TEMA 1 - CONCEPTOS BÁSICOS Y ORGANIZACIÓN DE DATOS

Estadística teórica (aspectos formales y normativos) y **aplicada** (aplicación a un campo concreto)

Estadística aplicada o análisis de datos: Niveles de medida

1.-*nominal*
2.-*ordinal*
3.-*de intervalo*
4.-*de razón*

Método científico: dar razón sistemática, empírica y experimental, de los fenómenos

Es **sistemático**→porque tiene etapas definidas
Es **replicable**→porque los datos obtenidos pueden ser replicados o refutados

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

Estadística: se ocupa de sistematización, recogida, ordenación y presentación de los datos referentes a un fenómeno que presenta variabilidad o incertidumbre para su estudio metódico, con objeto de hacer previsiones sobre los mismos, tomar decisiones u obtener conclusiones.

Estadística descriptiva→ Se organizan y resumen conjuntos de observaciones procedentes de una muestra. Cuantitativa (tablas, gráficos, valores numéricos)

Con 1 variable:
Índices para valores más habituales (índices de **tendencia central**)
Hasta que punto son similares o diferentes entre si (estadísticos de **variabilidad**)
Gado en que las observaciones se representan por encima o debajo de la tendencia central (estadísticos de **asimetría**)

Con 2 variables:
Relacionar variables entre sí (**coeficientes de correlación**)
Predecir el valor de una variable en función de otra (**ecuaciones de regresión**)

Estadística Inferencial→Inferencias a cerca de una población basándose en datos obtenidos de una muestra. Se utiliza el cálculo de probabilidades.

En una investigación se pretende conocer un **parámetro** (una característica) de una **población**, y como es demasiado amplia, se realiza un muestreo con el que se obtiene una **muestra de elementos que la representan**. Se estudia la característica deseada en la muestra mediante **estadísticos** que estiman los parámetros de la población.
Queremos conocer un parámetro "porcentaje de individuos que responden si" (y como no es posible por lo extensa de la población) conocemos la estimación de ese parámetro→el estadístico o porcentaje de la muestra que responden si.

POBLACION→conjunto de todos los elementos que cumplen una determinada característica objeto de estudio.
MUESTRA→subconjunto cualquiera de una población. Solo sirve para el total de la población si es representativa.
PARÁMETRO→propiedad descriptiva (medida) de una población
ESTADÍSTICO→propiedad descriptiva (medida) de una muestra

Para que una muestra sea representativa se deben utilizar métodos de muestreo probabilística (una **muestra probabilística** se elige mediante reglas matemáticas y una muestra no probabilística no, ej. Muestras de conveniencia o incidental (conformada por personas de fácil acceso para el investigador, o la bola de nieve (un elemento lleva a otro y así sucesivamente)

MEDICIÓN Y ESCALAS DE MEDIDAS

Medición: Proceso por el cual se asignan números a objetos o características según determinadas reglas
Objetos físicos→medición directa
Variables no directamente observables→??

Característica: cualquier propiedad de un objeto
Modalidad: formas en las que se presenta la característica (se asigna un nº a cada una de las modalidades de una característica)

Se utilizan diferentes escalas (*conjunto de reglas o modelos desarrollados para la asignación de nº a los valores de las variables*) en función de la variable a medir (timidez en clase, tiempo en realizar una tarea, nacionalidades de un conjunto, etc.)

Según Stevens (1946) Cuatro tipos de **escalas de medidas**:
1) **Nominal** (igualdad o desigualdad, entre 2 modalidades)
2) **Ordinal** (además, se puede establecer un orden)
3) **De intervalo** (además, se usa una unidad y tienen sentido las diferencias)
4) **De razón** (además, se pueden comparar dos medidas mediante un cociente)

Escala nominal

Asignación *arbitraria* de números o símbolos a cada una de las diferentes modalidades de la característica. Relación de **igualdad o desigualdad**, que implica la pertenencia o no a una categoría determinada.
Ej.: Religión (practicantes, no practicantes)

Escala Ordinal

Asignación (*no arbitraria*, sino atendiendo el orden existente entre las categorías) de números a objetos para indicar la extensión relativa en que se posee una característica. Se clasifica a las personas, objetos o eventos en una posición con relación a cierto atributo, pero sin indicar la distancia entre las posiciones. Solo se **indica el orden**. Permite la identificación, diferenciación y el establecimiento de relaciones de tipo “mayor que” o “menor que”.
Ej.: Estatus (alto, medio, bajo)

Escala de intervalo

Ordena los objetos o eventos según la magnitud del atributo que presentan y proveen intervalos entre las unidades de medida. Origen **arbitrario** y no refleja la ausencia de la magnitud que estamos midiendo. Se puede saber si un objeto es igual o diferente, si posee en mayor o en menor grado la característica de interés y estos números se pueden restar y sumar y las diferencias entre esos números se pueden multiplicar y dividir.
Su característica es la existencia de una **unidad de medición** común y constante, que permite asignar un nº real a todos los pares de objetos del conjunto ordenado.
Ej. Inteligencia (0,90, 160, etc.)

Escala de razón

Los números asignados admiten como válidas las relaciones de igualdad-desigualdad, orden, suma, resta, multiplicación y división. Tiene todas las características de la medida de intervalo y se suma que se le puede asignar un punto de origen verdadero, un **valor absoluto** (valor cero= ausencia de la magnitud).
Ej.: Altura

NOMINAL	Los números identifican y clasifican objetos	Igual-desigual	Sexo, estado civil, raza,
ORDINAL	+, los números indican las posiciones relativas de los objetos	mayor que- igual que	Grado de satisfacción, dureza
INTERVALO	+, hay una unidad de medición común	+, igualdad-desigualdad de diferencias	Temperatura, inteligencia
RAZÓN	+, el punto cero es absoluto.	+, igualdad-desigualdad de razones	Longitud, peso, altura

VARIABLE: CLASIFICACIÓN Y NOTACIÓN

Característica con 1 sola modalidad→**constante**

Variable: Representación numérica de una característica que presenta más de una modalidad (valor) de un conjunto determinado.

Tres tipos**: 1) Cualitativa** (nominales)

En función del número de categorías o modalidades:
Variable dicotómica: 2 categorías (Ej.: el sexo)
Variable politómica: Más de 2 categorías (Ej.: nacionalidades)

2) Cuasicuantitativa (ordinales)

3) Cuantitativa (de intervalo y de razón)

En función de los valores numéricos que pueden asignarse:
Variable continua: valores en cualquier punto de la escala (Ej.: peso)
Variable discreta: valores aislados, sin valores intermedios (Ej.: nº de hijos)

Variable independiente→suceso causa de otro
Variable dependiente→efectos de la variable independiente
Variable extraña→las que influyen sobre la variable independiente, pero que no se estudian.

Notación de la variable

Letras latinas mayúsculas, con un subíndice i

X_i, siendo i=1, 2, 3, …, n (siendo n, el número de elementos que componen la muestra)

DISTRIBUCIÓN DE FRECUENCIAS

Los datos con los que se trabaja pueden provenir de la medición directa de las variables o de frecuencias que provienen de un proceso de conteo. Normalmente se organiza la información mediante una **distribución de frecuencias** (*representación de la relación entre un conjunto de medidas exhaustivas y mutuamente excluyentes y la frecuencia de cada una de ellas*)

Organiza los datos
Da información para la representación gráfica
Facilita los cálculos para estadísticos muestrales

Frecuencia absoluta→(n_i) número de observaciones en cada categoría

Frecuencia relativa o proporción de cada categoría →(p_i) se obtiene dividiendo la (n_i) , entre el número total de observaciones.

En **porcentaje (P_i)** multiplicando cada proporción por 100.

Variable cualitativa (nominal)

X	n_i	p_i	P_i
Hombres	24	0,6	60
Mujeres	16	0,4	40
n=	40	1	100

Variable cuasicuantitativa (ordinales)

Igual pero respetando el orden predeterminado. Y se añaden la **frecuencia absoluta acumulada (n_a), frecuencia relativa acumulada o proporción acumulada (p_a)** y el **porcentaje acumulado (P_a),** para cada una de las categorías o modalidades de respuesta, y se obtienen acumulando (sumando) desde la categoría de menor valor de la variable a la de mayor valor, las frecuencias absolutas, proporciones o porcentajes, de cada categoría de respuesta**.**

X	n_i	p_i	P_i	n_a	p_a	P_a
Primaria	13	0,33	33	13	0,33	33
ESO	11	0,28	28	24	0,60	60
FP	7	0,18	18	31	0,78	78
Diplomatura	4	0,10	10	35	0,88	88
Licenciatura	5	0,13	13	40	1,00	100
n=	40	1,00	100			

Frecuencia absoluta→(n_i) Nº de veces que se repite cada uno de los valores de una variable. La suma de todas las frecuencias absolutas representa el **total de la muestra (n)**

Frecuencia relativa o proporción de cada categoría →(p_i) Cociente entre la frecuencia absoluta de cada variable **(n_i)** y Nº total de observaciones **(n)→** **(p_i)= (n_i) / (n)**

