables). Se define como la media aritmética del producto de las puntuaciones diferenciales de dos variables medidas conjuntamente. Uno de sus inconvenientes es que depende de las unidades de medida de las variables.

Cálculo → Cov (X, Y) = S_{XY} = $\{\Sigma (X_i - \bar{X})(Y_i - \bar{Y})\} / n$; También → **$S_{XY}$** = $(\Sigma X_i \cdot Y_i / n) - (\bar{X} \cdot \bar{Y})$

Interpretaciones:

Cov(X, Y) positiva ➔ Expresa una relación directa entre X e Y (valores altos en X se corresponden con valores altos en Y, en promedio).

Cov(X, Y) negativa ➔ Expresa una relación inversa entre X e Y (valores altos en X se corresponden con valores bajos en Y, en promedio).

Problema Ejemplo: A partir de la siguiente tabla, averiguar la covarianza entre X e Y.

X	Y	x	y	x · y	X · Y
88	**6**	**-12**	**-1**	**12**	**528**
96	**4**	**-4**	**-3**	**12**	**384**
100	**7**	**0**	**0**	**0**	**700**
104	**8**	**4**	**1**	**4**	**832**
112	**10**	**12**	**3**	**36**	**1120**
500	**35**	**0**	**0**	**64**	**3564**

Media de $\overline{X}$ = 500/5 = 100 Media de $\overline{Y}$ = 35/5 = 7 Aclaraciones → **x = (X – $\overline{X}$) // y = Y – $\overline{Y}$)**

S_{XY} = {Σ $(X_i - \overline{X})$ $(Y_i - \overline{Y})$} / n ➔ **S_{XY} =** Σ x·y / n ➔ 64/5 = 12,8

S_{XY} = (Σ $X_i \cdot Y_i$ / n) – ($\overline{X} \cdot \overline{Y}$) ➔ **$S_{XY}$ =** (3564 / 5) – (100 · 7) = 12,8 (La más utilizada)

Interpretación ➔ Relación lineal positiva (relación directa entre X e Y). Sin embargo la covarianza presenta el problema de la cuantificación de la correlación. Para superarlo, disponemos del coeficiente de correlación de Pearson.

EL COEFICIENTE DE CORRELACIÓN DE PEARSON ➔ (r_{XY}) = S_{XY} / $S_X \cdot S_Y$

Se trata de la covarianza (**S_{XY}**) dividida entre el producto de las desviaciones típicas de X e Y **($S_X \cdot S_Y$).** En general, indica la fuerza e intensidad de la relación LINEAL entre dos variables cuantitativas (X e Y)

Cálculo ➔ **Puntuaciones directas**:

$$r_{XY} = \frac{n(\Sigma XY) - (\Sigma X)(\Sigma Y)}{\sqrt{n\Sigma X^2 - (\Sigma X)^2}\ \sqrt{n\Sigma Y^2 - (\Sigma Y)^2}}$$

Problema Ejemplo: La siguiente tabla representa las puntuaciones de 6 sujetos en dos tests (X e Y)

SUJETOS ➔	1	2	3	4	5	6
X ➔	8	11	9	9	11	12
Y ➔	38	36	38	42	42	44

Situamos en forma de columna los diferentes elementos para facilitar el cálculo:

X	Y	XY	X^2	Y^2
8	38	304	64	1444
11	36	396	121	1296
9	38	342	81	1444
9	42	378	81	1764
11	42	462	121	1764
12	44	528	144	1936
60	**240**	**2410**	**612**	**9648**

Media $\overline{X}$ = 60/6 =10 Media $\overline{Y}$ = 240/6 = 40

➔ Correlación de Pearson en **puntuaciones directas:**

$$r_{XY} = \frac{n(\Sigma XY) - (\Sigma X)(\Sigma Y)}{\sqrt{n\Sigma X^2 - (\Sigma X)^2}\ \sqrt{n\Sigma Y^2 - (\Sigma Y)^2}}$$

$$r_{XY} = \frac{6(2410) - (60 \cdot 240)}{\sqrt{6 \cdot 612 - 60^2}\ \sqrt{6 \cdot 9648 - 240^2}} = \mathbf{0,42\ (aprox.)}$$

Propiedades de la correlación de Pearson:

- No puede alcanzar nunca un valor menor que (-1) ni mayor que (+1). Cuando las variables no están relacionadas $\mathbf{r_{XY} = 0}$

|----------|----------|
(-1) 0 (+1)

- Las transformaciones lineales de la variable no alteran $\mathbf{r_{XY}}$ **(**$\mathbf{r_{XY} = \pm 1}$**)**

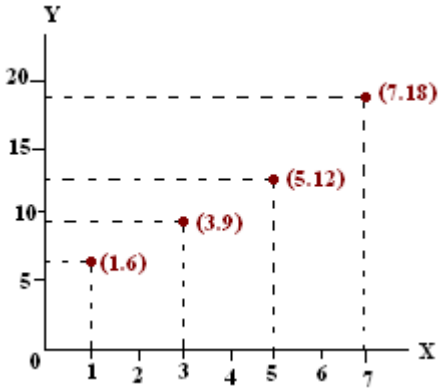
Interpretación: Se consideran el Signo y la cuantía.
Signo: Negativo ($r_{XY} < 0$) Relación lineal inversa // Positivo ($r_{XY} > 0$) Relación lineal directa.
Cuantía: No se pueden establecer normas generales. Para obtener una idea de la importancia cuantitativa de $\mathbf{r_{XY}}$ se debe recurrir al Coeficiente de determinación **(**$\mathbf{r^2_{XY}}$**)** que al multiplicarlo por 100 da un porcentaje de asociación. En nuestro ejemplo ➔ $\mathbf{r_{XY} = 0,42}$
Signo + ($\mathbf{r_{XY} > 0}$**)** Al aumentar o disminuir (X), aumenta o disminuye (Y)
Cuantía $\mathbf{r^2_{XY} =}$ 0,1764 ➔ 17,64 % de varianza común entre X e Y.

Correlación y Causalidad: El coeficiente de correlación de Pearson **nunca indica una relación causal**, únicamente alude a la covariación entre las variables **(relación circular**). Cuando entre dos variables hay una alta correlación y ésta se debe a la presencia de una tercera variable se denomina **correlación espuria.**

3.- <u>ANÁLISIS DE REGRESIÓN LINEAL</u> ➔ El hecho de que dos variables estén relacionadas puede servir para pronosticar o predecir cómo se comportará una variable tomando en consideración cómo se comporta la variable con la que está relacionada. Esto es lo que se consigue a través de los modelos de regresión, que sirven para representar la dependencia lineal de una variable (dependiente), respecto de otra variable (independiente). Ambas variables X e Y han de ser cuantitativas (de intervalo o de razón)

<u>LA RECTA DE REGRESIÓN O PREDICCIÓN</u> ➔ $\mathbf{Y'_i = a + b\ X_i}$

Y'_i = Variable dependiente (valores aleatorios no fijados de antemano)
a = Ordenada en el origen o interceptal (valor que toma Y' cuando X = 0).
b = Pendiente de la regresión (variación en Y cuando X varía 1 unidad).
X_i = Variable independiente o predictora (valores fijados de antemano)



X	Y	X^2	Y^2	XY
3	9	9	81	27
5	12	25	144	60
4	0	16	0	0
7	18	49	324	126
1	6	1	36	6
20	**45**	**100**	**585**	**219**

Datos ➔ $\mathbf{\bar{X} = \sum X / n}$ ➔ **20 / 5** = 4 $\mathbf{\bar{Y} = 45 / 5}$ = 9 $Y'_i = a + b\ X_i$ ➔ $\mathbf{Y'_i = 1,2 + 1,95 \cdot X_i}$

b = S_{XY} / S^2_x ➔ 7,8 / 4 = **1,95** **a =** $\bar{Y} - b\bar{X}$ ➔ 9 – (1,95 · 4) = **1,2**

$\mathbf{S_{XY} = (\sum XY / n) - \bar{X} \cdot \bar{Y}}$ ➔ **(219 / 5) – (4 · 9)** = 7,8 $\mathbf{S^2_X = (\sum X^2 / n) - \bar{X}^2}$ ➔ **(100 / 5) –** $\mathbf{4^2}$ = 4

Habitualmente nos encontramos con una serie de valores entre X e Y que forman una nube de puntos o diagrama de dispersión y, suponiendo una relación lineal entre las variables, estimamos una recta que representa a la nube de puntos (ajuste a la recta de regresión); a partir de ella realizamos las predicciones.

Una vez construida, es el instrumento que permite realizar predicciones de una variable (dependiente) a partir del conocimiento que se tiene de la otra (independiente) con la que está relacionada. Las predicciones que se realizan con la recta de regresión son predicciones promedio, las cuales se aproximan, pero no coinciden con los valores reales (error residual ➔ diferencia entre el pronóstico y el valor real). La predicción, salvo el caso de una relación lineal perfecta (+1 ó –1), no será exacta debido a los errores residuales, por ello llamaremos **Y' a las predicciones e Y a las puntuaciones reales** (la media de las puntuaciones pronosticadas coincide con la media de las puntuaciones reales)

Error residual $Є_i = (Y'_i - Y_i)$ diferencia entre el valor predicho con la recta de regresión y el valor real de Y (**la media de los errores siempre es = 0 y su varianza S^2_{XY}**). Según los datos anteriores: Para (X = 5) → Y´= 1,2 + (1,95 · 5) → Y´= 10,95 (la predicción para un sujeto que obtiene X = 5 es Y´= 10,95. Como puede apreciarse en la tabla, le corresponde una Y = 12. Podemos afirmar que pronosticamos esa puntuación con un error residual de (12 – 10,95 = 1,05).

