Analisis-de-datos.pdf

Albialbita

Introducción Al Análisis de Datos

1º Grado en Psicología

Facultad de Psicología
Universidad Nacional de Educación a Distancia

INTRODUCCIÓN
al
análisis de datos

ÍNDICE

T1: CONCEPTOS BÁSICOS Y ORGANIZACIÓN DE DATOS

Algunos usos, aplicada en casi todas las disciplinas (CCSS y salud, ↑ exponencialmente los últimos 30 años):
- Estudios epidemiológicos (medicina).
- Estudios toxicológicos relacionados con la eficacia de medicamentos (farmacia).
- Estudios genéticos y de impacto ambiental (biología).
- Muestreos en prospecciones petrolíferas o hidráulicas (geología).
- Censos de población e inf. demográfica (sociología).
- Estudios sobre la optimización del coste-beneficio (economía).
- **Medición de V y evaluación diagnóstica de tratamientos, programas educativos, sociales… (Psicología).**

- **Estadística teórica**: ocupa aspectos matemáticos formales y normativos.
- **Estadística aplicada**: aplicación a un campo concreto. *Bioestadística, psicoestadística, socioestadística…* Algunos, le han puesto el nombre de Análisis de datos.

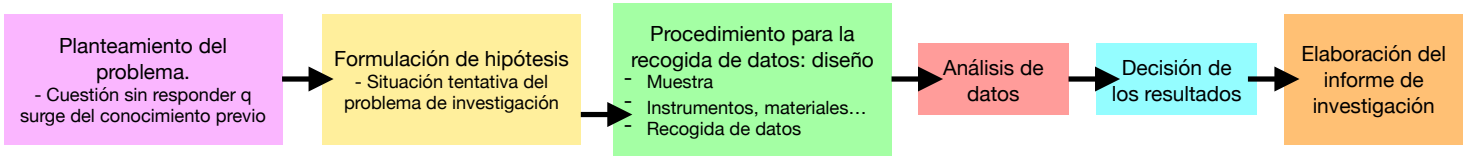
Se tiene la imagen como una rama de las matemáticas de difícil comprensión y ajena a nuestro día a día. Sin embargo, diariamente estamos sometidos a un bombardeo de datos estadísticos. El no ser capaz de distinguir una interpretación rigurosa de unos datos de una defectuosa, hace q se sea vulnerable a la manipulación. En ocasiones, las estadísticas presentadas en distintos medios (de com., políticos, publicidad, entorno laboral...) son incorrectas o engañosas por falta de preparación o por voluntad de «maquillar». Benjamín Disraeli (1er ministro del RU): *hay 3 tipos de mentiras: las mentiras, las grandes mentiras y las estadísticas*.

Nos proporciona las herramientas necesarias para valorar de manera crítica la inf. q recibimos.

1. LA INVESTIGACIÓN EN PSICOLOGÍA

Con la aparición de la ciencia moderna en el siglo XVII, el método científico pasó a ser la fuente de conocimiento más utilizada (hasta entonces eran el mito, sentido común o folclore popular). Las ciencias se distinguen entre sí por su objeto de estudio, pero tienen en común el método científico. La Psicología se sirve para acercarse a la conducta.

Método científico: procedimiento estructurado q utiliza la ciencia para la ampliación de sus conocimientos. Sistemático (tiene unas etapas definidas) y replicable (los datos obtenidos tienen q poder ser refutados -en las mismas circunstancias- por cualquier investigador interesado). Proporciona una manera de actuar para afrontar una investigación, a través de las sig. fases interdependientes:



1. Se define un problema, q puede surgir de teorías ya establecidas, de la lectura de la bibliografía o de la exp. directa con los hechos. En la mayoría de casos de lagunas o contradicciones en investigaciones anteriores.
2. Se plantea una hipótesis, q no es más que una solución tentativa al problema.
3. Las siguientes 3 fases contrastan si la hipótesis planteada es compatible con los hechos. Necesario establecer un procedimiento adecuado de recogida de inf., analizar los datos obtenidos y discutir los resultados en busca de conclusiones.
4. Elaborar un informe para dar a conocer los resultados.