Porcentaje→ (P_i) Valor de la frecuencia relativa multiplicado por 100. **(P_i)= (p_i).100**

Frecuencia absoluta acumulada→(n_a) Nº 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 total de observaciones **(n)→** **(p_a)= (n_a) / (n)**

Porcentaje acumulado→(P_a), Valor de la frecuencia relativa acumulada multiplicado por 100. **(P_a)= (p_a).100**

Variable cuantitativa (de intervalo y de orden)

1) **Nº de valores** de la variable **reducido** (Ej.: nº de hijos)→**Igual** que con variables **ordinales**
2) **Nº de valores amplio** (Ej.: edad, altura)→agrupar en **intervalos** (grupos de valores consecutivos) al establecer intervalos siempre se pierde información y se puede optar por la **amplitud** que más se ajuste al estudio (equilibrio entre la precisión que se necesite y la manejabilidad de los datos.
   Limites de los intervalos: hay que tratar de que el **límite superior exacto** de un intervalo coincidan con el **límite inferior exacto** del siguiente. Cuando no es así, se los llama: **límites informados o aparentes** (Ej.: edades entre 26 - 35, debe ser entre 25,5 - 35,5)

   Limites exactos= valor informado+- 0,5 x I (siendo I la unidad del instrumento de medida)
   Punto medio: semisuma *((a+b)/2*) del límite superior e inferior del intervalo de los límites exactos o de los aparentes
   Intervalo abierto: que no tiene límite inferior o superior (76 años o más)

Intervalo→cada uno de los grupos de valores que ocupan una fila en una distribución de frecuencia.

Límites aparentes, virtuales o informados→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 fuera exacto.

Punto medio del intervalo→semisuma de los límites exactos o de los límites aparente.

Amplitud del intervalo→diferencia entre el límite exacto superior y el límite exacto inferior

REPRESENTACIONES GRÁFICAS

Eje vertical→ordenada (o eje de las Y)
Eje horizontal→abscisa (o eje de las X)

1º cuadrante: +x, +y
2º cuadrante: -x, +y
3º cuadrante: -x, -y
4º cuadrante: +x, -y

a) Diagrama de barras (variables nominales, ordinales y cuantitativas discretas)
Abscisa (X)→valores de la variable
Ordenada (Y)→ frecuencias
En las ordinales y cuantitativas discretas, se puede utilizar también un **diagrama de barras acumulativo.**

b) **Diagrama de sectores** (variables cualitativas (nominal) y cuasicuantitativas (ordinal))
Forma de círculo, cuya superficie es proporcional a la frecuencia de la modalidad correspondiente. El ángulo total representa el nº total de observaciones y para determinar el ángulo de los sectores se multiplica la frecuencia relativa (proporción) por 360

c) **Pictograma** (variables cualitativas (nominal))
Dibujos alusivos cuya área es proporcional a la frecuencia de la modalidad que representa.

d) **Histograma** (variables cuantitativas continuas con datos agrupados en intervalos)
Abscisa (X)→intervalos con limites exactos (todos con la misma amplitud) o los puntos medios y sobre ellos se levantan rectángulos cuyas áreas sean proporcionales a la frecuencia correspondiente.
Ordenada (Y)→ frecuencias

e) **Polígono de frecuencias** (variables discretas y continuas)
Se unen los extremos superiores de lo que serían las barras (si se hubiera hecho un diagrama de barras) o de un histograma en los puntos medios de las bases superiores (variable continua)

REPRESENTACIONES GRÁFICAS DE DOS VARIABLES

a) **Diagrama de barras conjunto** (al menos una de las dos variables es cualitativa (nominal))
Cuando las dos son cualitativas conviene organizar los datos en una **tabla de doble entrada.**

X	Hombre	Mujer	
Casado	12	12	24
Divorciado	4	2	6
Soltero	4	2	6
Viudo	4	0	4
	24	16	40

Deben representarse en el mismo gráfico ambas situaciones.
Abscisa (X)→estados civiles
Ordenada (Y)→ porcentaje
Es importante que el nº de sujetos sea el mismo para utilizar las frecuencias absolutas, de lo contrario es recomendable utilizar las frecuencias relativas o porcentajes.

b) **Diagramas de dispersión o nube de puntos** (dos variables cuantitativas)
Dando idea de la relación que existe entre ambas variables.
Abscisa (X)→una variable
Ordenada (Y)→ la otra
Para cada par de datos se localiza la intersección y se marca con un punto
Se pueden establecer **relaciones lineales** entre variables.

PROPIEDADES DE LA DISTRIBUCIÓN DE FRECUENCIAS

Tendencia general: Lugar donde se centra una distribución particular en la escala de valores.

Variabilidad: Grado de concentración de las observaciones en torno al promedio.

Homogénea (poca variabilidad) si los valores estás cercanos al promedio.
Heterogénea (mucha variabilidad) si los valores se dispersan mucho con respecto al promedio.

Asimetría o sesgo: Grado en que los datos se reparten equilibradamente por encima y por debajo de la tendencia general.

Distribución simétrica: cuando al dividirla en dos a la altura de la media, las dos mitades se superponen.
Asimetría positiva: cuando la mayor concentración está en la parte baja de la escala (test difíciles)
Asimetría negativa: cuando la mayor concentración está en la parte alta de la escala (test fáciles)

TEMA 2 – MEDIDAS DE TENDENCIA CENTRAL Y POSICIÓN

MEDIDAS DE TENDENCIA CENTRAL

La tendencia central de una distribución de frecuencias se puede resumir en un valor o puntuación, las medidas o índices de puntuación de tendencia central indican sobre que puntuación se concentran las observaciones.

Media aritmética
Mediana
Moda

Media aritmética ($\bar{X}$)

Promedio o medio más conocido y usado. **Valor central alrededor del cual están la mayoría de las observaciones. Solo para variables cuantitativas.**

$$\bar{X} = \frac{\text{suma de todos los valores } (X_1, X_2, X_3 \ldots + X_n)}{\text{n= nº total de observaciones}} = \frac{\mathbf{E\ X_i}}{\mathbf{n}}$$

Cuando el nº de observaciones es elevado:

A partir de las **Frecuencias absolutas (n_i):**

$$\bar{X} = \frac{\mathbf{E\ n_i\ X_i}}{\mathbf{E\ n_i}} = \frac{\mathbf{E\ n_i\ X_i}}{\mathbf{n}}$$

n= nº total de observaciones

X_i= el valor i en la variable X (o punto medio del intervalo)

n_i= frecuencia absoluta del valor o intervalo i.

o de las **Frecuencias relativas (p_i):**

$$\mathbf{\bar{X} = E\ p_i\ X_i}$$

X_i= el valor i en la variable X (o punto medio del intervalo)

p_i= frecuencia relativa o proporción de observaciones del valor o intervalo i.

Propiedades matemáticas:

1) La suma de las desviaciones de cada valor con respecto a su media es igual a cero.

$$\mathbf{\sum_{i=1}^{n} (X_i - \bar{X}) = 0}$$

2) Si a los valores de la variable X le aplicamos la siguiente transformación lineal: $Y_i = bX_i + a$, la media de los nuevos valores **Y** → $\mathbf{\bar{Y} = b\bar{X} + a}$

Límites:

a) Cuando los datos están agrupados en intervalos, la media no se puede calcular si el intervalo máximo no tienen límites superior o el intervalo mínimo no tiene límite inferior
b) Sensible a valores extremos (no se recomienda en distribuciones asimétricas)

Mediana (Md):

Buena para representaciones asimétricas. No es sensible a **valores** extremos porque en su cálculo no entran todos los valores (como en la media aritmética) sino únicamente los que ocupan las **posiciones centrales**. En todo tipo de variables, menos en las cualitativas. **Valor de la variable que divide la distribución de frecuencias en dos partes iguales, conteniendo un 50% de las observaciones**.

Se ordenan las puntuaciones de mayor a menor, si es nº impar, la mediana es la observación que ocupa la posición central; si es nº par la mediana es la media aritmética de los dos valores centrales.

Cuando el nº de observaciones es elevado:
Intervalo en el que se encuentra la mediana→**intervalo crítico** y corresponde con aquel en el que la **frecuencia absoluta es igual o superior a n/2.**

$$\mathbf{Md = L_i + \left(\frac{\frac{n}{2} - n_d}{n_c} \right) . I}$$

L_i = Limite exacto inferior del intervalo crítico
n = nº de observaciones
n_d =Frecuencia absoluta acumulada por debajo del intervalo crítico
n_c = Frecuencia del intervalo crítico
I = Amplitud del intervalo crítico

Se asume que la distribución de las frecuencias dentro de cada intervalo es homogénea. Ej.: sabemos que el nº de observaciones totales es de 50 y por tanto la media dividirá en 25 sujetos a esta observación, si el límite superior del intervalo crítico es de 22, faltan 3 observaciones para llegar al 50% en el que se encuentra la mediana. Asumimos que estas puntuaciones se reparten homogéneamente dentro del intervalo.

Si los datos no están ordenados en intervalos:
Se genera un caso particular en el que I (amplitud del intervalo crítico) es =1

No se puede utilizar cuando el intervalo donde se encuentra la mediana es abierto.