Cálculo de las Constantes de la Regresión: para construir la recta de regresión debemos averiguar las constantes de la regresión (a y b):

b = PENDIENTE DE LA REGRESIÓN (Procedimientos equivalentes)	
$b = (n \Sigma XY) - (\Sigma X \Sigma Y) / n \Sigma X^2 - (\Sigma X)^2$	(con puntuaciones directas)
$b = r_{XY} \cdot (Sy / Sx)$	(a partir de la correlación de Pearson)
a = ORDENADA EN EL ORIGEN / INTERCEPTAL	
$a = \overline{Y} - b\overline{X}$	(a partir de la pendiente → b)
Aclaraciones → La recta de regresión pasa por el punto ($\overline{X}$, $\overline{Y}$)	

Problema Ejemplo: Algom y otros (1986) en un artículo sobre predicción del dolor ante un sock eléctrico (medido en miliamperios mA) obtuvieron los siguientes resultados:

INTENSIDAD DEL SOCK (Y):	**3,5**	**2,5**	**3,0**	**4,5**	**1,5**
DOLOR ESTIMADO (X):	**9**	**10**	**11**	**13**	**7**

Y	**X**	**XY**	**X^2**	**Y^2**	**$x = X-\overline{X}$**	**$y = Y-\overline{Y}$**	**xy**	**x^2**	**y^2**
1,5	7	10,5	49	2,25	-3	-1,5	4,5	9	2,25
3,5	9	31,5	81	12,25	-1	0,5	-0,5	1	0,25
2,5	10	25	100	6,25	0	-0,5	0	0	0,25
3,0	11	33	121	9	1	0	0	1	0
4,5	13	58,5	169	20,25	3	1,5	4,5	9	2,25
N =15	**n = 50**	**158,5**	**520**	**50**			**8,5**	**20**	**5**

Media de $\overline{X}$ = 50/5 = 10 Media de $\overline{Y}$ = 15/5 =3

(Desv. Típica de X) $Sx = \sqrt{(20/5)} = 2$; (Desviación Típica de Y) $Sy = \sqrt{(5/5)} = 1$

(Covarianza) Sxy = 8,5/5 = 1,7 Correlación Pearson = $r_{XY} = 1,7 / (2 \cdot 1) = 0,85$

b = PENDIENTE DE LA REGRESIÓN (Procedimientos equivalentes)	
$b = (n \Sigma XY) - (\Sigma X \Sigma Y) / n \Sigma X^2 - (\Sigma X)^2$	**b =** $(5 \cdot 520) - (50 \cdot 15) / (5 \cdot 520)^2 - (50)^2$ **= 0,425**
$b = r_{XY} \cdot (Sy / Sx)$	**b =** $0´85 \cdot (1/2)$ **= 0,425**

a = ORDENADA EN EL ORIGEN / INTERCEPTAL	
$\mathbf{a = \bar{Y} - b\bar{X}}$	**a =** $3 - (0{,}425 \cdot 10)$ **= (- 1,25)**
LA RECTA DE REGRESIÓN	
$\mathbf{Y'_i = a + b\ X_i}$	$\mathbf{Y'_i = -1{,}25 + 0{,}425\ X}$

Predicción de valores: Obtenida la recta de regresión, se pueden predecir valores de la variable Y para cada uno de los valores de la variable X. Considerando la ecuación de la recta de regresión de nuestro problema ejemplo, procedamos a realizar predicciones para un sujeto que dice sentir un dolor 8 y pretendemos estimar la intensidad del sock que se le ha aplicado.

En puntuaciones directas: $\mathbf{Y'_i = (- 1{,}25) + 0{,}425\ X}$ **//** $\mathbf{Y'_i = -1{,}25 + 0{,}425\ X \rightarrow\ (-1{,}25) + (0{,}425 \cdot 8) = 2{,}15}$ **mA**

RELACIONES FUNDAMENTALES:

- ➔ La varianza de la variable dependiente $\mathbf{(S^2_Y)}$ es igual a la varianza de las puntuaciones pronosticadas $\mathbf{(S^2_{Y'})}$ más la varianza de los errores residuales $\mathbf{(S^2_{y \cdot x})}$ ➔ $\mathbf{S^2_Y = S^2_{Y'} + S^2_{y \cdot x}}$
  Varianza Total (S^2_Y) = Varianza asociada a X ($S^2_{Y'}$) + Varianza no asociada a X ($S^2_{y \cdot x}$)
- ➔ El cuadrado del coeficiente de correlación de Pearson $\mathbf{(r^2_{xy})}$ se designa habitualmente por (R^2) y se denomina Coeficiente de Determinación (proporción de **varianza explicada** por la regresión lineal)
- ➔ $\mathbf{(1 - r^2_{xy)} = S^2_{yx} / S^2_Y}$ ➔ Es la proporción de la variabilidad de Y (VD) que **no es explicada** por (o no está asociada a) la variabilidad de X (VI) ➔ % Debido al error.

TEMA 5 → NOCIONES BÁSICAS DE PROBABILIDAD

OBJETIVOS DE APRENDIZAJE:

- Conocer los conceptos de experimento aleatorio y espacio muestral.
- Distinguir los distintos tipos de sucesos que forman parte del espacio muestral y las operaciones fundamentales que con ellos pueden realizarse.
- Adquirir un concepto preciso de probabilidad.
- Saber resolver aquellos problemas en que se nos presentan probabilidades condicionadas.

INTRODUCCIÓN: En cualquier investigación intentamos aceptar o rechazar una hipótesis con un determinado nivel de probabilidad y, asimismo, procedemos a generalizar nuestros resultados a la población (Inferencia Estadística) y esta generalización siempre es probabilística. Lo probable (que puede suceder) implica que disponemos de datos suficientes para suponer que determinado suceso ocurrirá. La probabilidad se aplica a la probabilidad de ocurrencia de determinadas proposiciones, juicios o acontecimientos.

1.- CONCEPTOS PREVIOS:

EXPERIMENTO ALEATORIO O ESTOCÁSTICO → Cualquier experimento realizado al azar que se puede repetir indefinidamente en las mismas condiciones y cuyo resultado no se puede predecir con certeza (ejemplo: lanzar una moneda o un dado al aire).

Condiciones básicas de un experimento aleatorio:

- Se puede repetir indefinidamente en las mismas condiciones.
- El resultado de cada ensayo pertenece al conjunto de todos los resultados posibles.
- Antes de cada ensayo no se puede predecir con certeza el resultado que obtendremos.
- Al aumentar el número de ensayos la frecuencia relativa o proporción de cada resultado tiende a aproximarse a un valor fijo.

ESPACIO MUESTRAL → Conjunto formado por todos los resultados posibles de un experimento aleatorio. (E = Universo o población del experimento)
Lanzar un dado: E = {1, 2, 3, 4, 5, 6}; Lanzar una moneda: E = {cara, cruz)

SUCESOS → Los distintos resultados de un experimento aleatorio (subconjunto del espacio muestral)

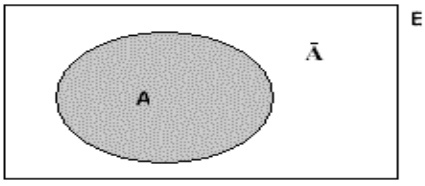
Tipos de sucesos:

- **Elementales o Simples** (implican un solo resultado del espacio muestral E). Lanzar un dado: Obtener un tres (elemental o simple) A = {3}
- **Compuestos** (implican dos o más resultados del espacio muestral). Obtener un número par (compuesto) A = {2, 4, 6}

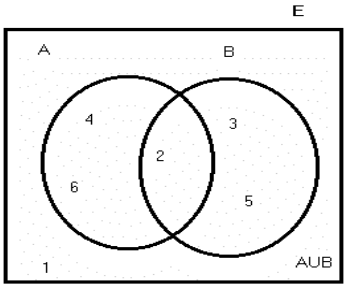
Los sucesos también se pueden clasificar atendiendo a las siguientes características:

- **Suceso Seguro →** Siempre se verifica (Lanzar un dado y obtener puntuación menor que siete)
- **Suceso imposible →** Nunca se verifica (conj. vacío Ø). (Lanzar un dado y obtener diez puntos)

Operaciones con sucesos:
Unión de sucesos A ó B = P (aparezca A ó aparezca B ó ambos a la vez) = P (A U B)
Intersección de sucesos A y B = P (aparezca A y aparezca B) = P (A ∩ B). Cuando no contiene ningún elemento, los sucesos son incompatibles o excluyentes y no pueden ocurrir simultáneamente.
Complementario de A → Subconjunto formado por los sucesos que no pertenecen a A → (No A)

Lanzamos un dado→ Suceso A = {5, 6} Complementario de $\bar{A}$ = {1, 2, 3, 4}	

Problema Ejemplo: Lanzar un dado es una experiencia aleatoria (nunca podemos asegurar el valor que se obtiene al lanzarlo). El conjunto de los resultados posibles constituye el Espacio Muestral.



Espacio Muestral → E = {1, 2, 3, 4, 5 y 6}

Sucesos → A = {cifra par} A **U** B = {2, 3, 4, 5, 6}
B = {número primo} A ∩ B = {2}
Complementario de $\overline{A}$ = {1, 3, 5}

A y B son **Compatibles** → A ∩ B = {2}

Problema Ejemplo: Dados los conjuntos A, B y C (resultados del lanzamiento de un dado) determinar los elementos que pertenecen a cada una de las expresiones siguientes:

A = {1, 2, 3, 4} B = {2, 3, 5} C = {1, 3, 5, 6}

(A ∩ C) ∩ $\overline{B}$ = {1, 2, 3, 4} ∩ {1, 3, 5, 6} ∩ {1, 4, 6} = {1, 3} ∩ {1, 4, 6} = **{1}**

(A ∩ C) - B = {1, 3} - {2, 3, 5} = **{1}** (Suprimo en A ∩ C los sucesos de B)

C ∩ ($\overline{A \cup B}$) = {1, 3, 5, 6} ∩ {6} = **{6}**

C - (A U B) = {1, 3, 5, 6} - {1, 2, 3, 4, 5} = **{6}**

2.- DEFINICIÓN DE PROBABILIDAD (ESPACIO MUESTRAL DISCRETO)

Se consideran tres definiciones. En cualquiera de las tres, la probabilidad se cuantifica como un número comprendido entre 0 y 1 → 0 para el suceso imposible // 1 para el suceso seguro // cualquier otro suceso, asignando un número entre 0 y 1, en función de la cuantía de su probabilidad de ocurrencia.

DEFINICIÓN CLÁSICA (LAPLACE / A PRIORI): P (A) = n_A / N → La probabilidad de un suceso A es el cociente entre el número de casos favorables y el número de casos posibles (suponemos que todos los sucesos tienen la misma probabilidad de ocurrencia → Equiprobabilidad)
El gran problema es aplicarla a sucesos cuya condición de equiprobabilidad no está garantizada. Una salida es aplicar la Definición Estadística.

DEFINICIÓN ESTADÍSTICA (RICHARD VON MISES / A POSTERIORI): P (A) = lim N→ ∞ n_A / N → Límite al que tiende la frecuencia relativa de aparición de un suceso A cuando el número de ensayos (N) tiende a infinito. Von Mises (cuando se realiza un experimento aleatorio un número suficiente de veces, la frecuencia de aparición de un determinado suceso tiende a aproximarse a un valor constante que es la probabilidad de aparición de ese suceso)
Al igual que la anterior tampoco es una definición satisfactoria. A veces no es posible, ni práctico, repetir un experimento aleatorio un gran número de veces (no es posible aproximarse a infinito). Por ello los matemáticos cambiaron el rumbo de su pensamiento, lo que originó la Definición Axiomática.

DEFINICIÓN AXIOMÁTICA (KOLMOGOROV) → Dado un espacio muestral E, llamamos probabilidad de un suceso A, designado P (A) a un número real que asignamos al suceso A, tal que cumple los siguientes axiomas (afirmaciones que se aceptan sin demostración):

AXIOMAS	
0 ≤ P (A) ≥ 1 **P (E) = 1**	Las dos propiedades indican que la probabilidad se cuantifica mediante un número positivo comprendido entre 0 y 1 (cero cuando no puede ocurrir nunca y uno cuando ocurre seguro)
P (A) = 1 – P ($\overline{A}$)	La probabilidad de A puede obtenerse restando de 1 la probabilidad de su complementario (ambos son exhaustivos y mutuamente excluyentes)

Problema ejemplo → Lanzamos dos monedas y consideramos el suceso:

A = {Probabilidad de obtener 1 cara y 1 cruz}; Espacio Muestral **E = {cc, cx, xc, xx}**

P (A) + P ($\bar{A}$) = 1 → P (A) = 2/4 = 1/2 = **0, 5 → P ($\bar{A}$) =** 2/4 = 1/2 = **0, 5**

En la práctica se aplica la **definición clásica** → Probabilidad de un suceso = $\frac{\text{Nº Casos Favorables}}{\text{Nº Casos Posibles}}$

3.- TEOREMAS DE PROBABILIDAD: Los Teoremas de Probabilidad permiten calcular probabilidades de sucesos que reúnen una serie de condiciones (dependientes, independientes, mutuamente excluyentes,...). La aplicación de los teoremas más importantes de la probabilidad de sucesos requieren distinguir, previamente, entre las características de dichos sucesos:

TEOREMA DE LA SUMA: Permite resolver preguntas del tipo → ¿Cuál es la probabilidad de que aparezca un suceso o el otro? La letra (o) implica alternativa entre dos opciones (unión de sucesos).