Nos ocuparemos de la 4ª y 5ª fase; aprender a procesar los datos recogidos en una investigación con el fin de obtener la inf. q se precisa para contrastar la hipótesis formulada, y poder dar R al problema.

Diversos estudios ponen de manifiesto el efecto q la ansiedad ante los exámenes puede tener en la calificación PAU. Equipo investigador diseña un programa de tratamiento para paliar este efecto, q combina técnicas de estudio y relajación. Se ha selecciona a 40 estudiantes con este problema, voluntarios. 1/2 se ha asignado aleatoriamente al G1 (sin tratamiento) y 1/2 al G2 (tratamiento). Al finalizar el curso, se recogieron datos sobre las V relevantes de la investigación, además de algunas sociodemográficas, como sexo, nivel de estudios de la madre, opción de bachillerato elegido y h de estudio semanales.
¿cómo relacionaría los datos de este ejemplo con las fases de una investigación?

El problema objeto de estudio es se comprobar si el tratamiento influye en la calificación obtenida en la prueba. Hipótesis; el grupo investigador espera q sea eficaz, q el G2 tenga ↑ rendimiento en el examen q el G1. En la sig. fase se encontraría la det. de un plan de trabajo o procedimiento para la recogida de datos; decide escoger como muestra a 40 estudiantes con problemas de ansiedad ante los exámenes asignándolos de manera aleatoria a los grupos 1 y 2 comparando después sus resultados = análisis de los datos obtenidos y la discusión de dichos resultados. Se analizarían las calificaciones por ambos G. Se calcularía y compararía la media de ambos. Además, en otras investigaciones con otros objetivos, podría ser interesante plantear otro tipo de análisis, como cuantificar la

relación entre el n° de h estudiadas y la calificación en el examen de Lengua, o realizar pronósticos en el rendimiento en el examen en función de la ansiedad ante los exámenes y el n° de horas estudiadas (ambos procedimientos se estudiarán en el Tema 5 del programa). Por último, para difundir los resultados se elabora un informe.

El análisis de datos constituye una parte integral no solo de la activ. investigadora, tb en la práctica profesional. Resulta crucial tener unos conocimientos básicos para evaluar los resultados de una investigación, y leer de forma crítica las publicaciones de carácter psicológico *(artículos científicos, libros, informes de investigación o notas de prensa)*.

2. CONCEPTO Y FUNCIONES DE LA ESTADÍSTICA: DESCRIPCIÓN E INFERENCIA

Es la rama de las matemáticas q se encarga del estudio de det. carac. en una población, recogiendo los datos, agrupándolos, organizándolos en tablas, representándolos gráficamente y analizándolos para sacar conclusiones de dicha población.

2 grandes áreas: la Estadística Descriptiva y la Estadística Inferencial.

- **Estadística Descriptiva:** organiza y resumen conjuntos de observaciones cuantificadas procedentes de una muestra o de la población total. Mediante tablas, gráficos o valores numéricos. Se dispone de distintos procedimientos q nos permiten estudiar las cara. de 1 o + V:
  - **1V**: estadísticos q nos indicarán cuáles son los valores + habituales de esa V (de tendencia central), hasta q punto esos valores son similares o dif. entre sí (de variabilidad), en q grado las observaciones se reparten equilibradamente por encima y por debajo de la tendencia central (de asimetría) y cómo de apuntada es la distribución de las puntuaciones de la V (de curtosis).
  - **2V:** índices q nos indiquen hasta q punto están ambas V relacionadas entre sí (de asociación), procedimientos para predecir el valor de una V en función de otra (ecuaciones de regresión).
- **Estadística Inferencial**: se realizan inferencias acerca de una población basándose en los datos obtenidos a partir de una muestra. Estas generalizaciones a la población se basan en el cálculo de probabilidades.