Moda (Mo):
Se puede utilizar en variables cuantitativas y cualitativas.
Cualitativa→la moda es la categoría con la máxima frecuencia.
Cuantitativa sin intervalos→la moda es el valor con mayor frecuencia absoluta (n_i)
Cuantitativa con intervalos→se localiza el intervalo modal que es el intervalo con la frecuencia máxima y la moda es el punto medio de dicho intervalo.
Si un único valor con la frecuencia máxima, una moda→unimodal
Son dos o más valores con la frecuencia máxima→bimodal, trimodal, etc.

Características:
a) Cálculo sencillo y de fácil interpretación.
b) Cuando la variable es cuantitativa con intervalo, la moda no se puede calcular si el intervalo modal está en un intervalo abierto.

Elección de una medida de tendencia central
Se recomienda la **media aritmética** (se desaconseja cuando la distribución de las frecuencias es muy asimétrica) y **no** se puede cuando el nivel de medida es **nominal** u **ordinal** ni en datos agrupados con **intervalos abiertos** en sus extremos.
La siguiente es la **mediana**, resistente a los valores extremos, **si** se puede con niveles **ordinales y** en datos agrupados con **intervalos abiertos. No** en variables **nominales, cuando la mediana se encuentra en el intervalo abierto.**
Moda, no se puede cuando la frecuencia sea **amodal o el intervalo modal** coincida con el **intervalo abierto.**

CUALITATIVA (nominal)→ **MODA**
CUASICUANTITATIVA (ordinal)→**MODA, MEDIANA**
CUANTITATIVA ((de intervalo y de razón)→**MODA, MEDIANA Y MEDIA ARITMÉTICA**

(CUANTITATIVA, SIMETRICA Y UNIMODAL→MEDIA, MEDIANA Y MODA= VALOR)

MEDIDAS DE POSICIÓN

Medidas o índices de posición o cuantiles: Informan acerca de la posición relativa de un sujeto con respecto a su grupo de referencia, dentro de la distribución de frecuencias de la variable (situación de una puntuación con respecto a un grupo, utilizando a éste como referencia).
Dividir la distribución en un nº de partes o secciones iguales entre sí en cuanto al nº de observaciones (la mediana divide en dos partes, 50%) dependiendo de cuantos valores utilicemos para dividir la distribución:

Percentiles
Cuarteles
Deciles

Percentiles (o centiles)→ k→(P_k)
99 valores que dividen en 100 partes iguales la distribución de frecuencias.
Ej.: percentil 50→ (P_{50}): Divide a la distribución de frecuencia en 50%, igual que la mediana**.** **P_{50}= Md**
Cálculo:
Frecuencias absolutas (n_i) en intervalos
Intervalo donde está el percentil $_k$ →intervalo crítico
Intervalo crítico =frecuencia absoluta acumulada (n_a) es igual o superior a $\frac{n.k}{100}$

$$P_k = L_i + \left(\frac{\frac{n.k}{100} - n_d}{n_c} \right) . I$$

n_d: Frecuencia absoluta acumulada por debajo del intervalo crítico
n_c: Frecuencia absoluta del intervalo crítico
L_i: Límite inferior exacto del intervalo crítico
I: Amplitud del intervalo

Datos agrupados sin intervalos:
Misma formula con (I=0)
Este método es para calcular el valor de cualquier de los 99 valores (valor de X, dado k)

Para calcular que posición ocupa un valor de la variable X_i (valor de k, dado X)

$$k = \left(\frac{\frac{(P_k - L_i).n_c + n_d}{I}}{n} \right) . 100$$

n_d: Frecuencia absoluta acumulada por debajo del intervalo crítico
n_c: Frecuencia absoluta del intervalo crítico
L_i: Límite inferior exacto del intervalo crítico
I: Amplitud del intervalo

Si el resultado es con decimales se toma la cantidad entera más próxima.

Cuartiles (Q_1) (Q_2) (Q_3)

3 valores que dividen en 4 partes iguales la distribución de frecuencias:

Primer cuartil (Q_1) por debajo 25%, por encima 75% → $\boxed{Q_1=P_{25}}$

Segundo cuartil (Q_2) por debajo 50%, por encima 50% → $\boxed{Q_2= P_{50}= Md}$

Tercer cuartil (Q_3) por debajo 75%, por encima 25% → $\boxed{Q_3=P_{75}}$

Igual forma de cálculo que los percentiles.

Se utilizan para construir índices para el estudio de la variabilidad de una distribución de frecuencias.

Deciles (D_1) (D_2) (D_3) (D_4) (D_5) (D_6) (D_7) (D_8) (D_9)

9 valores que dividen en 10 partes iguales la distribución de frecuencias:

Primer decil (D_1) por debajo 10%, por encima 90%
Primer decil (D_2) por debajo 20%, por encima 80%
Primer decil (D_3) por debajo 30%, por encima 70%
…
Primer decil (D_9) por debajo 90%, por encima 10%

Igual forma de cálculo que los percentiles.

TEMA 3 – MEDIDAS DE VARIABILIDAD Y ASIMETRÍA

Dos nuevas propiedades de una distribución de frecuencias:

Variedad o dispersión: grado en que las puntuaciones se asemejan o diferencian entre sí, o se aproximan o se alejan de una medida de tendencia central como la media aritmética.

Índices de medida: **Amplitud total**
Varianza
Desviación típica
Amplitud semi-intercuartil
Coeficiente de covariación: para comparar distintas distribuciones de frecuencias en términos de su variabilidad

Asimetría o sesgo de la distribución:

Índice de asimetría de Pearson: resultado numérico sobre el grado y tipo de asimetría de la distribución.

Puntuaciones directas:

Para comparar a los sujetos entre sí y en diferentes variables.

Puntaciones diferenciales
Puntaciones típicas

Medidas de variabilidad

Variabilidad o dispersión: grado de variación en un conjunto de puntuaciones. Puntuaciones muy próximas entre sí (concentradas alrededor de la media)→ poca dispersión; puntuaciones alejadas entre sí→ más dispersión = mayor variabilidad. Cuanta menos variabilidad más homogénea es la muestra. Cuanta más variabilidad más heterogénea es la muestra.

Para cuantificarlo → medidas o índices de variabilidad:

Amplitud total o de rango y la amplitud semi-intercuartil: Los que miden el grado en que las puntuaciones se asemejan o diferencian entre si

Varianza y desviación típica: Los que miden la dispersión con respecto a alguna medida de tendencia central (media aritmética)

Amplitud total (rango o recorrido) (A_T): Distancia que hay en la escala numérica entre los valores que representan la puntuación máxima (*límite exacto superior del intervalo máximo*) y la puntuación mínima (*límite exacto inferior del intervalo mínimo*): $\mathbf{A_T = X_{máx} - X_{mín}}$

EJ.: $A_T = X_{máx} - X_{mín}$ → $9,5 - 4,5 = 5$

X_i	n_i
5	135
6	66
7	45
8	36
9	18
Σ	300

Inconvenientes: sensible únicamente a los calores extremos, por lo que no captura la poca o mucha dispersión entre los restantes valores.

Varianza y desviación típica: Distancia entre las puntuaciones y un valor central (media aritmética)
Poca variabilidad: medidas muy cercana a la media; mucha variabilidad: medidas alejadas de la media.

Promedio de las desviaciones o diferencias de cada puntuación respecto a su media ($\bar{X}_d$):

$$\bar{X}_d: \frac{\Sigma d_i}{n} = \frac{\Sigma (X_i - \bar{X})}{n}$$

Como se vio en la primera propiedad matemática de la media, el sumatorio del numerador siempre es igual a cero ($\Sigma (X_i - X)$), por lo que no sería una buena medida para la variabilidad. Para poder utilizar un índice con estas desviaciones evitando el cero:

1. **Desviación media (DM)**: calcular el valor absoluto de cada desviación antes de realizar la suma:

   $$DM = \frac{(X_1 - \bar{X}) + (X_2 - \bar{X}) + \ldots\ldots + (X_n - \bar{X})}{n} = \frac{\Sigma (X_i - \bar{X})^2}{n}$$

   Se usa muy poco porque es poco manejable matemáticamente.