(A ó B) Sucesos mutuamente excluyentes → P (A U B) = P (A ó B) = P (A) + P (B)
Los sucesos mutuamente excluyentes o incompatibles no pueden ocurrir simultáneamente. La aparición de uno excluye al otro (obtener cara y cruz en el mismo lanzamiento **P (A U B) =** (conjunto vacío Ø) **= 0.**

Problema ejemplo → En una bolsa hay 15 bolas rojas y 10 verdes. Extraemos una bola de la bolsa. Cuál es la probabilidad de que sea roja o verde → P (Roja) = 15 / 25 y P (Verde) = 10 / 25. Se trata de sucesos excluyentes (si es roja no puede ser verde)

P (Roja U Verde) = P (Roja ó Verde) = (15 / 25) + (10 / 25) = 25 / 25 **= 1**

(A ó B) Sucesos simultáneos → P (A U B) = P (A) + P (B) – P (A ∩ C)
Los sucesos simultáneos o compatibles son los que pueden ocurrir a la vez (ejemplo: ser hombre y conducir) → **P (A U B) ≠** (conjunto vacío Ø).

Problema ejemplo → A los 70 años la probabilidad de ser miope es 0,3, la de tener cataratas es 0,15 y la de ser miope y tener cataratas es 0,1. La probabilidad de ser miope o tener cataratas a esta edad sería:
Ser miope (M) y Tener cataratas (C) son sucesos compatibles (simultáneos) y su probabilidad de ocurrencia conjunta es 0,1 **→ P (M U C) =** P (M) + P (C) – P (M ∩ C) = 0,3 + 0,15 – 0,1 **= 0,35**

TEOREMA DEL PRODUCTO (PROBABILIDAD CONDICIONADA P (A/B): La aparición de uno de los sucesos depende de la aparición del otro. Dados dos sucesos (A y B), se llama probabilidad de A condicionada por B, a la intersección, dividida por la probabilidad de la condición B y viceversa:

P (A / B) = P (A ∩ B) / P (B) Supuesto P (B) ≠ 0
P (B / A) = P (A ∩ B) / P (A) Supuesto P (A) ≠ 0

(A y B) Sucesos dependientes → P (A ∩ B) = P (A) · P (B / A)
Por tanto la probabilidad de que se presenten simultáneamente **P (A ∩ B)** se obtiene despejando el numerador del teorema del producto. En las extracciones sin reposición los sucesos son siempre dependientes.

Problema ejemplo → En un colectivo profesional formado a partes iguales por ambos sexos, el estrés afecta a un 35% de los hombres y a una de cada cuatro mujeres. Elegida al azar una persona, cuál es la probabilidad de que tenga estrés y la probabilidad de que no padeciendo estrés sea mujer.

E = Padecer estrés → $\bar{E}$ = No padecer estrés; H → Hombre M → Mujer

P (H) = P (M) = 0,5 P (E / H) = 0,35 P (E / M) = 0,25

$P(\bar{E} / M) = 1 - P(E / M) = 1 - 0,25 = 0,75$

Elegida al azar una persona, cuál es la probabilidad de que tenga estrés →

$P(E) = P(E \cap H) + P(E \cap M) =$ $P(H) \cdot P(E / H) + P(M) \cdot P(E / M) = (0,5 \cdot 0,35) + (0,5 \cdot 0,25)$ **= 0,3**

Probabilidad de que no padeciendo estrés sea mujer → Hemos averiguado la probabilidad de padecer estrés P (E) = 0,3, luego la probabilidad de no padecer estrés será:

$P(\bar{E}) =$ $1 - P(E) = 1 - 0,3$ **= 0,7**

$P(M / \bar{E}) =$ $P(M \cap \bar{E}) / P(\bar{E}) \rightarrow P(M) \cdot P(\bar{E} / M) / P(\bar{E})$

$P(M / \bar{E}) =$ $(0,5 \cdot 0,75) / 0,7$ **= 0,536**

(A y B) Sucesos independientes → $P(A \cap B) = P(A) \cdot P(B)$
La probabilidad de que se produzca uno de ellos no depende del otro. En las extracciones con reposición los sucesos son siempre independientes.

Problemas ejemplo → Se lanzan dos monedas y analizamos las siguientes probabilidades**:**
Espacio muestral = {cc, cx, xc, xx}

Obtener dos caras → P (2 caras) = (1 / 2) · (1 / 2) = 1 / 4 **= 0,25**
Obtener dos cruces → P (2 cruces) = (1 / 2) · (1 / 2) = 1 / 4 **= 0,25**
Obtener al menos una cara → P (al menos 1 cara) = 1 – P (ninguna cara) = 1 – 1 / 4 = 3 / 4 **= 0,75**

Un examen tipo test consta de tres preguntas, todas ellas con cuatro alternativas de respuesta de las que sólo una es correcta. Si un alumno responde al azar cuál es la probabilidad de que acierte las tres preguntas.

$P(A \cap B \cap C) =$ $P(A) \cdot P(B) \cdot P(C) = (1 / 4) \cdot (1 / 4) \cdot (1 / 4) = 1 / 64$ **= 0, 0156**

<u>TEOREMA DE BAYES</u> → $P(B_K / A) = P(B_K) \cdot P(A / B_K) \; // \; \Sigma P(B_I) \cdot P(A / B_I)$
También se denomina probabilidad de las causas. Si llamamos B_K a las causas, podemos determinar la probabilidad de que el suceso A (consecuencias), se haya realizado bajo cada una de las causas B_K.

$P(B_K)$ → Probabilidades a priori **(conocidas)**
$P(A / B_K)$ → Probabilidades condicionadas o verosimilitudes **(conocidas)**
$P(B_K / A)$ → Probabilidades a posteriori (**desconocidas** por lo que debemos calcularlas)

Problema ejemplo → En una sala de un hospital hay tres pacientes con la enfermedad B_1, dos pacientes con la enfermedad B_2 y cinco pacientes con la enfermedad B_3. Sabemos que la probabilidad de curación completa (C) de la enfermedad es: para la enfermedad B_1 = 1/3; para la enfermedad B_2 = 2 / 3 y para la enfermedad B_3 = 1 / 7

a) ¿Cuál es la probabilidad total de curarse?: (Nº total de enfermos → 10)

Probabilidades A priori	**Probabilidades Condicionadas**	**Probabilidades Condicionadas**
$P(B_1)$ = 3/10 (0,3) **$P(B_2)$** = 2/10 (0,2) **$P(B_3)$** = 5/10 (0,5)	**$P(C/B_1)$** = 1/3 (0,333) **$P(C/B_2)$** = 2/3 (0,667) **$P(C/B_3)$** = 1/7 (0,143)	**$P(NC/B_1)$** = 2/3 (0,667) **$P(NC/B_2)$** = 1/3 (0,333 **$P(NC/B_3)$** = 6/7 (0,857)

Teorema de la probabilidad total (denominador del teorema Bayes)

$\Sigma P(B_I) \cdot P(A / B_I)$ → Probabilidad total de curarse

Probabilidades a priori (multiplicadas por) probabilidades condicionadas

$P(C) = P(B_1) \cdot P(C / B_1) + P(B_2) \cdot P(C / B_2) + P(B_3) \cdot P(C / B_3)$

P (C) = (0,3 · 1/3) + (0,2 · 2/3) + (0,5 · 1/7) = **0,305**

Un enfermo fue dado de alta sano ¿cuál es la probabilidad de que sufriese la enfermedad B_3?:

$$P(B_3 / C) = \frac{P(B_3) \cdot P(C / B_3)}{P(B_1) \cdot P(C/B_1) + P(B_2) \cdot P(C/B_2) + P(B_3) \cdot P(C/B_3)}$$

$$P(B_3 / C) = \frac{0,5 \cdot 1/7}{(0,3 \cdot 1/3) + (0,2 \cdot 2/3) + (0,5 \cdot 1/7)} = \mathbf{0,234}$$

<u>Problema ejemplo (repaso de los conceptos expuestos en el tema)</u>: Sabemos que el 85 % de los enfermos depresivos se recupera con un determinado tratamiento psicológico. De los sujetos que se recuperan, el 10 por 100 habían sido tratados médicamente con anterioridad. De los que no se recuperan el 25 % también habían sido tratados médicamente con anterioridad:

Probabilidades A priori	**Probabilidades Condicionadas**	**Probabilidades Condicionadas**
P (R) = 0,85 **P (NR)** = 0,15	**P (T/R)** = 0,10 **P (T/NR)** = 0,25	**P (NT/R)** = 0,90 **P (NT/NR)** = 0,75

P (R) = probabilidad de que un enfermo se recupere con el tratamiento
P (NR) = probabilidad de que un enfermo no se recupere con el tratamiento
P (T/R) = probabilidad de que un enfermo (que se ha recuperado) haya sido tratado médicamente
P (NT/R) = probabilidad de que un enfermo (que se ha recuperado) no haya sido tratado médicamente
P (T/NR) = probabilidad de que un enfermo (que no se ha recuperado) haya sido tratado médicamente
P (NT/NR) = probabilidad de que un enfermo (que no se ha recuperado) no haya sido tratado médicamente

a) Elegido al azar un enfermo depresivo, ¿cuál es la probabilidad de que se haya recuperado con el tratamiento? →**P(R) = 0'85** (definición clásica, página 23 formulario)

b) Elegido un enfermo depresivo al azar que se haya recuperado, ¿cuál es la probabilidad de que no haya sido tratado médicamente con anterioridad? → **P (NT/R) = 0'90** (probabilidad condicionada).

c) Elegido un enfermo al azar, ¿cuál es la probabilidad de que haya sido tratado médicamente y no se haya recuperado con el tratamiento psicológico? →**P (NR y T) = P(NR) · P(T/NR)** = 0'15 · 0'25 **= 0'0375** (teorema del producto)

d) Elegido un enfermo al azar, ¿cuál es la probabilidad de que haya sido tratado médicamente? →
P(T) = P(R) · P(T/R) + P(NR) · P(T/NR) = 0'85 · 0'10 + 0'15 · 0'25 = **0'1225** (teorema probabilidad total)

e) Elegido un enfermo al azar resulta que ha sido tratado médicamente con anterioridad, ¿cuál es la probabilidad de que se recupere con el tratamiento psicológico?:

$$P(R/T) = \frac{P(R \text{ y } T)}{P(T)} = \frac{P(R) \cdot P(T/R)}{P(T)} = \frac{0'85 \cdot 0'10}{0'1225} = \mathbf{0'6939}$$ (Teorema de Bayes)

TEMA Nº 6 → DISTRIBUCIONES DISCRETAS DE PROBABILIDAD

OBJETIVOS DE APRENDIZAJE:

- Ser capaz de definir correctamente una o más variables aleatorias sobre los resultados de un experimento aleatorio y determinar los valores que toma una determinada variable aleatoria previamente definida.
- Conocer las propiedades que deben cumplir la función de probabilidad y de distribución de una variable aleatoria discreta.
- Obtener la función de probabilidad y la función de distribución de una variable aleatoria discreta y realizar su representación gráfica.
- Calcular e interpretar la media y la varianza de una variable aleatoria discreta.
- Conocer las condiciones de aplicación de la distribución binomial, su media y su varianza.
- Manejar las tablas de la distribución binomial para la resolución de problemas concretos.