En una investigación cualquiera, lo habitual es q se desee conocer un parámetro o carac. de los elementos de una población; sin embargo, suele ser demasiado extensa para estudiarla al completo *(conllevaría un coste inabordable)*. Por esto, se realiza un muestreo con el q se obtiene un conjunto de elementos q la representan y se estudia la carac. deseada en la muestra mediante estadísticos q se utilizarán para estimar los parámetros de la población.

- **Población**: conjunto de todos los elementos q cumplen una det. carac. objeto de estudio. *Niños con Trastorno por Déficit de Atención e Hiperactividad (TDAH) de la CCAA de Madrid.*
- **Muestra**: subconjunto cualquiera de una población. Personas, animales o cosas q cumplan una definición compartida por la población.

¿Por qué elegir un subconjunto y no trabajar con la población? Cuestión de viabilidad, no es posible trabajar con la población completa. El nº puede ser demasiado grande, puede haber familias (elementos de la población) quno deseen participar… Se trata de caracterizarla población. *Nos puede interesar conocer la inteligencia de los niños con TDAH. Utilizaríamos un test de inteligencia, el WISC-IV; al no tener acceso a la población completa de niños con TDAH se extrae una muestra de dicha población para obtener el nivel de inteligencia de cada niño de la muestra.*

- **Parámetro**: índice medido en una población q la describe de alguna manera.
  Propiedad descriptiva (una medida) de una población. Se denota con letras griegas; μ para la media, o^2_X para la varianza y n para la proporción)
- **Estadístico**: índice medido en una muestra. Utilizando la estadística inferencial se pronostica el valor de los parámetros poblacionales a partir de los estadísticos muestrales. *Niños con TDAH se calcula la media en inteligencia de los de la muestra, q es el estadístico X, para pronosticar el valor medio en inteligencia de la población, parámetro μ (el valor q nos interesa)*.
  Con letras latinas; X para la media, S^2_X para la varianza y P para la proporción.

3. VARIABLES: MEDICIÓN Y CLASIFICACIÓN

Variable: conjunto de valores resultantes de medir una carac. de interés sobre cada elemento indiv. de una población o muestra. Para representarlas se utilizan letras latinas mayúsculas. Para referirnos a un valor cualquiera de la variable X se utiliza el subíndice i (X_i); *n* el nº de elementos q componen la muestra. De manera genérica, se la designa como:

X_i siendo i = 1,2,3…, *n*

Si se trata de objetos físicos, el proceso de medición es directo y generalmente sencillo. Es cuestión de seguir unas reglas prescritas expresadas mediante det. escalas. *Medir la estatura de una persona asignando el nº correspondiente de la cinta métrica a la distancia q hay desde sus pies hasta su cabeza.* Cuando se trata de medir la timidez de un estudiante en una situación de interacción social, ya no es tan sencillo. El reto de la Psicología es su necesidad de medir en muchas ocasiones variables q no son directamente observables.

Medición: proceso por el cual se asignan nº a objetos o sucesos según det. reglas. Previo al análisis de datos, especifica el procedimiento de asignación de nº a los valores de la variable. *2 valores de la variable sexo (hombre y mujer) —> nº 1 y 2, y al peso de una rata el nº en gramos q da la balanza.* Para medir V psicológicas en muchas ocasiones se utilizan test psicológicos diseñados para ese fin. Su aplicación proporciona una puntuación para cada persona en esa variable.

Valoración de la calidad de vida, medida a través de una pregunta q forma parte de un test amplio y q se incluye en bastantes investigaciones sobre salud:
¿cómo calificarías tu calidad de vida? A) Muy mala. B) Regular. C) Normal. D) Bastante buena. E) Muy buena.
Asignar un nº a cada una de las opciones de R.