2. **Varianza (S^2_x)**: otra alternativa para el problema del signo es el promedio de los cuadrados de las desviaciones de las puntuaciones respecto a la media

   $$S^2_x = \frac{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \ldots\ldots + (X_n - \bar{X})^2}{n} = S^2_x = \frac{\Sigma (X_i - \bar{X})^2}{n}$$

   Ó

   $$S^2_x = \frac{\Sigma X^2_i}{n} - \bar{X}^2$$

   Primero de elevan al cuadrado las diferencias y luego se obtiene el promedio de esas desviaciones al cuadrado

   Datos en frecuencias agrupadas o sin agrupar en intervalos:

   $$S^2_x = \frac{\Sigma n_i (X_i - \bar{X})^2}{\Sigma n_i} = \frac{\Sigma n_i (X_i - \bar{X})^2}{n}$$

   Ó

   $$S^2_x = \frac{\Sigma n_i X_i^2}{\Sigma n_i} - \bar{X}^2 = \frac{\Sigma n_i X_i^2}{n} - \bar{X}^2$$

   n= nº total de observaciones
   X_i= valor de i en la variable X o el punto medio del intervalo
   n_i= es la frecuencia absoluta del valor o intervalo i

   Datos en frecuencias relativas:

   $$S^2_x = \Sigma p_i X_i^2 - \bar{X}^2$$

   pi= frecuencia relativa o proporción de observaciones del valor o del intervalo i

3. *La varianza al trabajar con nº al cuadrado siempre es positiva que se expresa en las unidades de la variable al cuadrado, para lograr una medida de dispersión en las mismas unidades que la variable, se calcula la* ***raíz cuadrada de la varianza*** → ***desviación típica.***

$$S_x = \sqrt{\frac{\Sigma (X_i - \bar{X})^2}{n}}$$

La desviación típica se utiliza más que la varianza porque se expresa en las mismas unidades de medida que la variable objeto de estudio.

Propiedades de la varianza y de la desviación típica

.- Usan todas las puntuaciones observadas

.- Miden la variabilidad respecto a la media aritmética, solo deben utilizarse si se usa la media como medida de tendencia central.

.- Iguales o mayores (positivas) que cero. = 0 si todas las puntuaciones son iguales entre sí (no hay variabilidad o dispersión). Nunca negativas.

.- Si a una variable *X* se le suma o se le resta una constante *a*, la varianza y la desviación típica de la variable original no se ven afectadas, siguen siendo las mismas. Pero cuando multiplicamos los valores de las *X* por una constante *b*, la varianza queda multiplicada por la constante b^2 y la desviación típica por *b*.

4. **Cuasivarianza:** se divide por *n-1* en lugar de *n* como en la varianza

$$S^2_{n-1} = \frac{\Sigma (X_i - \bar{X})^2}{n-1}$$

5. **Cuasi desviación típica**: raíz cuadrada de la cuasivarianza

$$\sqrt{S^2_{n-1}} = \sqrt{\frac{\Sigma (X_i - \bar{X})^2}{n-1}}$$

Coeficiente de variación: Comparación del grado de variabilidad o dispersión entre dos conjuntos de puntuaciones en una misma o distintas variables. Por lo general, las variables se miden en unidades distintas y es necesario definir un índice de variabilidad relativa que no dependa de las unidades de medida.

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

Está definido para variables con la media $\bar{X} > 0$ y es recomendable que su resultado se acompañe de la media y de la desviación típica de la distribución a partir de las que se ha calculado.
Solo se puede utilizar cuando la media de ambos grupos es la misma.

Amplitud semi-intercuartil:

Varianza, desviación típica, media aritmética → Estadísticos para estudiar la variabilidad y la tendencia central.

Distribución asimétrica → mediana y amplitud semi-intercuartil

Distancia media entre el tercer y el primer cuartil → $Q = \frac{Q_3 - Q_1}{2} = \frac{P_{75} - P_{25}}{2}$

Informa sobre la variabilidad del 50 % de las puntuaciones, precisamente las comprendidas entre el percentil 25 y el 75 de la distribución.

INDICE DE ASIMETRÍA DE PEARSON

Asimetría: Grado en que las puntuaciones se reparten por debajo y por encima de la medida de tendencia central→ mediante la representación gráfica (positiva o negativa)→Índice numérico que lo cuantifica: **Índice de asimetría de Pearson: relación entre la media ($\bar{X}$) y la moda (Mo)**

$$A_s = \frac{\bar{X} - Mo}{S_x}$$

Índice adimensional (no tiene unidades de medida) que se aplica a distribuciones unimodales.
Distribución simétrica→ Media y Moda iguales y se anulan→$A_s = 0$
Asimetría positiva→Media mayor que la Moda→$A_s > 0$
Asimetría negativa→ Media menor que la Moda →$A_s < 0$

PUNTUACIONES TÍPICAS

Las puntuaciones directas (test, etc.) nos ofrecen poca información, conocida una puntuación directa no sabemos si se trata de un valor alto o bajo porque depende del promedio del grupo.
Puntuación diferencial (x_i)→ permiten comparar las puntuaciones de un sujeto en dos variables distintas. A una puntuación directa X le restamos la media de su grupo.

$$x_i = X_i - \bar{X}$$

Esta información nos permite saber si la puntuación coincide con la media de su grupo, es inferior o superior

Propiedades:

a) Su media es cero: $\bar{x} = 0$
b) La varianza de las puntuaciones diferenciales es = a la varianza de las puntuaciones directas

Puntuación típica (Z_x)→permiten comparar las puntuaciones de un sujeto en dos variables distintas y comparar dos sujetos distintos en dos variables distintas. Al proceso de obtenerlas se le llama tipificación. Indica el nº de desviaciones típicas que se apartan de la media de una determinada puntuación.

$$Z_x = \frac{x}{S_x} = \frac{X - \bar{X}}{S_x}$$

Propiedades:

a) Su media es cero: $\bar{Z} = 0$
b) Su varianza es 1

Reflejan las relaciones entre las puntuaciones con independencia de la unidad de medida. Permiten comparaciones entre distintos grupos y entre distintas variables.

TEMA 4 –ANÁLISIS CONJUNTO DE DOS VARIABLES

Dos variables con dos medidas c/u, en una muestra de 100 sujetos. Se obtiene una lista de 4 columnas y 100 filas

Caso	Nombre y Apellido	Género (X)	Estrés (Y)
1	Pepe Pérez	Varón	Sí
2	Quica, Chonco	Mujer	No
…	…	…	…
99	Ines Ayala	Mujer	Sí
100	Pablo Jota	Varón	Sí

Asociación o relación de dos variables

Dos variables están relacionadas entre sí cuando ciertos valores de una de las variables, se asocian con ciertos valores de la otra variable.

ASOCIACION ENTRE DOS VARIABLEA CUALITATIVAS

Variable cualitativa→se mide en escala nominal o de clasificación.

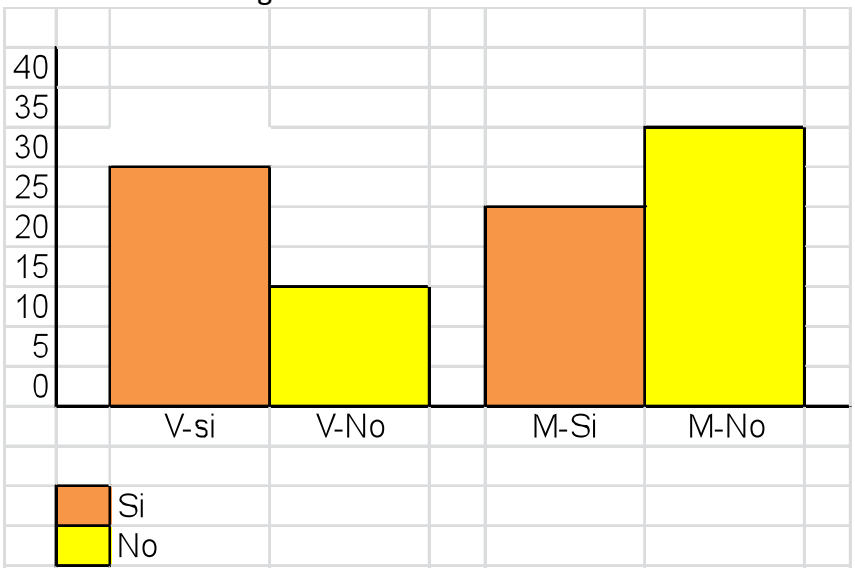
- ➔ Pueden ser dicotómicas (dos categorías)
- ➔ O politómicas (más de dos categorías)
- ➔ También son cualitativas variables que presentan un mayor nivel de medida 8intervalos o razón) pero han sido categorizadas.

Tabla de contingencia→ con los datos de dos variables cualitativas para todos los sujetos de una muestra.

Frecuencias observadas o empíricas (n_e) en X e Y

		Y		
		Sí	No	
X	V	30	10	40
	M	25	35	60
		55	45	100

Que se representa gráficamente en un diagrama de barras:



Para poder saber si existe o no **relación entre las variables**, se utiliza el **estadístico X^2**, asociado a una distribución de probabilidad (Chi cuadrado χ^2). **X^2** se define en función de las **frecuencias empíricas (n_e)** y de las **frecuencias teóricas (n_t)**→ que se calculan asumiendo que ambas variables son independientes o no relacionadas y son el producto del total de su fila por el total de su columna, dividido por la frecuencia total.

$$\text{Frecuencia teórica } (n_t) = \frac{\text{Total fila X Total columna}}{N}$$

Ej.:

		Y				
		Sí		No		
X	V	$\frac{40 \times 55}{100}$	22	$\frac{40 \times 45}{100}$	18	40
	M	$\frac{60 \times 55}{100}$	33	$\frac{60 \times 45}{100}$	60	60
		55		45		100

Y luego se elabora la diferencia entre las **frecuencias empíricas (n_e)** y las **frecuencias teóricas (n_t)**

		Y	
		Sí	No
X	V	8	-8
	M	-8	8

Que siempre tiene que dar cero. El valor -, nos indica una relación negativa.