1.- VARIABLE ALEATORIA (VA) ➔ Se trata de un conjunto de números diferentes que se asignan de forma específica a cada uno de los resultados posibles del experimento aleatorio como consecuencia de aplicar una función o regla de asignación (se construye un modelo de distribución de probabilidad).

Definición (VA) = Función o regla que asigna un número real, y sólo uno, a cada uno de los resultados de un experimento aleatorio (a cada suceso del espacio muestral (E).

TIPOS DE VARIABLES ALEATORIAS:

→ **V. Aleatoria discreta:** Cuando toma un número finito de valores (casos posibles susceptibles de ser contados). Entre dos valores consecutivos no existen valores intermedios. Ejemplos: número de hijos de determinadas familias, número de asignaturas de primer curso, etc. La distribución discreta más importante es la Binomial.

→ **V. Aleatoria continua:** Cuando puede tomar cualquier valor numérico de un conjunto infinito (casos posibles no numerables). Entre dos valores podemos encontrar infinitos valores intermedios. Ejemplos: Tiempo, CI, etc. Los modelos de distribución continua más importantes son: Distribución Normal Tipificada, la Distribución Chi-Cuadrado de Pearson, la Distribución "t" de Student y la Distribución "F" de Snedecor.

2.- CONCEPTOS BÁSICOS (VV AA Discretas): Supongamos que el director de un gabinete de psicología clínica tiene tres mujeres y dos hombres trabajando con él. Desea escoger a **dos personas** para un trabajo especial de selección. A fin de no introducir sesgos en su elección, decide escogerlos al azar, sucesivamente y sin reposición. Consideramos M (Mujer) y H (Hombre).

Llamamos X a la Variable aleatoria = {Número de mujeres seleccionadas}
Sucesos del Espacio muestral → E = {MM; HM; MH; HH}
Probabilidades (Casos favorables / Casos Posibles):

3 M → P (M) = 3/5 → 0,6 Probabilidad de elegir una Mujer
2 H → P (H) = 2/5 → 0,4 Probabilidad de elegir un Hombre

Función de probabilidad ➔ De una variable discreta X, y se representa por f (x), a aquella función que asocia a cada valor de la variable la probabilidad de que ésta adopte ese valor **[f (x) = P (X = x)]**
La función de probabilidad de una variable aleatoria es la definición de su comportamiento matemático. Supone calcular la probabilidad asociada a cada elemento del Espacio muestral.
En nuestro ejemplo: E = {M ∩ M; H ∩ M; M ∩ H; H ∩ H}

Ninguna mujer: **f (0) =** P (X = 0) = **P (H ∩ H)** → P (H) · P (H / H) = (2/5) · (1/4) = 2/20 **= 0,1**
Una mujer: **f (1)** = P (X = 1) → **P (H ∩ M) U P (M ∩ H)** → P (H) · P (M / H) + P (M) · P (H / M) = (2/5) · (3/4) + (3/5) · (2/4) = 12/20 **= 0,6**
Dos mujeres → **f (2) =** P (X = 2) = **P (M ∩ M) → P (M) · P (M / M) =** (3/5) · (2/4) = 6/20 **= 0,3**

→ Representación de la Función de Probabilidad

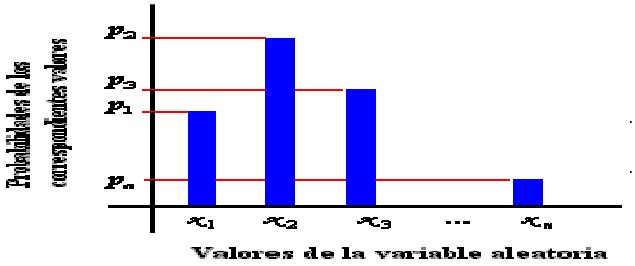
x	0	1	2
f (x)	0´1	0´6	0´3

x	f (x)
0	0´1
1	0´6
2	0´3

Propiedades de la función de probabilidad:

Para cualquier valor de x, siempre toma valores positivos o nulos → **∀ x ε X f (x) ≥ 0**
La suma de todas las probabilidades correspondientes a cada valor de x es igual a 1 →
$\sum$ f(x) = f(x_1) + f(x_2) + f(x_3) +…+ f(x_n) = 1

Representación Gráfica: Para variables aleatorias discretas adopta la forma de un diagrama de barras, con los valores de la variable en el eje de abscisas (horizontal) y las probabilidades de cada valor en el eje de ordenadas (vertical).



<u>Función de Distribución</u> → Supone calcular la probabilidad de que la variable aleatoria X tome un valor menor o igual que un valor concreto de x. Se obtiene acumulando (sumando) los valores de la Función de Probabilidad. Se representa por **F (x) = P (X ≤ x).** La suma de probabilidades debe ser uno.

Siguiendo el ejemplo del gabinete de psicología (Función de Distribución):

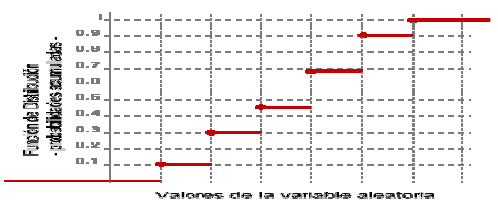
x	0	1	2
F (x)	0´1	0´7	1

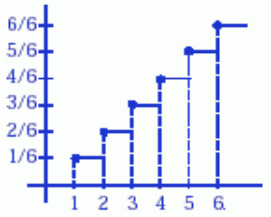
x	F (x)
0	0´1
1	0´7
2	1

Propiedades de la función de distribución:

Todos los valores de la función de distribución son positivos o nulos→ **∀ x F (x) ≥ 0**
La función de distribución es igual a 0 para todo valor fuera del límite inferior e igual a 1 para todo valor fuera del límite superior de la variable aleatoria → **F (x) = 0** (Si x < a) y **F (x) = 1** (Si x > b) .
La función **F (x) es no decreciente** (al ir acumulando probabilidades)
La probabilidad de que la variable aleatoria tome valores entre x_1 y x_2 es la diferencia entre el valor superior y el inferior → **P (x_1 ≤ x ≤ x_2) = F (x_2) - F (x_1)**

Representación Gráfica: Para la función de distribución





Problema Ejemplo → En un concurso de tiro al plato, un concursante dispara dos veces consecutivas. La probabilidad de acertar el primer disparo es 0,60 y el segundo 0,80. Si el participante no acierta ningún disparo debe pagar 2000 €. Si acierta uno de los dos gana 100 €. Si acierta los dos gana 200 €. Calcular la función de probabilidad y la función de distribución de la variable aleatoria X "euros ganados por el participante".

	Acierto	**Fallo**	P (X = - 2000) = (0,4 · 0,2) **= 0,08** P (X = 100) = (0,6 · 0,2) + (0,4 · 0,8) **= 0,44** P (X = 200) = (0,6 · 0,8) **= 0,48**
Primer disparo	**P (0,6)**	**P (0,4)**	
Segundo disparo	**P (0,8)**	**P (0,2)**	

x	- 20000	100	200
f (x)	0´08	0´44	0´48

Función de Probabilidad

x	- 20000	100	200
F (x)	0´08	0´52	1

Función de Distribución

Media y Varianza de una variable aleatoria:

Media, esperanza matemática o valor esperado de X → Promedio teórico que tomaría la variable aleatoria si se repitiera el experimento aleatorio infinitas veces**.**

Se representa por → **E (X) = Σ x · f (x)** Suma de los productos de cada uno de los valores, x, que toma la variable aleatoria, por sus respectivas probabilidades, f (x).

Varianza → Esperanza matemática de los cuadrados de las diferencias entre los valores de la variable y la media. Se designa con la letra griega σ^2 o con la expresión **V (X)**

Se representa por → $\sigma^2 = E(X^2) - [E(X)]^2$ Esperanza de los cuadrados de X {E (X^2)}, menos el cuadrado de la esperanza de X $[E(X)]^2$

En consecuencia, la **Desviación típica** será: $\sigma = \sqrt{E(X^2) - [E(X)]^2}$

Problema ejemplo: La primera prueba presencial de una determinada asignatura consta de dos problemas (A y B). Supongamos que es obligatorio responder a los dos problemas. Las probabilidades de responder correctamente a cada uno de ellos es respectivamente: 0,7 y 0,4. Suponiendo que las respuestas dadas a los problemas son independientes, definimos la variable aleatoria X = {Número de problemas resueltos correctamente}

Función de Probabilidad y Función de Distribución de la variable X.

P (A) = 0,7 y P (B) = 0,4 son las probabilidades de responder correctamente los problemas (A y B)

P (X = 0) = P ($\overline{A}$) · P ($\overline{B}$) = 0,3 · 0,6 **= 0,18** (No responder correctamente ninguno)

P (X = 1) = P ($\overline{A}$) · P (B) + P (A) · P ($\overline{B}$) = 0,3 · 0,4 + 0,7 · 0,6 **= 0,54** (Responder correctamente uno)

P (X = 2) = P (A) · P (B) = 0,7 · 0,4 **= 0,28** (Responder correctamente los dos)

X	**0**	**1**	**2**
f (X)	0,18	0,54	0,28
F (X)	0,18	0,72	1

Media de la variable X → E (X) = Σ x · f (x) → (Cada suceso por su probabilidad de ocurrencia)

E (X) = (0 · 0,18) + (1 · 0,54) + (2 · 0,28) **= 1,1**

Varianza y desviación Típica de la variable X

Varianza de la variable X → $\sigma^2 = E(X^2) - [E(X)]^2$ → (1,66) - (1,21) **= 0,45**

$E(X^2) = E X^2 \cdot f(x)$ **=** $(0^2 \cdot 0{,}18) + (1^2 \cdot 0{,}54) + (2^2 \cdot 0{,}28) =$ **1,66** $[E(X)]^2 = (1{,}1)^2 =$ **1,21**

Desviación Típica = $\sqrt{0{,}45}$ **= 0,67**

3.- <u>DISTRIBUCIONES DISCRETAS DE PROBABILIDAD:</u> Estos modelos tienen una complejidad matemática menor que los modelos de distribución continua. Las probabilidades de variables aleatorias discretas se pueden calcular a partir de sus expresiones matemáticas o con ayuda de las tablas (debemos manejar ambos procedimientos).
Trabajaremos con distribuciones discretas que sólo pueden tomar dos valores (dicotómicas) → (sucesos tipo Bernouilli) y que habitualmente denominaremos 1 (éxito) y 0 (fracaso o error). La más importante es la Distribución Binomial.

<u>DISTRIBUCIÓN BINOMIAL</u> ➔ B (n, p)

Definición → Probabilidad de obtener en N ensayos (tipo Bernouilli) un número determinado (x) de éxitos. La Distribución Binomial depende de los valores que tome N (número de ensayos) y p (probabilidad de éxito)

Características:
→ Se trata de N ensayos independientes tipo Bernouilli
→ Cada ensayo tiene dos posibles resultados que se representan por 0 y 1.
→ La probabilidad p, permanece constante en cada ensayo.

Parámetros → Media **= E (X) = N · p** Varianza **= σ^2 = N · p · q** Desviación Típica = **$\sigma = \sqrt{N \cdot p \cdot q}$**

Función de Probabilidad → La variable aleatoria es "nº de éxitos en N ensayos" (N es fijo, y "x" es variable). La función de probabilidad nos permite calcular la probabilidad de que en N ensayos aparezcan "x" éxitos. El número combinatorio **"n sobre x"** es igual a: **N ! / x ! (N – x) !**

$$f(x) = P(X = x) = \binom{N}{x} p^x \cdot q^{N-x}$$

Además de la fórmula expuesta, las probabilidades pueden obtenerse con la **Tabla I** de las páginas 21 a 25 del formulario, para n ≤ 20 y algunos valores de p ≤ 0'50. Permite determinar la probabilidad de que en N ensayos independientes aparezca x veces el suceso A (suceso favorable o éxito)

Función de Distribución → $F(x) = P(X \le x) = \sum \binom{N}{x} p^x \cdot q^{N-x}$

Se pueden utilizar la **Tabla II** de las páginas 26 a 31 del formulario para calcular directamente la función de distribución.

<u>Problemas Ejemplo</u> → Un niño lanza al aire una moneda imparcial en 10 ocasiones y recibe un caramelo cada vez que sale cara. Calcular:

a) ¿Cuál es la probabilidad de que obtenga 4 caramelos (cuatro caras)?:

N = 10 y p = 0,5 → **Distribución Binomial** = **B (10, 0,5)**

$P(X = 4) = \binom{10}{4} 0{,}5^4 \cdot 0{,}5^6 = [(10!) / (4!) \cdot (6!)] \cdot (0{,}5)^4 \cdot (0{,}5)^6 =$ **0,205**

Utilizando la **Tabla I,** con **B (10, 0,5)** → P(X = 4) = **0,2051**

b) ¿Cuál es la probabilidad de que obtenga más de 5 y menos de 8 caramelos?:

P (5 < X < 8) = [P (X = 6) + P (X = 7)] **= 0,322**

$\binom{10}{6} 0,5^6 \cdot \square 0,5^4 = [(10!) / (6!) \cdot (4!)] \cdot (0,5)^6 \cdot (0,5)^4 =$ **0,205**

$\binom{10}{7} 0,5^7 \cdot 0,5^3 = [(10!) / (7!) \cdot (3!)] \cdot (0,5)^7 \cdot (0,5)^3 =$ **0,117**

Utilizando la **Tabla I,** con **B (10, 0,5)** → 0,2051 + 0,1172 **= 0,3223**

c) Número más probable de caramelos que obtendrá (valor esperado / media).

Esperanza matemática **E (X) = N · p** = 10 · 0,5 **= 5**

Se sabe que la probabilidad de que una rata aprenda a elegir el lado izquierdo de un laberinto en forma de T, donde se encuentra la comida, va creciendo a medida que aumenta el número de ensayos de la siguiente manera:

ENSAYO	1	2	3	4	5
PROBABILIDAD	0,5	0,7	0,8	0,9	1

Si colocamos en la salida del laberinto a diez ratas:

a) ¿Cuál es la probabilidad de que más de 4 ratas elijan el camino adecuado en el primer ensayo?: N = 10 y p = 0,5 → **Distribución Binomial =** B (10, 0,5)

P (X > 4) = 1 – P (X ≤ 4) = 1 - [f (0) + f (1) + f (2) + f (3) + f (4)]

Utilizando la **Tabla II** con **B (10, 0,5)** → P(X **≤** 4) **= 0,3770 →** Solución → 1 - 0,3770 **= 0,6230**

b) ¿Cuál es la probabilidad de que más de 2 y menos de 5 elijan el camino erróneo en el segundo ensayo?: P (Ensayo 2 → camino correcto = 0,7; camino erróneo = 0,3)

P (2 < X < 5) = [P (X = 3) + P (X = 4)]

$\binom{10}{3} 0,3^3 \cdot \square 0,7^7 + \binom{10}{4} 0,3^4 \cdot \square 0,7^6 = 0,267 + 0,2 =$ **0,467**

Utilizando la **Tabla I,** con **B (10, 0,3)** → 0,2668 + 0,2001 **= 0,4669 → 0,467**

c) ¿Cuál es la probabilidad de que las 10 ratas elijan el camino correcto en el tercer ensayo?: P (camino correcto → ensayo 3) = 0,8. Como no existe en las Tablas la probabilidad (0,8), se razona aplicando "que ninguna rata elija el camino erróneo" y utilizamos la probabilidad (0,2).

$$P (X = 10) \rightarrow \binom{10}{10} 0,8^{10} \cdot 0,2^0 =$$ **0,1074**

Tabla I, con B (10, 0,2) → P (X = 0) **= 0,1074**

TEMA Nº 7 → DISTRIBUCIONES CONTINUAS DE PROBABILIDAD

OBJETIVOS DE APRENDIZAJE:

- Conocer las características de la distribución normal como distribución de probabilidad de una variable y la aproximación de la binomial a dicha distribución.
- Utilizar las tablas de la distribución normal para obtener probabilidades.
- Conocer las características de las distribuciones t, chi-cuadrado y F: su media y varianza.
- Utilizar las tablas de las distribuciones t, chi-cuadrado y F, para obtener probabilidades asociadas a unos determinados valores e, inversamente, obtener los valores de estas variables asociados a unas determinadas probabilidades.

INTRODUCCIÓN: La dificultad a la hora de describir la función de probabilidad de una variable continua es que no podemos construir una tabla incluyendo los valores que toma la variable, debido a que el número de valores que puede tomar es, por su propia definición, infinito.

Las funciones de densidad de probabilidad de los modelos de distribución continua se definen mediante ecuaciones o fórmulas matemáticas concretas, y se utilizan para calcular las probabilidades asociadas a los intervalos de valores de dichas distribuciones (la función de densidad de probabilidad para un valor concreto en una distribución continua es cero). En todos los casos, prescindimos de las ecuaciones que representan las funciones de probabilidad, podemos servirnos de las tablas para sustituir estos cálculos.

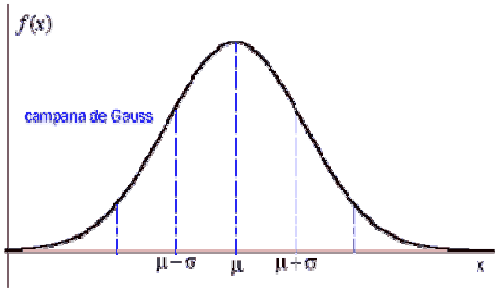
Cada una de las variables continuas se definen en función de otras variables y todas ellas convergen, en algún momento, con la Distribución Normal:

Chi-Cuadrado (varias variables con Distribución Normal)

t de Student (dos variables: una con Distribución Normal y otra Chi-Cuadrado)

F de Snedecor (dos variables con distribución Chi-Cuadrado)

1.- DISTRIBUCIÓN NORMAL (CAMPANA DE GAUSS): Se obtiene a partir del Histograma y el polígono de frecuencias considerando un número amplio de casos y una pequeña amplitud de los intervalos que lo conforman.



Propiedades básicas:

→ Es **simétrica** y coinciden la Media, la Mediana y la Moda. Divide a la distribución en dos partes iguales, cada una comprende el 50 % de las observaciones (proporción = 0,5).
→ Es **asintótica** respecto al eje de las X, se acerca a la abscisa pero nunca la toca (se extiende desde - ∞ a +∞).
→ Tiene **dos puntos de inflexión** (la curva pasa de ser cóncava a convexa) situados a una desviación típica por encima y por debajo de la media. El punto más alto de la curva es la media.
→ Si aplicamos una **transformación lineal** a una variable distribuida normalmente, la transformación lineal también se distribuye normalmente. Cuando se trata de la suma o la resta de dos variables distribuidas normalmente, la suma o la resta también se distribuye según la curva normal.

$\overline{Y} = b \cdot \overline{X} + a$ $\quad \overline{Y} = \overline{X}_1 \pm \overline{X}_2$ $\quad S_Y = |b| \cdot S_X$ $\quad S_Y = \sqrt{S^2_1 + S^2_2}$

Tipificación: Para cada valor (μ y σ) tendremos curvas normales distintas, con Puntuaciones Típicas (μ = 0 y σ = 1). Se denomina unitaria porque el área que comprende es uno.

Distribución Normal estándar o tipificada Z ➔ **N (0,1)**

Tipificar → Convertir una puntuación directa en típica mediante la fórmula: $\mathbf{Z = (X_i - \overline{X}) / S_X}$

Áreas y proporciones bajo la curva normal:

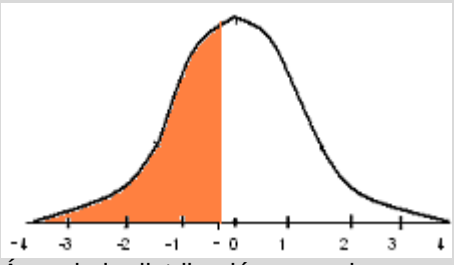
→ Las **Tablas III y IV** del formulario recogen la función de distribución de la Curva Normal. Columna (Z): Puntuaciones Típicas (unidades y décimas), en la fila superior centésimas. Cada valor Z deja por debajo un área, frecuencia, probabilidad o porcentaje según se expresen sobre 1 ó sobre 100.

→ Los valores **Z positivos** (superiores a la media: zona a la derecha de la curva). Los valores **Z negativos** (inferiores a la media: zona a la izquierda de la curva). Los **valores interiores** indican áreas bajo la curva (probabilidades, frecuencias, porcentajes y proporciones)

Problemas ejemplo:

1.- Las puntuaciones en una determinada asignatura (X) de un grupo de 500 niños se distribuyen normalmente con media 6 y desviación típica 2. Queremos determinar el número de niños que obtiene puntuaciones menores o iguales a 5´5.