Cálculo del **estadístico** $X^2 \rightarrow \frac{\Sigma (n_e - n_t)^2}{n_t}$

Inconvenientes: difícil interpretación, ya que desconocemos su límite superior. Sabemos que tiene valor cero cuando no hay relación entre las variables (*cuando las frecuencias empíricas y teóricas son iguales*).

Para resolver este problema se utiliza el **índice o Coeficiente de Contingencia, C** que toma valores 0 <= C < 1

$$C = \sqrt{\frac{X^2}{X^2 + n}}$$

El valor obtenido se puede comparar, dado que la tabla de contingencia tiene igual nº de filas que de columna (K) con una **C máximo** definido como:

$$C \text{ máx} = \sqrt{\frac{K - 1}{K}}$$

Ídem con dos variables cualitativas con más de dos categorías

Característica del coeficiente C:

1. Valores mayores, iguales a 0y menores que 1.
   0 cuando $X^2 = 0 \rightarrow$ dos variables sin relación (Frec. Empíricas = Frec. Teóricas)
   1 cuando $n = 0 \rightarrow$ no hay observaciones (nunca se puede dar)

2. A mayor valor de C, mayor es la relación entre las dos variables y al revés. Para usar el valor de C para comparar la relación entre dos variables de diferentes investigaciones es necesario que tengan el mismo nº de filas, de columnas y de datos.
3. Fundamentar la causalidad en un coeficiente de contingencia (hay variables que se relacionan entre sí porque existe otra variable ajena que tiene una relación clara con ambas
4. Se puede estimar un valor máximo de C si la tabla de contingencia tiene el mismo nº de filas que de columnas.

CORRELACIÓN ENTRE DOS VARIABLES CUANTITATIVAS

Requisitos: Muestra grande

Representaciones gráficas: **Diagrama de depresión o nube de puntos** (se puede apreciar si existe una relación lineal entre X e Y)

Índices para cuantificar la relación lineal:

Covarianza: Variación conjunta de dos variables→ ***Cov (X, Y)*** ó S_{XY}.

$$S_{XY} = \frac{\sum_{i=1}^{n} X_i Y_i}{n} - \bar{X}\,\bar{Y}$$

Xi= Valor de la variable X en el caso i
Yi= Valor de la variable Y en el caso i
$\overline{X}$ *= Media de la variable X*
$\overline{Y}$ *= Media de la variable Y*
n= número de casos de la muestra

El signo + - indica si la relación entre ambas variables es directa o inversa.

Relación lineal directa→ mayores valores de X, mayores valores de Y; menores valores de X, menores valores de Y (y viceversa) (+/+; -/-)

Relación lineal inversa→mayores valores de X, menores valores de Y; menores valores de X, mayores valores de Y (+/-; -/+)

Problemas *(Igual que el coeficiente X2, en las cualitativas)*→ Se desconoce su rango, sus valores máximos y mínimos, para evitar este problema→**Coeficiente de correlación de Pearson (r_{XY})→**

$$r_{XY} = \frac{S_{XY}}{S_X S_Y}$$

SX= Desviación típica de la variable X
SY= Desviación típica de la variable Y
SXY= Covarianza entre X e Y

Cociente entre la covarianza entre X e Y, y el producto de la desviación típica de X y de la desviación típica de Y.

Propiedades:

.- Solo toma valores comprendidos entre -1 y 1. Cero: cuando no exista relación entre X e Y.

.- $r_{XY} = +\_ 1$ si una variable es una transformación lineal de otra

Fórmula alternativa:

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

Para interpretar los resultados hay que tener en cuenta:

a) Valor absoluto→ a mayor valor absoluto, relación lineal entre las dos variables más fuerte.
b) Signo→positivo (+/+, -/-)→ relación directa
   negativo (+/-, -/+) → relación inversa

Problemas:

a) Solo detecta relaciones lineales, un coeficiente de correlación lineal cercano a cero indica que no existe correlación, pero pueden existir otro tipo de relaciones de carácter no lineal (relación curvilínea)
b) No tiene una comparación directa entre resultados de estudios diferentes rXY= 0 , no hay relación y r_{XY}=+_ 1, relación directa.
c) Dificultad para fundamentar la causalidad, cuando existe un coeficiente de correlación elevado entre dos variables, no se puede afirmar que una variable sea la causante de la otra.

REGRESIÓN LINEAL

Recta de regresión, para efectuar pronósticos de los valores de una variable a partir de la otra variable→ **Y=a + bX** (*b, pendiente; a, ordenada*)

Puntuaciones en Y a partir de puntuaciones en X→ $b = \frac{n \Sigma (XY) - \Sigma X \Sigma Y}{n \Sigma X^2 - (\Sigma X)^2}$

$$a = \overline{Y} - b\overline{X}$$

A las puntuaciones de **Y** obtenidas a través de esta ecuación →**Puntuaciones pronosticadas**

Propiedades:

a) La media de los errores es cero.
b) La media de las puntuaciones pronosticadas coinciden con la media de las puntuaciones reales de Y
c) La varianza de las puntuaciones en Y es igual a la suma de la varianza de los pronósticos, más la varianza de los errores

TEMA 5 –NOCIONES BÁSICAS DE PROBABILIDAD

CONCEPTOS

Experimento aleatorio: Proceso que puede repetirse indefinidamente en las mismas condiciones, cuyo resultado no se puede predecir con certeza.

Obtenemos un resultado (experimento) aleatorio (porque interviene el azar)→

a) Todos los resultados posibles son conocidos con anterioridad
b) No se puede predecir con certeza cuál será el resultado que se obtendrá
c) Puede repetirse cuantas veces se desee.

Espacio muestral (E) o *suceso seguro*: conjunto de todos los resultados posibles.

Sucesos (A, B, …): Resultados del experimento aleatorio o subconjuntos del espacio muestral.

Elementales (simple)→ un solo resultado del espacio muestral

Compuestos→Dos o más resultados del espacio muestral

Suceso Imposible (ϕ) o *conjunto vacío*: Suceso que no puede ocurrir nunca

Operaciones con sucesos

a) **Unión AUB→** Subconjunto de **E** formado por los elementos que pertenecen a A y pertenecen a B o a ambos a la vez.
b) **Intersección A∩ B→**Subconjunto de **E** formado solamente por los elementos pertenecientes a A y a B. Cuando la intersección no contiene ningún elemento, los sucesos son incompatibles o excluyentes, no pueden verificarse simultáneamente.
c) **Complementario $\overline{A}$→** Subconjunto de E formado por los elementos que no pertenecen al suceso A. se representa con el Diagrama de Venn.

DEFINICION DE PROBABILIDAD

Calcular la probabilidad de la ocurrencia de un suceso. Cero→ Suceso imposible, Uno→ Suceso seguro, otro suceso→entre 0 y 1

Clásica (Laplace): la probabilidad de un suceso es igual al cociente entre el nº de casos favorables de que ocurra ese suceso y el nº de casos posibles, en el supuesto de que todos los casos tengan la misma oportunidad de ocurrir (sean igualmente probables)

$$\textbf{Probabilidad de suceso} = \frac{\textbf{Nº de casos favorables}}{\textbf{Nº de casos posibles}}$$

Es necesario que los sucesos sean equiprobables

Estadística

Si repetimos el experimento aleatorio muchas veces y anotamos las frecuencias relativas (*Frecuencia absoluta (n_i) de una variable estadística X_i, es el número de veces que aparece en el estudio este valor, Frecuencia relativa (f_i), es el cociente entre la frecuencia absoluta y el tamaño de la muestra (N))*de un suceso, tiende a estabilizarse en un valor entre 0 y 1, que se denomina **probabilidad de suceso→***el límite al que tiene la frecuencia relativa de aparición de un suceso A cuando el nº de ensayos, n tiende a infinito.*

$$P(A)=\lim_{n\to\infty} \frac{n_A}{n}$$

Problema: Muchas veces no se puede o no es práctico, repetir el experimento un gran nº de veces.

Axiomática: Dado un espacio muestral E, se llama probabilidad de un suceso A, definido en el espacio muestral E, que se designa por P(A), a un nº real que se asigna al suceso A, que cumpla las siguientes condiciones:

a) 0= o < P(A) = o < 1 (propiedad cuantificable entre 0 y 1)
b) P(E)=1 (Cero cuando no puede ocurrir nunca y 1 cuando el suceso se produce con seguridad)
c) P(A)=1 – (□) La probabilidad de A se puede obtener restando de uno la probabilidad de su complementario, □.
d) Teorema de la suma: la probabilidad de que ocurra el suceso A o el suceso B es igual a la suma de la probabilidad de que ocurra el suceso A más la probabilidad de que ocurra el suceso B, menos la probabilidad de que ocurran ambos: P(AUB)= P(A) + P(B) –P(A□B)

A y B incompatibles→ P(AUB)= P(A) + P(B) (ya que P(A□B) = ~~o.~~

PROBABILIDAD CONDICIONADA

Cuando la aparición de un suceso A, depende de la aparición de otro B. Los sucesos A y B son **dependientes**.