Primero, se calcula la puntuación típica correspondiente a **5,5** (esta puntuación será menor que cero, negativa, puesto que la puntuación directa 5,5 es menor que la media = 6) → Z = (5,5 – 6) / 2 = (- 0,25). Según la Tabla III → proporción que corresponde a la puntuación típica **(- 0,25):** buscamos en la columna (z) el valor (- 0.2) y en la primera fila el valor 0´05. En el cruce de la fila con la columna encontramos 0´4013, que es la proporción buscada **(0,4013).**



Área de la distribución normal
bajo la puntuación z = (- 0.25)

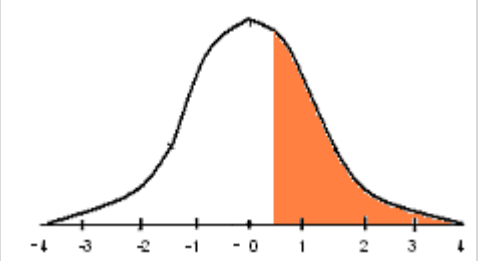
z	0	1	.....	5	....	9
-3.5						
-3.4						
.						
.						
.						
-0.2				. 4013		
-0.1						
-0.0						

La proporción de casos con puntuaciones típicas menores o iguales a (- 0.25) vale 0´4013. Si multiplicamos esta proporción por 100 obtenemos el porcentaje (40´13%). Mediante una regla de tres, obtenemos el número de niños que han obtenido puntuaciones típicas menores o iguales que (- 0.25):

$$\left.\begin{array}{l} 500 \rightarrow 100\% \\ X \rightarrow 40´13\% \end{array}\right\} \Rightarrow x = \frac{500 \cdot 40´13}{100} = 200´65 \cong \textbf{201 niños}$$

2.- La distribución de la variable "peso al nacer" tiene una media de 3.400 gr. ($\bar{X}$ = 3.400) y una desviación típica de 400 gr. (S_X = 400). Suponiendo que esta distribución es normal, resolver las siguientes cuestiones:

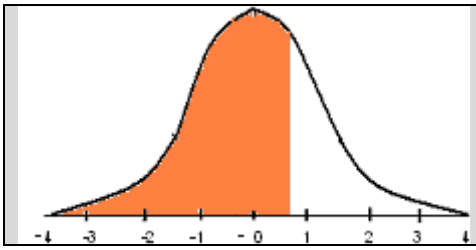
a) ¿Cuál es el porcentaje de recién nacidos con peso superior a 3.600 gr?:



$Z = (X - \bar{X}) / S_X$ → Z = (3600 - 3400) / 400 = 0,5
$Z_{0,5}$ = 0,6915 (probabilidad por debajo)
1 - 0,6915 = 0,3085 (probabilidad por encima)

Porcentaje = 30,85%

b) ¿A qué peso corresponde el percentil 75 (P_{75})?:

	$P_{75} = Z_{0,67}$ 0,67 = (X - 3400) / 400 → X = 3.668 gr

c) Averiguar el intervalo de pesos que incluye el 90% central de los recién nacidos.

Z_1 = 45% (probabilidad por debajo el 5%) Z_1 = (- 1,64)	Z_1 = (- 1,64) = (X - 3400) / 400 X = 2.744 gr
Z_2 = 45% (probabilidad por debajo el 95%) Z_2 = 1,64	Z_2 = 1,64 = (X - 3400) / 400 X = 4.056 gr

d) Averiguar Q (Amplitud Semi-intercuartílica) / 50% de las puntuaciones centrales):

Q = (P_{75} - P_{25}) / 2 → (3668 - 3132) / 2 = 0,268	3400 ± 0,268 = (3.132 y 3.668)

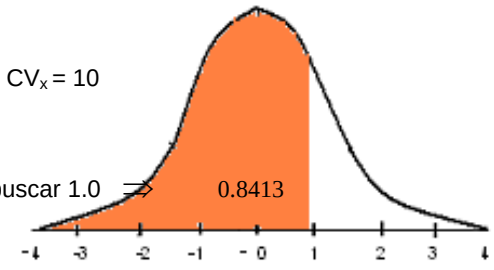
e) Averiguar el porcentaje de sujetos entre las Puntuaciones Directas 3.000 y 3.500.

Z_1 = (3000 - 3400) / 400 Z_1 = - 1 (probabilidad por debajo) → 0,1587	Z_2 = (3500 - 3400) / 400 Z_2 = 0,25 (probabilidad por debajo) → 0,5987
Probabilidad entre = Z_2 - Z_1 → (0,5987 - 0,1587) = 0,44 (44%)	

f) El peso de un grupo de 1000 niños se distribuye normalmente con un coeficiente de variación de 10 (CV_x = 10). Si el 84.13% de ellos no supera los 33 kg. ¿Cuánto vale la media y la desviación típica de la distribución?:

$CV_x = 10$

Es el nº que encontramos en la tabla al buscar 1.0 ⇒ 0.8413



Establecemos el sistema de ecuaciones y resolvemos:

$$\frac{S_x}{\bar{X}} \cdot 100 = 10$$

$$84.13\% \rightarrow z = 1$$

$$1 = \frac{33 - \bar{X}}{S_x}$$

$$\Rightarrow \begin{cases} S_x \cdot 100 = 10\bar{X} \\ S_x = 33 - \bar{X} \end{cases} \Rightarrow (33 - \bar{X}) \cdot 100 = 10\bar{X} \Rightarrow \begin{cases} \bar{X} = 30 \\ S_x = 3 \end{cases}$$

Aproximación de la Binomial a la Curva Normal: Cuando aumenta N (N > 20) y p no toma valores extremos (0,1 < p < 0,9), la Distribución Binomial se aproxima a la Distribución Normal y podemos utilizar la Normal para resolver problemas [las tablas de la Binomial no recogen N mayores de 20, por lo que la utilizaremos a partir de N > 20.

C. Normal (Tipificación) → $Z = (X - \overline{X}) / Sx$ // **C. Binomial** (Tipificación) → $Z = (X - N \cdot p) / \sqrt{N \cdot p \cdot q}$
Para tipificar en la Binomial se utiliza la **corrección por continuidad (X ± 0,5)** que permite utilizar las puntuaciones discretas como si fueran continuas.

Problema ejemplo → De los aspirantes presentados a una prueba de selección, el 0,5% fue elegido para un puesto de trabajo. De un grupo de 25.000 aspirantes, determinar la probabilidad de que el número de seleccionados esté comprendido entre 120 y 132.

Datos → N = 25.000; p = 0,005; Media = N · p → (25.000 · 0,005) = 125

Desviación típica = $\sqrt{N \cdot p \cdot q} = \sqrt{25.000 \cdot 0,005 \cdot 0,995} = 11,15$

Tipificamos los valores (120 y 132) → Utilizando corrección por continuidad

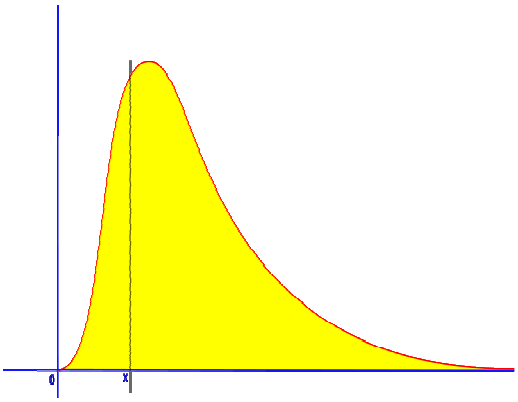
Z_{120} = (119,5 – 125) / 11,15 = (- 0,49) → Probabilidad Tabla III = 0,3121

Z_{132} = (132,5 – 125) / 11,15 = (0,67) → Probabilidad Tabla IV = 0,7486

P (120 < X < 132) = 0,7486 – 0,3121 **= 0,4365**

2.- DISTRIBUCIÓN CHI – CUADRADO DE PEARSON → $\chi^2 = Z_1^2 + Z_2^2 + Z_3^2 + ... + Z_n^2$

Definición → Suma de los cuadrados de n variables normales tipificadas N (0, 1) e independientes. Depende del número de grados de libertad / Nº observaciones que pueden elegirse libremente → (n – 1)



Características de la Distribución Chi – Cuadrado:
→ Nunca puede tomar valores negativos (lo impiden los cuadrados). **Toma valores positivos** en el intervalo desde cero (0) hasta infinito (∞)
→ Si una variable X se distribuye según χ^2 con n grados de libertad, sus parámetros son:
$\mu = E(X) = n$ (nº de grados libertad)
$\sigma^2 = V(X) = 2 \cdot n$ (dos veces el nº de grados libertad).
→ Es una distribución **asimétrica positiva** y cuando aumentan los grados de libertad, la distribución Chi – Cuadrado **se aproxima a la Curva Normal**

→ Para **n > 30** la podemos aproximar a una distribución normal **N (n, 2n)**

Problema Ejemplo: Sea X una variable aleatoria distribuida según **Chi – Cuadrado con 20 grados de libertad** (Nota: la **Tabla V** siempre indica la proporción de casos que un valor deja por debajo):

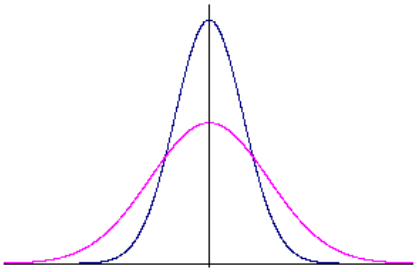
a) Probabilidad de un valor menor o igual que 9,59 → **$P(\chi^2_{20} \leq 9,59) = 0,025$** → Por tanto, mayor o igual que 9,59 → **$P(\chi^2_{20} \geq 9,59) = 1 - 0,025 = 0,975$**

b) Probabilidad entre 37,57 y 28,41 →

$P(28,41 \leq \chi^2_{20} \leq 37,57) =$ $P(\chi^2_{20} \leq 37,57) - P(\chi^2_{20} \leq 28,41) = 0,990 - 0,900 =$ **0,09**

3.- DISTRIBUCIÓN t DE STUDENT → $\{t = Z / \sqrt{(\chi_n^2 / n)}\}$

Definición → Cociente entre una variable normal estándar (Z) y la raíz cuadrada de una variable Chi – Cuadrado (χ^2) dividida por sus grados de libertad. Tiene el mismo número de grados de libertad que la χ^2 de la que surge → (n)



Características de la Distribución t de Student: Es una distribución **acampanada y simétrica** y a medida que aumentan los grados de libertad, la distribución t de Student converge con la Curva Normal Tipificada (la aproximación es buena a partir de 100 grados de libertad (n ≥ 100).
→ La curva es asintótica al eje de abscisas.
→ Su rango puede tomar **valores positivos o negativos** en el intervalo (- ∞**,** ∞)

→ Sus parámetros son: Media **μ = E (t) = 0** y Varianza **σ^2 = V (t) = n / (n – 2)**

Problema Ejemplo: Sea X una variable aleatoria distribuida según **t con 24 grados de libertad** (Nota: la **Tabla VI** siempre indica la probabilidad que deja "por debajo de" el valor dado. Como su media es cero la tabla sólo informa de los percentiles mayores que 50, el resto, menores que 50, son simétricos en negativo):

a) Probabilidad de obtener un valor menor o igual que 2,80 → **P (t ≤ 2,80) = 0,995;** Por tanto, mayor o igual que 2,80 → **P (t ≥ 2,80) = 1 - 0,995 = 0,005**

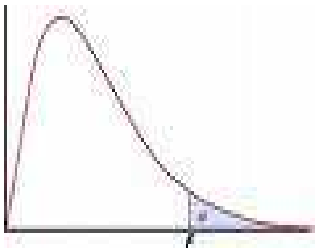
c) Probabilidad entre (– 2,06) y (+ 2,80) →

P (- 2,06 ≤ t ≤ 2,80) = P (t ≤ 2,80) - P (t ≤ - 2,06) = 0,995 – 0,025 = **0,97**

Nota: P (t ≤ - 2,06) = P (t ≥ 2,06) = 1 – P (t ≤ 2,06) = 1 - 0,975 = 0,025

4.- <u>DISTRIBUCIÓN F DE SNEDECOR ($F_{n, m}$)</u> → $\dfrac{\chi_n^2 / n}{\chi_m^2 / m}$

Definición → Cociente entre dos variables Chi – Cuadrado (χ^2) dividida cada una de ellas por sus respectivos grados de libertad.



Características Distribución F: Son distribuciones acampanadas y asimétricas positivas (parecida a la χ^2).
→ Su rango siempre toma **valores positivos** en el intervalo (0, ∞). Sus grados de libertad se corresponden con los de la χ^2 del numerador y la χ^2 del denominador ($F_{n,m}$)
→ Sus parámetros son:
μ = m / m – 2; (para m > 2)
$\sigma^2 = 2m^2 (n + m – 2) / n (m – 2)^2 \cdot (m – 4)$; (para m > 4)
→ **Propiedad recíproca:**
1-p $F_{n \cdot m} = 1 / p F_{m \cdot n}$ (p = probabilidad asociada al valor de la variable)

<u>Problemas ejemplo:</u> Sean X e Y variables aleatorias distribuidas según **$F_{8 \text{ y } 12 \text{ gl}}$.**

Calcular: (X / 8) / (Y / 12) ≤ 2,85 → P (**$F_{8 \text{ y } 12}$** ≤ 2,85)

a) Probabilidad de obtener un valor menor o igual que 2,85 → **P ($F_{8 \text{ y } 12}$ ≤ 2,85) = 0,95**

El valor de F (2,85) deja "por debajo" una probabilidad de 0,95.

b) Probabilidad de obtener un valor mayor o igual que 2,85 → **P (**$F_{8 \text{ y } 12}$ **≥ 2,85) = 1 - 0,95 = 0,05**

El valor de F (2,85) deja "por encima" una probabilidad de 0,05.

c) La media y la varianza de $F_{8\ y\ 12}$

$\mu = m / m - 2$; $\mu = 12 / (12 - 2) =$ **1,2**
$\sigma^2 = 2m^2 (n + m - 2) / n (m - 2)^2 \cdot (m - 4) \rightarrow$ **σ^2** $= 2 \cdot 12^2 (8 + 12 - 2) / 8 (12 - 2)^2 \cdot (12 - 4) =$ **0,81**

d) Valores que limitan el 90% central de esta distribución ($F_{8\ y\ 12}$) →

5%	90%	5%

$F_1 =$ 0,305 $F_2 =$ 2,85

Buscando F_2 en las **Tablas VII** que dejan por debajo 0,95. **El valor de F_1** puede obtenerse aplicando **la propiedad recíproca:**

$$_{0,05}F_{8\ y\ 12} = 1 / {}_{0,95}F_{12\ y\ 8} = 1 / 3,28 = 0,305$$

Nota: Todas las Tablas de la distribución VII indican valores F mayores que 1.

TEMA Nº 8 → ESTIMACIÓN

OBJETIVOS DE APRENDIZAJE:

- Conocer las relaciones entre muestra, análisis estadístico descriptivo y análisis estadístico inferencial.
- Conocer los conceptos de muestra aleatoria y muestra representativa así como las características fundamentales de algunos tipos de muestreo.
- Conocer los aspectos básicos de la inferencia estadística (distribución muestral,..) y la estimación por intervalos.
- Saber realizar inferencias mediante intervalos de confianza para responder a problemas de investigación.

1.- INFERENCIA ESTADÍSTICA → Estudio de las muestras para conocer la población a la que representan. Se trata de estimar, con cierta probabilidad, un parámetro desconocido a partir de una muestra aleatoria extraída de la población. La inferencia siempre se hace en términos probabilísticos (afirmamos con una cierta probabilidad de éxito). **El error muestral** es la diferencia entre el resultado obtenido en la muestra y el que habríamos obtenido si se hubiese trabajado con la población.

FASES PARA INFERIR UN PARÁMETRO A PARTIR DE UNA MUESTRA	
1	**Especificación de la población:** delimitarla claramente (únicamente se pueden hacer inferencias a la población de la que procede la muestra.
2	**Obtención de la muestra:** determinar el número de elementos que la componen.
3	**Medición de la variable:** se obtiene una medida de cada elemento en las mismas condiciones.
4	**Análisis descriptivo y verificación de datos:** los datos se analizan de manera analítica y gráfica y se verifican para detectar errores en la recogida.
5	**Inferencia estadística sobre un parámetro:** se utilizan dos procedimientos, la estimación por intervalo (intervalos de confianza) y el contraste de hipótesis.

2.- CONCEPTOS PREVIOS →

Población (Universo): Conjunto de elementos finito o infinito, definido por una o más características de las que gozan todos los elementos que lo componen y sólo ellos (ejemplo: estudiantes de Psicología en Albacete). Se denomina **N** al **número total** de elementos de la población. Los términos **individuo, sujeto o caso** son los elementos de la población. El **Censo alude al** listado de todos los elementos que componen la población.

Muestra: Subconjunto de los elementos representativos de la población. La muestra representativa recoge el grado de diversidad de la población y la muestra aleatoria la extracción al azar de los elementos de la población (ambos aspectos son deseables para una muestra)

Muestreo: Proceso para extraer muestras representativas de la población. Lo que realmente interesa al investigador es trabajar con muestras reducidas pero que sean representativas. Para obtener una muestra debemos tener en cuenta la variabilidad de la población; cuanto mayor sea la variabilidad, mayor será el número de elementos que deberemos seleccionar de la población para configurar la muestra. Los principales tipos de muestreo son:

a) **Probabilístico (aleatorio):** Cada elemento tiene igual probabilidad de ser elegido. Es el único capaz de darnos la probabilidad de error que cometemos en la inferencia; es decir, la representatividad de la muestra. El **muestreo aleatorio simple** es el más utilizado y la base de todos los demás.
b) **No probabilístico:** La representatividad está sujeta al criterio del investigador. Puede ser **por cuotas o accidental** (estratos o individuos más representativos para los fines de la investigación); **intencional u opinático** (responde al interés del investigador por incluir grupos típicos); **casual o incidental** (de fácil acceso) y **bola de nieve** (un elemento lleva a otro, a otro, etc.)

Muestreo aleatorio simple: Los métodos de muestreo probabilístico son aquellos que se basan en el principio de equiprobabilidad. Es decir, todos los individuos tienen la misma probabilidad de ser elegidos para formar parte de una muestra y, consiguientemente, todas las posibles muestras de tamaño "n" tienen la misma probabilidad de ser elegidas. Sólo estos métodos de muestreo probabilístico nos aseguran la representatividad.

El Muestreo aleatorio simple consiste en tomar de una población de tamaño N, una muestra de tamaño n de forma aleatoria. Se puede hacer con reposición **(muestreo aleatorio simple:** el tamaño de la población es el mismo para cada extracción**)** o sin reposición **(muestreo irrestrictamente aleatorio:** el tamaño de la población cambia en cada extracción). Todas las muestras son equiprobables.
La forma de obtener la muestra es: Determinar el tamaño de la muestra, n. Enumerar de 1 a N todos los elementos de la población (por lo tanto, se requiere de un listado). Extraer "n" números comprendidos entre el 1 y el N, con la ayuda de las tablas de números aleatorios, de un bombo, etc.

Muestreo aleatorio sistemático: Se utiliza cuando los elementos de la población están ordenados o pueden ordenarse (alumnos de un determinado centro). Parte de un arranque aleatorio y se toman los elementos de k en k, siendo **k = N / n**. También exige un listado de los elementos de la población.

Muestreo aleatorio estratificado: Si se sabe que una población puede dividirse en partes o estratos, de forma que, en cada uno de ellos, los elementos poseen una gran homogeneidad con respecto al carácter que se estudia, entonces se aumenta la precisión de las estimaciones tomando una muestra en cada estrato, es decir, actuando separadamente en cada estrato (alumnos de infantil, primaria y secundaria de un centro). También se necesita un listado de los elementos de la población.

Muestreo por conglomerados: En este método la unidad muestral es un grupo de elementos de la población, a la que llamamos conglomerado. Normalmente estos grupos tienen una existencia real (manzanas de viviendas, distritos universitarios, etc.). Con gran frecuencia los conglomerados son áreas geográficas; por ello el muestreo por áreas no es más que un caso particular del muestreo por conglomerados. La ventaja de este método es que no se necesita un listado de todos los elementos de la población, sólo hay que conocer los elementos de los conglomerados seleccionados.

Muestreo polietápico: Es una combinación del muestreo estratificado y por conglomerados: En la primera etapa se selecciona una serie de conglomerados o unidades muestrales primarias. En una segunda etapa se selecciona conglomerados más pequeños, o unidades muestrales secundarias. Y así sucesivamente, tantas etapas como sea necesario. Tiene la ventaja de que en cada etapa se puede aplicar el Muestreo Aleatorio (simple, estratificado, sistemático) que se considere más adecuado al tipo de conglomerados de que se trate.

3.- <u>ESTIMACIÓN DE PARÁMETROS</u> ➔ Generalmente se desconocen los parámetros de la población por lo que se hace necesario estimarlos a partir de los valores muestrales. Un estimador es un estadístico que utilizamos para estimar parámetros.

Distribución Muestral de un Estadístico: Distribución de todos los estadísticos obtenidos de todas y cada una de las posibles muestras que se pueden extraer de una población. La Distribución Muestral es una Distribución de Probabilidad de un estadístico calculado en todas las posibles muestras del mismo tipo y tamaño constante.

<u>Distribución muestral de la media:</u> Distribución de todas las medias obtenidas de todas las posibles muestras que se pueden extraer de una población (media de todas las medias posibles).

Propiedades:

- La media de la distribución muestral de medias **(μ_x)** coincide con la media de la población **(μ)**
- La varianza de la distribución muestral de la media es igual a → **(σ^2 / n)** y la desviación típica de la distribución muestral de medias **($\sigma_x = \sigma / \sqrt{n}$)** se denomina **Error típico de la media**.
- La forma de la distribución muestral de la media es normal cuando lo es la distribución de la variable estudiada (al margen de n, tamaño de la muestra) y tiende a la normal a medida que n va aumentando (al margen de la distribución de la variable) → **Teorema central del límite** (la aproximación es buena para $n \geq 30$)

	Población	Muestra	D. Muestral (Media)
Media	$\mu = (\sum X / N)$	$\bar{X} = (\sum X / n)$	$\mu_x = \mu$
Varianza	$\sigma^2 = \sum (X - \mu)^2 / N$	$S^2_{n-1} = \sum (X - \bar{X})^2 / n-1$ Cuasivarianza	$\sigma^2_x = \sigma^2 / n$
Desviación Típica	$\sigma = \sqrt{\sum (X - \mu)^2 / N}$	$S^2_{n-1} = \sqrt{\sum (X - X)^2 / n-1}$ Cuasidesviación típica	$\sigma_x = \sqrt{\sigma^2 / n} = \sigma / \sqrt{n}$ Error Típico de la Media

La media de la muestra es un estimador puntual insesgado (sin error) para estimar la media de la población.
El error típico de la media (desviación típica de la distribución muestral de la media) es un indicador de la precisión al estimar la media. Depende de la desviación típica de la población y del tamaño de la muestra.

A menor desviación típica **(σ)** ó mayor tamaño de la muestra **(n)** → menor error típico **(σ_x)**

Distribución muestral de la proporción: Distribución de todas las proporciones obtenidas de todas las posibles muestras que se pueden extraer de una población.