P (A□B) donde B es la condición requerida→"probabilidad de A condicionada a B": Es igual a la probabilidad de la intersección dividido por la probabilidad de la condición B:

$$P(A \square B) = \frac{P(A \square B)}{P(B)} \quad \text{siempre que } P(B) \neq 0$$

$$P(B \square A) = \frac{P(B \square A)}{P(A)} \quad \text{siempre que } P(A) \neq 0$$

Si A y B independientes→ P (A□B)= P(A) y P(B□A) = P(B)

LA REGLA DEL PRODUCTO Y EL TEOREMA DE BAYES

Varios experimentos simultáneos…

Probabilidad condicionada:

$$P(B \square A) = \frac{P(A \square B)}{P(A)}$$

Si despejamos P(A□B) → **P(A□B)= P(A). P (A□B)**

La probabilidad de que ocurra A y B es igual a la probabilidad de la ocurrencia de A por la probabilidad de la ocurrencia de B, dado que A ha ocurrido previamente

P(A□B)→probabilidad de que ocurra B dado que ha ocurrido A

Si A y B independientes→ P(A□B)= P(A). P (B)

Se representa gráficamente con el **diagrama del árbol,** (*da todo lo que puedes combinar)* donde los nº corresponden a las probabilidades condicionadas al suceso que aparece antes. Se debe cumplir siempre que las sumas de las probabilidades que salgan de un mismo punto deben sumar 1.

Para calcular las posibilidades de intersección de dos sucesos hay que ir multiplicando las probabilidades de cada "rama", hasta que se llegue al extremo del árbol.

Teorema de Bayes:

A partir de que ha ocurrido el suceso B (ha ocurrido un accidente) deducimos las probabilidades del suceso A (¿estaba lloviendo o hacía buen tiempo?)

$$P(A|B) = \frac{P(A).\ P(B|A)}{P(B)}$$

Su importancia radica en los trabajos que ha generado y en la corriente denominada bayesiana.

Parte de una situación en la que es posible conocer las probabilidades de que ocurran una serie de sucesos. A esta se añade un suceso B cuya ocurrencia proporciona cierta información, porque las probabilidades de ocurrencia de B son distintas según el suceso A que haya ocurrido. Conociendo que ha ocurrido el suceso B, la fórmula del teorema de Bayes nos indica como modifica esta información las probabilidades de los sucesos A.

TEMA 6 DISTRIBUCIONES DISCRETAS DE PROBABILIDAD

En experimentos en los que no se pueden predecir los resultados.

Variable aleatoria

Función que asigna un número real (y solo uno) a cada uno de los resultados de un experimento aleatorio. Se puede definir de la manera que consideremos oportuna. Una vez definida la variable y obtenido el resultado, la función asigna un valor numérico inequívoco a este resultado. El resultado es aleatorio no la variable o función.

Se representa por letras mayúsculas: X; Y;

Y letras minúsculas para referirnos a los valores concretos que toman esas variables: x_2, y_1,

Discretas→ cuando solo puede tomar un conjunto infinito y numerable de valores (*Ej. nº naturales*) o finito de valores (*Ej. nº de caras al lanzar una moneda*)

Continuas→cuando puede tomar infinitos y no numerable.

Variables aleatorias discretas

1) **Función de probabilidad** ***f(x)***

X→ viene dada por los valores que puede tomar la variable aleatoria

Asocia a cada valor de la variable la probabilidad de que ésta adopte ese valor.

$$f(x) = P(X = x)$$

Representación gráfica→diagrama de barras.

Propiedades fundamentales:

1. Cualquier valor de x, siempre toma valores positivos o nulos.
2. La suma de todas las probabilidades es igual a 1.

2) **Función de distribución** ***F(x)***

Indica cual es la probabilidad de que la variable aleatoria tome un valor menor o igual que un valor concreto x.

Asocia a cada valor de la variable la probabilidad de que ésta adopte ese valor u otro inferior.

$$F(x) = P(X \leq x)$$

Si ordenamos de menor a mayor los valores x de la variable aleatoria discreta, se obtiene acumulando (sumando) los valores de la función de probabilidad:

$$F(x) = P(X \leq x) = f(x1)+f(x2)+...+f(x)$$

Representación gráfica→va dando saltos.

Propiedades fundamentales:

1. Todos los valores son positivos o nulos.

2. $F(x)$ es nula (vale 0) para todo valor inferior al menor valor de la varianza aleatoria.

$$F(x) = 0 \text{ si } x < x_1 \text{ (representa al menor valor)}$$

3. $F(x)$ es = 1 para todo valor igual o superior al mayor valor de la variable aleatoria.

$$F(x) = 1 \text{ si } x > x_k \text{ (representa al mayor valor)}$$

4. La función F(x) es no decreciente ya que es una acumulación o suma de probabilidades que son siempre positivas o nulas.

5. La probabilidad, P, de que la variable aleatoria X, tome valores x comprendidos entre x_1 y x_2 ($x_1 < x < x_2$) es la diferencia entre los valores de la función de distribución correspondientes a su valor superior menos su valor inferior.

$$P\ (x_1 < x < x_2) = F(x_2) - F(x_1)$$

Media y varianza de la variable aleatoria

Media (μ)→ *E(X)*, Esperanza matemática o Valor esperadosumatorio de cada uno de los valores que toma la variable por su función de probabilidad:

$$\mu = \Sigma\, x\,.\, f(x)$$

Promedio teórico que tomaría la variable aleatoria si se repitiese el experimento aleatorio infinitas veces.

Varianza σ^2 → *V(X):* Sumatorio del producto de cada uno de los valores que toma la variable menos su media elevada al cuadrado por su correspondiente valor de la función de probabilidad.

$$\sigma^2 = \Sigma\ (x - \mu)^2\,.\, f(x)$$

o

$$\sigma^2 = E(X^2) - [E(X)]^2$$

Dónde:

$E(X^2) = \Sigma\ x^2\ .f(x)$

$[E(X)]^{2=}$ la media elevada al cuadrado.

Desviación típica σ: raíz cuadrada de la varianza.

$$\sigma = \sqrt{\sigma^2}$$

DISTRIBUCIONES DISCRETAS DE PROBABILIDAD

Distribución Binomial *B(n, p)*→Variables aleatorias discretas que toman solo dos valores (dicotómicas) representados por 0 y 1.

Experimentos Bernouilli o binomial (éxito - fracaso) se repite "n" veces y de forma independiente.

Una variable aleatoria X sigue una distribución binomial (con parámetros n y p) si expresa el número de realizaciones independientes "n" con la probabilidad "p" y por tanto (1 – p) de obtener fracaso. Se representa por B(n, p), dónde **B** indica **binomial**, **n** el **número de ensayos** y **p** la **probabilidad de éxito**.

(Ej.: Ejemplo: Si tiramos tres veces la moneda al aire y definimos X como el número de caras, esta variable seguirá los parámetros n = 3 y p = 0,5. Lo mismo que B(3; 0,5))

Características Fundamentales:

1. *Función de probabilidad: $F(x) = P\ (X=x) = \binom{n}{x} p^x\, q^{n-x}$*

2. *Función de distribución:* $F(x) = P(X \leq x) = \binom{n}{x} p^x q^{n-x}$

3. *Media:* $\boldsymbol{\mu = np}$

4. *Varianza :* $\boldsymbol{\sigma = npq}$

Dónde x es el número de aciertos, n el número de ensayos, p la probabilidad de éxito de cada ensayo, q la probabilidad de fracaso (1-p) y el número combinatorio $\binom{n}{x}$ *que se lee "n sobre x" es igual a :*

$$\frac{n!}{x!\,(n-x)!}$$

Se utilizan las tablas II y III (en esta se presentan las probabilidades acumuladas) si tenemos una p ≥ 0,5, hay que intercambiar las condiciones de éxito y fracaso.

<u>Otras distribuciones discretas</u>

Existen otros modelos de distribuciones discretas. El modelo Poisson de los "sucesos raros", que se utilizan en condiciones similares a las binomiales pero con un elevado número de ensayos y un valor p muy pequeño.

TEMA 7 DISTRIBUCIONES CONTINUAS DE PROBABILIDAD

Modelos en los que se ajustan las variables con las que trabajamos→modelo normal

Modelos con implicación como instrumentos estadísticos→Chi-cuadrado de Person
t de Sudent
F de Snedecor

La distribución normal

Variable aleatoria que toma infinitos valores→variable aleatoria continua y ya no se puede hablar de que la variable tome un valor en concreto, sino que este dentro de un determinado intervalo.

Características y propiedades

La siguiente fórmula recoge la función:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$ Para - ∞ < x < ∞

Dónde μ y σ (media y desviación típica) son sus parámetros, π = 3,1416 y e = 2,718 y (base de los logaritmos neperiano).

Si una variable X tiene una distribución que se ajusta a la fórmula anterior, es una distribución normal y se expresa X→***N (μ y σ)*** indicando que tiene una distribución normal N con parámetros μ y σ.
Forma una campana que es más apuntada cuanto menor es su desviación típica.