Propiedades:

- Cuando la variable sólo toma valores 0 y 1 (dicotómica), la proporción de la muestra se define como **$P = (\sum X / N)$.**
- El estadístico P, se distribuye según la Binomial con: **$\mu_p = \pi$** y **$\sigma^2_p = \pi (1- \pi) / n$**
- Según el Teorema central del límite, a medida que el tamaño de la muestra crece, la distribución muestral de la proporción tiende a la normal Con media **π** y varianza **$\pi (1- \pi) / n$** (cuanto más se aleje π de 0´5, más elementos debe tener la muestra para hacer aproximación a la normal)

	Población	Muestra	D. Muestral (Proporción)
Media	$\pi = (\sum X / N)$ Donde X = 0 y 1	$P = (\sum X / n)$ Donde X = 0 y 1	$\mu_p = \pi$
Varianza	$\sigma^2 = \pi (1- \pi)$	$S^2 = P (1 – P)$	$\sigma^2_p = \pi (1- \pi) / n$
Desviación Típica	$\sigma = \sqrt{\pi (1- \pi)}$	$S = \sqrt{P (1 – P)}$	$\sigma_p = \sqrt{\pi (1- \pi) / n}$ Error Típico de la Proporción

La media de la distribución muestral de la proporción **(μ_p)** es un estimador puntual insesgado (sin error) para estimar la proporción poblacional **(π)**
El error típico de la proporción (desviación típica de la distribución muestral de la proporción) es un indicador de la precisión al estimar la proporción. Depende de la desviación típica de la población y del tamaño de la muestra.

A menor desviación típica **(σ)** ó mayor tamaño de la muestra **(n)** → menor error típico **(σ_p)**

Estimación por intervalos: En lugar de un único valor como estimación del parámetro, se elabora un intervalo en el que se espera que se encuentre el parámetro desconocido, con una cierta probabilidad. La estimación por intervalos **establece un rango de valores** dentro del cual estaría el valor del parámetro. La estimación por intervalos requiere una probabilidad (nivel de confianza). La amplitud del intervalo nos indicará su precisión. **A menor amplitud, más precisión**, más informativo es, más útil.

Una estimación por intervalos depende de: una estimación puntual del parámetro (media o proporción); una medida de variabilidad (error típico); una probabilidad (nivel de significación o nivel de confianza) y un supuesto acerca de la distribución en la población (o el tamaño de la muestra)

Cálculo de los intervalos de confianza para la media: para simplificar la exposición, partiremos del valor de la **media en la muestra**, a la que sumaremos y restaremos el **Error de estimación máximo** (depende de la desviación típica de la población, del nivel de confianza y del tamaño de la muestra) para averiguar los límites del intervalo de confianza.

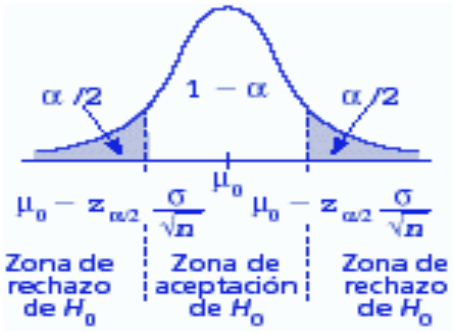
$\bar{X} \pm E_{máx}$ = Límites del intervalo de confianza // $E_{máx} = Z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}$ // $E_{máx}$ = (Z · Error típico)

Los límites son simétricos respecto a la Media Muestral, por tanto → $\bar{X}$ = (Lím Inf + Lim Sup) / 2

A partir de este planteamiento general, se pueden presentar tres casos:

INTERVALOS DE CONFIANZA PARA LA MEDIA	
Muestreo aleatorio simple Varianza poblacional conocida **(σ^2)** Distribución Normal ó n ≥ 30	**$\bar{X} \pm E_{máx}$ = Límite inferior y superior** **$E_{máx} = Z_{1-\alpha/2}$ ($\sigma / \sqrt{n}$)**
Muestreo aleatorio simple Varianza poblacional desconocida Distribución Normal y n < 30	**$\bar{X} \pm E_{máx}$ = Límite inferior y superior** **$E_{máx} = T_{n-1\ y\ 1-\alpha/2}$ ($S_{n-1} / \sqrt{n}$)**
Muestreo aleatorio simple Varianza poblacional desconocida Distribución Normal ó n ≥ 30	**$\bar{X} \pm E_{máx}$ = Límite inferior y superior** **$E_{máx} = Z_{1-\alpha/2}$ ($S_{n-1} / \sqrt{n}$)**

Valores α / 1- α / Z más frecuentes



α	0´05	0´01	0´001
1 - α	0´95	0´99	0´999
$Z_{\alpha/2}$	- 1´96	- 2´58	- 3´29
$Z_{1-\alpha/2}$	1´96	2´58	3´29
α → Nivel de significación			
1 – α → Nivel de confianza			
$Z_{\alpha/2}$ → Z (Tabla III negativa)			
$Z_{1-\alpha/2}$ → Z (Tabla IV positiva)			

Problemas Ejemplo:

Problema 1.- Se midieron los niveles de depresión en una muestra de 36 personas. Asumiendo que la variable se distribuye normalmente en la población, se calculó la media de las puntuaciones y se obtuvo un valor de 8 y una desviación típica igual a 2. Hallar los límites del intervalo de confianza para la media de la población, con un nivel de confianza del 0´99.

Datos→ $\bar{X}$ = 8 // Distr. Normal (n ≥ 30) // 1-α = 0,99 // S_{n-1} = 2 (Varianza poblacional desconocida)

Intervalo de confianza → $\bar{X} \pm E_{máx}[Z_{1-\alpha/2} (S_{n-1} / \sqrt{n})]$ = [Límite superior y Límite inferior]

Nivel de significación (α = 0,01) → (α/2) = (0,01/2) = **0,005** → (1 - α/2) = **0,995**

Las probabilidades (0,005 y 0,995) corresponden a los valores → **Z = ± 2,58**

Error Típico → $\sigma_x = (S_{n-1} / \sqrt{n})$ → $(2 / \sqrt{36})$ = **0,33**

Error de estimación máximo → $(Z\alpha \cdot \square\ \sigma_x) = (2{,}58 \cdot \square 0{,}33) =$ **0,85**

Límite Superior = 8 + (2,58 · □0,33) **= 8,85**
Límite Inferior = 8 - (2,58 · □0,33) **= 7,15**

Límites del intervalo de confianza → [7,15 y 8,85]

Conclusión: La media de los niveles de depresión aplicada a la población general es, con una probabilidad del 99%, un valor comprendido entre 7,15 y 8,85 → **P (7,15 ≤ μ ≤ 8,85)**

Problema 2.- Hallar los límites del intervalo de confianza suponiendo que la muestra está compuesta por 25 personas, con un nivel de confianza del 0´95.

Datos→ $\overline{X} = 8$ // Distr. Normal (n < 30) // 1-α = 0,95 // $S_{n-1} = 2$ (Varianza poblacional desconocida)

Intervalo de confianza → $\overline{X} \pm E_{máx}[T_{n-1\ y\ 1-\alpha/2}\ (S_{n-1} / \sqrt{n})]$ = [Límite superior y Límite inferior]

Nivel de significación (**α = 0,05**) → (α/2) = (0,05/2) = **0,025** → (1 - α/2) = **0,975**

(0,025 y 0,975) corresponden a los valores → **T de Student con (25-1 gl) = ± 2´064**

Error Típico → $\sigma_x = (S_{n-1} / \sqrt{n})$ → $(2 / \sqrt{25}) =$ **0,4**

Error de estimación máximo → $(T\alpha \cdot \square\ \sigma_x) = (2{,}064 \cdot \square 0{,}4) =$ **0,8256**

Límite Superior = 8 + (2,064 · □0,4) **= 8,8256**
Límite Inferior = 8 - (2,064 · □0,4) **= 7,1744**

Límites del intervalo de confianza → [7,1744 y 8,8256]

Tamaño de la muestra para estimar la media: Cuanto mayor es el tamaño de la muestra mayor es la precisión del intervalo y mayor la precisión de la estimación. Cuanto menor es el error típico, menor es el intervalo de confianza y, por tanto, más preciso (para reducirlo se aumenta el tamaño muestral). La fórmula para averiguarlo (varianza poblacional conocida) se obtiene despejando n de la ecuación sobre el Error de estimación máximo:

$$E_{máx} = Z_{1-\alpha/2}\ (\sigma / \sqrt{n}) \rightarrow \mathbf{n = (\sigma^2 \cdot Z^2_{1-\alpha/2}) / E^2_{máx}}$$

Nota: Cuando se desconoce la varianza poblacional se sustituye en la formula **($Z_{1-\alpha/2}$)** por (**$T_{n-1\ y\ 1-\alpha/2}$**) y **(σ^2)** por **(S^2_{n-1})** → El nivel de confianza se distribuye según T de Student y la varianza poblacional se estima a través de la varianza de la muestra (cuasivarianza)

Problema Ejemplo: Por experiencias anteriores se sabe que las estaturas de los soldados tienen una varianza poblacional de 64 cm. ¿Qué tamaño debe tener la muestra para que la media estimada no se aleje más de ± 1´5 puntos de la media poblacional?, considere (α = 0,05)

Para (1- α/2 = 0,975) → $Z_{Tablas} = 1{,}96$ Varianza poblacional conocida **(σ^2)** = 64

$n = (\sigma^2 \cdot Z^2_{1-\alpha/2}) / E^2_{máx}$ → $n = (64 \cdot 1{,}96^2) / 1{,}5^2 = 109{,}27$ **≈ 109 soldados**

Con un margen de error del 5% debemos tomar una muestra de 109 soldados.

Cálculo de los intervalos de confianza para la proporción: de forma similar a la utilizada en la media, partiremos del valor de la proporción muestral (variable dicotómica o dicotomizada), a la que sumaremos y restaremos el **Error de estimación máximo** (depende del nivel de confianza **1-α**; del error típico de la proporción → $\sqrt{\pi\ (1- \pi) / n}$ y del tamaño de la muestra **n**. Se debe cumplir para aproximación a la normal → **n π (1- π) ≥ 5**

P ± $E_{máx}$ = Límites del intervalo de confianza // $\mathbf{E_{máx} = Z_{1-\alpha/2} \cdot \sqrt{P\ (1 - P) / n}}$

Los límites son simétricos respecto a la Proporción, por tanto → **P = (Lím Inf + Lim Sup) / 2**

Problema Ejemplo: Para comprobar la eficacia en la aplicación de un tratamiento, se someten al mismo 64 pacientes. Finalizado el periodo de aplicación, se observó que remitió la enfermedad en 50 casos. Con un nivel de significación (α = 0,05), estime por intervalo el porcentaje de efectividad del tratamiento objeto de estudio.

Datos: Proporción muestral → (50 / 64) **= 0,781** // Nivel de confianza → 1 – α = 0´95

Condición (aproximación a la normal → 64 · 0´781 · (1 – 0´781) = 10´946 ≥ 5

Para (α/2 = 0,025) → Z = **(- 1´96)** y Para (1 - α/2 = 0,975) → Z = **(+ 1´96)**

Intervalo de confianza → $P \pm |Z_{1-\alpha/2}| \cdot \sqrt{P\,(1-P)\,/\,n}$ = Límites Superior e Inferior

$\mathbf{E_{máx}} = Z_{1-\alpha/2} \cdot \sqrt{P\,(1-P)\,/\,n}$ → $\mathbf{E_{máx}} = 1´96 \cdot \sqrt{0´781 \cdot 0´219\,/\,64} = 1´96 \cdot 0´0517 =$ **0´1**

$0,781 \pm (1,96 \cdot 0,0517) =$ **[0´681 y 0,881]**

Con un margen de error del 5% el tratamiento será efectivo entre el 68% y el 88% de los casos.