Si una variable X le aplicamos una transformación lineal $Y = bX+a$, la nueva variable Y se distribuirá normalmente pero con media $b\mu x+ a$ y la desviación típica $|b|\sigma$. Si restamos la media y dividimos por la desviación típica obtenemos una nueva variable "*z*". Por tanto:

z→N (0,1)

Y su función de probabilidad:

$$f(z) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{z^2}{2}}$$ Para - ∞ < x < ∞ **Distribución normal tipificada (Tablas III y IV)**

Propiedades fundamentales:

a. Simétrica entorno a su media, μ, que coincide con su mediana y su moda.
b. La curva normal tiene dos puntos de inflexión; dos puntos donde la curva pasa de ser cóncava a convexa, situados a una desviación típica de la media.
c. Es asintótica en el eje de abscisas, se extiende desde - ∞ hasta + ∞ sin tocar nunca el eje.

Casos de utilización de las tablas

1. En el supuesto que la tabla no recoja el valor, podemos utilizar el más próximo.
2. Cálculo de la probabilidad para valores menores o iguales que una determinada puntuación típica: se mira directamente en la tabla.
3. Cálculo de la probabilidad para valores mayores que una determinada puntuación: se mira en la tabla la probabilidad que esa puntuación deja por debajo y se resta a 1.

4. Cálculo de la probabilidad entre dos puntuaciones determinadas: se restan las probabilidades que dejan por debajo de sí las dos puntuaciones típicas.

HISTOGRAMA Y DISTRIBUCION NORMAL

Si disponemos de los datos originales de una variable X, y su distribución es normal, utilizaremos las tablas III y IV, pero anteriormente transformaremos las puntuaciones directas en puntuaciones típicas:

$$z_i = \frac{X_i - \overline{X}}{S_x}$$

Siendo S la desviación típica

APROXIMACION DE LA BINOMIAL A LA NORMAL

Cuando las distribuciones binomiales superan sus valores de 20 ("n"< 20) se puede aproximar la binomial a la normal. Teniendo una variable X, con distribución binomial, su media es $\mu = \boldsymbol{np}$ y su desviación típica $\sigma = \boldsymbol{npq}$

Podemos realizar:

$$P(X = x) = P\left(\frac{(x-0,5) - \mu}{\sigma} \leq \frac{x-\mu}{\sigma} \leq \frac{(x+0,5) - \mu}{\sigma}\right)$$

$$P(X = x) = P\left(\frac{(x-0,5) - np}{\sqrt{npq}} \leq z \leq \frac{(x+0,5) - np}{\sqrt{npq}}\right)$$

A medida que aumenta n (intentos) mejora la aproximación.

Sumar y restar el valor 0,5 se llama **corrección por continuidad**, permitiendo utilizar las puntuaciones discretas como continuas. Se interpreta cada puntuación X como si fuera el punto medio de un intervalo, se intenta asegurar que el intervalo incluya los valores discretos de la binomial.

DISTRIBUCION CHI-CUADRADO DE PERSON

En la distribución de Chi cuadrado de Pearson una variable X con distribución X^2_1, X^2_2.....X^2_n pasa a ser $\boldsymbol{X = X^2_n}$

Su media y varianza valdrán $\boldsymbol{\mu = n}$ y

$$\boldsymbol{\sigma^2 = 2n}$$

Esta distribución se usa para contrastar si la distribución de una variable se ajusta a una distribución determinada.

Propiedades

1. Nunca adopta valores menores de 0.
2. Es asimétrica positiva pero a medida que aumentan sus grados de libertad se va aproximando a la distribución normal.
3. Para n > 30 la podemos aproximar a una distribución N(n, 2n).

En la **tabla V** se hallan algunos valores de las distribuciones X^2.

Ej.: En una variable con 5 grados de libertad, $X \rightarrow X^2_5$, el valor 11,07 deja por debajo de sí una proporción de 0,95, representándose de la siguiente manera: $_{0,95} X^2_5 = 11,07$

Ahora si quisiéramos calcular P (X > 11,07):

$$P (X > 11,07) = 1 – P (X < 11,07) = 1 – 0,95= 0,05$$

DISTRIBUCION "t" DE STUDENT

Siendo X e Y dos variables aleatorias independientes, donde X sigue una distribución N (0,1) e Y sigue una distribución X^2_n. La variable aleatoria $T = \frac{X}{\sqrt{Y/n}}$, sigue una distribución "t" con "n" grados de libertad y se

Expresa por $T \rightarrow t_n$

Su media siempre vale 0 $(\mu=0)$

Su varianza $\sigma^2 = \frac{n}{n-2}$

Cociente entre una variable N(0,1) y la raíz cuadrada de una variable X^2_n dividida por sus grados de libertad

Características:

1. Es simétrica, con μ = 0. Su forma es muy parecida a la N(0,1), aunque menos apuntada.
2. Puede tomar cualquier valor ($-\infty \leftrightarrow +\infty$).
3. A medida que aumentan los grados de libertad, la distribución se aproxima más a una distribución normal.
4. La curva es asintótica al eje de abscisas.

Se emplea en estadística inferencial en contrastes. En la **tabla VI** se muestran los valores de esta distribución.

DISTRIBUCION "F" DE SNEDECOR

Si X_1 y X_2 son variables aleatorias independientes, con distribución chi-cuadrdado con n_1 y n_2 grados de libertad respectivamente, entonces una nueva variable $F = \frac{X_1/n_1}{X_2/n_2}$

Sigue una distribución F con n_1 y n_2 grados de libertad $\rightarrow (Fn_1, n_2)$.

Siendo n_1 los grados de libertad del numerados y n_2 los grados de libertad del denominador.

Media: $\mu \rightarrow \frac{n_2}{n_2 - 2}$ para $n_2 > 2$

Varianza: $\sigma^2 \rightarrow \frac{2n^2_2 (n_1 + n_2 -2)}{n_1 (n_2 – 4) (n_2 – 2)^2}$ para $n_2 > 4$

Se emplea para el contraste de hipótesis.

Características:

1. Asimétrica positiva, nunca toma valores menores que 0.
2. Propiedad recíproca: Si X es una variable con distribución F con n_1 y n_2 grados de libertad, entonces la variable Y = 1/X es también una variable con distribución F con n1 y n2 grados de libertad: $_{1-p} F_{n1, n2} = \frac{1}{{}_p F_{n2, n1}}$ donde p es la probabilidad asociada al valor de la variable,

   Tabla VII sólo aparece la probabilidad de que X = 0,900; 0,950; 0,975 y 0,990.

TEMA 8 ESTIMACIÓN

Conceptos previos

Población se refiere al conjunto total de elementos que se quieren estudiar una o más características. Debe estar bien definida. Llamaremos **N** al número total de elementos de una población. También se suelen utilizar los términos **individuos, sujetos y casos** para referirnos a los elementos de la población.
Cuando se dispone de un censo (listado) de la población, se puede estudiar a todos ellos.
No siempre es factible estudiar a la totalidad de una población; por lo que se estudia un subconjunto de los elementos totales; es decir, una **muestra**. Llamaremos **n** al número de los elementos de una muestra.

El **muestreo** es un proceso de selección con el fin de obtener una muestra lo más semejante posible a la población y así obtener estimaciones precisas. El tamaño es una característica esencial; ya que debe ser lo suficientemente amplia para representar adecuadamente las propiedades de la población y reducida para que pueda ser examinada en la práctica.

Probabilistico: se conoce la probabilidad asociada a una muestra y cada elemento de la población tiene una probabilidad conocida de pertenecer a la muestra

Una forma de obtener una muestra de una población homogénea es utilizar:
1) El **muestreo aleatorio simple**; por el cual se garantiza que cada elemento de la población tenga la misma probabilidad de formar parte de la muestra. Primero se asigna un número a cada elemento y después mediante algún medio (sorteo, papeletas,...) se eligen tantos elementos como sea necesario para la muestra.

2) Cuando los elementos están ordenados o pueden ordenarse se utiliza el **muestreo sistemático**. Se selecciona al azar entre los que ocupan los lugares $\frac{N}{n}$

Ejemplo: N = 100; n = 5; 100/5= 20; escogeríamos los elementos situados en las posiciones 20. El riesgo de este muestreo es la falta de representación; que se pudiese dar, del total de los elementos.

3) Cuando topamos con una población heterogénea, utilizamos el **muestreo estratificado**. Se emplea cuando disponemos de información suficiente sobre alguna característica y podemos elegir una muestra en función del número de elementos según estas características o estratos.

4) Ante poblaciones desordenadas y conglomeradas en grupos, se emplea el **muestreo por conglomerados**; donde se van seleccionando de todos los grupos, subgrupos, clases… y finalmente de los elementos restantes la muestra.

5) De la unión del **estratificado y** del **conglomerado**, surge otro **muestreo el polietápico**.

No probabilístico: se desconoce, o no se tiene en cuenta, la probabilidad asociada a cada muestra y se selecciona la que más le parezca representativa al investigador.

1) El muestreo **por cuotas (accidental)** se basa en un buen conocimiento de los estratos o individuos más representativos para la investigación. Similar al estratificado pero carente del carácter aleatorio.

2) El muestreo **opinático (intencional)** muestra el interés por incluir en la muestra a grupos supuestamente típicos.

3) El **causal (incidental)** selección de los individuos de fácil acceso.

4) **Bola de nieve**; donde un elemento seleccionado lleva a otro y éste a otro y así sucesivamente hasta completar la muestra.

Una muestra es representativa si exhibe internamente el mismo grado de diversidad que la población y es aleatoria si los elementos han sido extraídos al azar de la población.

INFERENCIA ESTADÍSTICA

El valor estadístico obtenido de una muestra (como media) no será igual al valor del parámetro de población. Para inferir un parámetro a partir de un estadístico hay que aplicar herramientas estadísticas de tipo inferencial como la estimación por intervalo (intervalos de confianza) o contraste de hipótesis.

ESTIMACION DE LA MEDIA

La media muestral es una variable aleatoria que toma un valor u otro según la muestra (tendremos tantas medias como posibles muestras del mismo tamaño podamos extraer de la población. Su función de probabilidad es la **distribución muestral de la media.**

La **distribución muestral de un estadístico** es un concepto central, tanto de la estimación como del contraste de hipótesis.

Distribución muestral de la media

Una función de probabilidad queda caracterizada por su forma, su media y su varianza. La media de la distribución muestral de la **media** (μ_x) es igual a la media de la población (μ). La **varianza** de la distribución muestral de la media es $\frac{\sigma^2}{n}$ y la **desviación típica** de la distribución muestral de la media es:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

La forma de la distribución original de la media se parece a una distribución normal aunque la distribución original de la variable en la población no es normal.

Dado el muestreo aleatorio simple:

• Si la distribución de X en la población es normal con media μ y desviación típica σ, entonces la distribución muestral de la X es normal

$$\mu, \left(\frac{\sigma}{\sqrt{n}} \right)$$

• Si la distribución de X en la población no es normal con media μ y desviación típica σ, entonces la distribución muestral de la X tiende a la normal a medida que n crece (Teorema Central del Límite), siendo la aproximación buena para n ≥ 30.
La desviación típica de la población es una medida de la variabilidad de la variable X en la población
La desviación típica de la muestra (cuasidesviación típica) es una medida de la variabilidad de la variable X en la muestra.
La desviación típica de la distribución muestral de la media (error típico de la media) representa el grado de variabilidad entre los valores de las medias muestrales.
A mayor error típico, menor precisión en la estimación.

La media como estimador

Un estimador es un estadístico que se utiliza para estimar un parámetro.
Por lo que la media de la muestra es un estimador de la media poblacional; y el valor del estimador en una muestra se denomina estimación o estimación puntual.
La media muestral X es un estimador insesgado de la media poblacional (μ). El error típico de la media es un indicador de la precisión de la estimación de la media; Cuanto menor es la desviación típica de la población, menor será el error típico; cuanto mayor es "n", menor será el error típico; cuanto menor es el error típico, mayor es la precisión. Dependiendo de la desviación típica de la población y del tamaño de la muestra.

ESTIMACION DE LA PROPORCION

La obtención de la distribución muestral de la proporción es similar a la de la media.

Distribución muestral de la proporción

Sea X una variable que sólo toma valores 0 y 1, la proporción de la muestra P se define como:

$$P = \frac{\Sigma X}{n}$$

Dado el muestreo aleatorio simple, el estadístico proporción (P) se distribuye según una binomial con:

$$\mu_P = \pi \text{ y } \sigma^2_P = \frac{\pi(1-\pi)}{n}$$

Como P es la media de los valores de X en la muestra, según el **TeoremaCentral del Límite**, a medida que el tamaño crece, la distribución muestral de la proporción tiende a la normal con media π y varianza $\frac{\pi(1-\pi)}{n}$

Cuanto más alejado esté π de 0,5, más elementos debe tener la muestra para realizar la aproximación a la normal.

- La media de la distribución muestral de la proporción (μ_P) es igual a la proporción de la población (π)
- La varianza de la distribución muestral de la proporción es : $\sigma^2_P = \frac{\pi(1-\pi)}{n}$
- La desviación típica de la distribución muestral de la proporción (error típico de la proporción) es:

$$\sigma_P = \sqrt{\frac{\pi(1-\pi)}{n}}$$

La proporción como estimador

La proporción muestral (p) es un estimador insesgado de la proporción poblacional (π).

El error típico de la proporción, es un indicador de la precisión de la estimación de la proporción. Cuanto menor es el error típico, mayor es la precisión.

INTERVALOS DE CONFIANZA

La finalidad de un intervalo de confianza es estimar un parámetro desconocido de una población a partir de una muestra. Al estimar la media de la población a partir de una muestra, podemos cometer un error de estimación $|\bar{X} - \mu|$.

La estimación por intervalo consiste en acotar el error con una alta probabilidad ***1 – α (nivel de confianza)*** de forma que $|\bar{X} - \mu|$ no sea superior a un estimado máximo ***($E_{máx}$).***

El error de estimación máximo ($E_{máx}$) es función de la variabilidad de la variable en la población, del nivel de confianza (n.c.) y del tamaño de la muestra:

$$E_{máx} = z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}$$

dónde:

- $z_{1-\alpha/2}$ es función del n.c. = 1 – α y se obtiene en la tabla de la distribución normal tipificada (tabla IV).

- $\frac{\alpha}{\sqrt{n}}$ Es la desviación típica de la distribución muestral de la media, es decir, el error típico de la media.

- σ es la desviación típica de la población que es conocida

- n es el tamaño de la muestra.

A partir de esta ecuación deducimos tanto el tamaño de la muestra como los límites del intervalo de confianza. El tamaño de la muestra se obtiene despejando n de la ecuación.

$$n = \frac{z^2_{1-\alpha/2}\ \sigma^2}{E_{máx}^2}$$

Vemos que n depende de:

- La desviación típica de la población.
- El nivel de confianza.
- El error de estimación máximo.

Los **límites inferior (Li) y superior (Ls)** se obtienen a partir del $E_{máx}$:

$$L_i = X - E_{máx} \;//\; L_i = X - z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}$$

$$L_s = X + E_{máx} \;//\; L_s = X + z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}$$

El n.c. o probabilidad **1 – α** significa que si extrajésemos todas las muestras posibles de una población, calculásemos la media en cada una de ellas y el intervalo de confianza, una proporción **1 – α** de todos los intervalos de confianza contendrá la media poblacional y una proporción **α** no lo contendrá.

Tamaño de la muestra

Interesa que un intervalo sea lo más estrecho posible y con alta probabilidad. A mayor nivel de confianza mayor es el error de estimación máximo, por lo que más amplio será el intervalo y menos precisa será la estimación. Una forma de mantener y reducir el error de estimación máximo dado y aumentar el n.c., es aumentando n.

Otro factor que interviene es la variabilidad de la variable, cuanto mayor sea la desviación típica de la población, mayor debe ser n para alcanzar una misma precisión.

Para calcular el tamaño de la muestra desconociendo σ, hay que sustituir en la ecuación, la desviación típica por la cuasidesviación típica (S_{n-1}) y $z_{1-\alpha/2}$ por $t_{n-1,\ 1-\alpha/2}$ (**tabla VI**).

Aplicaciones

Los pasos para aplicar un intervalo de confianza son los siguientes:

- Establecer un error de estimación máximo para un nivel de confianza 1 – α.
- Obtener el tamaño de la muestra n para el error de estimación máximo especificado.
- Extraer una muestra aleatoria de tamaño n y medir la variable.
- Calcular el estadístico (es estimador del parámetro) con las medidas obtenidas.
- Calcular los límites del intervalo de confianza.

Intervalo de confianza para la proporción

El error de estimación máximo de la proporción es:

dónde:

- $z_{1-\alpha/2}$ es función del nivel de confianza 1 – α (**tabla IV**).
- $\sqrt{\frac{\pi(1-\pi)}{n}}$ es el error típico de la proporción: σ_p.
- π es la proporción de la población que no es conocida.
- n es el tamaño de la muestra y se debe cumplir $n\pi(1-\pi) \geq 5$ para la aproximación a la normal.

Los límites inferior y superior del intervalo de confianza se obtienen a partir del error de estimación máximo. Como desconocemos π, que es lo que precisamente queremos estimar, operamos con la proporción muestral P. Así, si en $E_{máx}$ sustituimos π por la proporción muestral P, los límites inferior y superior del intervalo de confianza son:

$$L_i = P - z_{1-\alpha/2}\sqrt{\frac{P(1-P)}{n}} = P - E_{máx}$$

$$L_s = P - z_{1-\alpha/2}\sqrt{\frac{P(1-P)}{n}} = P + E_{máx}$$

Y la probabilidad de obtener un intervalo de confianza que contenga al parámetro π es:

$$P\left(P - z_{1-\alpha/2}\sqrt{\frac{P(1-P)}{n}} \leq \pi \leq P + z_{1-\alpha/2}\sqrt{\frac{P(1-P)}{n}} \right) = 1 - \alpha$$