Psicología utiliza dif. escalas de medida (conjunto de reglas o modelos desarrollados para la asignación de números a las V) en función de la V a medir. *Escala centígrada de temperatura, q se basa en asignar 0º a la temperatura de congelación del agua y 100º a la de ebullición.*

En función de las relaciones matemáticas q puedan verificarse empíricamente entre los distintos valores de una V y, siguiendo la clasificación de Stevens (1946), pueden distinguirse 4 tipos de niveles o escalas de medida:

Escala nomial	Solo distinguiremos la igualdad o desigualdad entre 2 valores. Consiste en la asignación, puramente arbitraria de nº o símbolos a cada uno de los valores de la V. La única relación q se tiene en cuenta es la de igualdad (y desigualdad), q implica la pertenencia o no a una categoría det. Los valores de la variable se denominan **categorías**. Podemos decidir si un sujeto es igual o dif. a otro, pero no podemos establecer relaciones de orden respecto a esa V, ni de cantidad. *V enfermedad, distinguiendo entre: (1) «sanos» y (2) «enfermos», carece de sentido establecer relaciones entre estos 2 nº del tipo 1 + 1 = 2, sería considerar algo como q dos personas «sanas» = 1 «enferma».* Se puede asignar a cada valor de la V cualquier tipo de símbolo. **V cualitativas o categóricas**: Presentan un nivel de medida nominal. Se clasifican además, en función del nº de categorías q presentan. 2 = dicotómica *(sexo)*; +2 = politómica *(estado civil)*. En ocasiones se categorizan V q podrían medirse a un nivel superior; se ha dicotomizado si se han establecido 2 categorías, y politomizado si se han establecido +2 categorías. *V peso del roedor de un exp.: aunq podríamos medir exactamente su peso en gramos, puede resultar útil en una investigación dicotomizarla clasificando a las ratas en peso alto y bajo, o politomizarla, estableciendo 3/+ niveles de peso.*
Escala ordinal	Añade la posibilidad de establecer un orden. Se asignan nº a objetos para indicar la extensión relativa en q se posee una carac. sólo para indicar el orden de las posiciones. Los datos pueden utilizarse para jerarquizar u ordenar las observaciones, pero sin indicar la distancia q hay entre las posiciones. • Permite la identificación y diferenciación de los sujetos. • Permite establecer relaciones del tipo «mayor q» o «menor q», aunq no se plantea una distancia entre unas medidas y otras. La asignación de nº a las distintas categorías no puede ser completamente arbitraria, debe hacerse atendiendo al orden existente entre éstas. *V severidad de la enfermedad: 1 leve, 2 moderado y 3 grave. No es lo mismo padecer una con una intensidad leve o grave, y q la intensidad de la enfermedad en el caso de grave es mayor.* **V cuasicuantitativas** .
Escala de intervalo	Se usa 1 unidad y tienen sentido las diferencias. Ordenan los objetos según la magnitud del atributo q representan y proveen intervalos iguales entre las unidades de medida. Existencia de una unidad de medición común y constante. El origen es arbitrario, y no refleja en ningún momento ausencia de la magnitud que estamos midiendo. Los nº asignados: • Permiten decidir si un objeto es igual o dif. a otro. • Si posee en mayor o menor grado la carac. de interés. • La distancia entre los distintos valores consecutivos de la V es la misma. *La inteligencia medida con un test. 4 personas (A, B, C y D) han obtenido 80, 90, 150 y 160 puntos, podemos decir q la dif. en inteligencia entre A y B es la misma q entre C y D (90-80 = 160-150), el test proporciona una unidad de medida estable. Sin embargo, no se puede afirmar q D sea el doble de inteligente q A aunq tenga el doble de puntuación en el test, para realizar una afirmación de ese tipo sería necesario q el cero de la escala fuera absoluto*. En este caso es arbitrario pq obtener un 0 no refleja ausencia de la carac. medida, no significa q no se posea ni un ápice de inteligencia.
Escala de razón	Se pueden comparar 2 medidas mediante un cociente. Los nº asignados admiten como válidas las relaciones de igualdad-desigualdad, orden, suma, resta, multiplicación y división. Tiene todas las carac. de una medida de intervalo y se le puede asignar un punto de origen verdadero de valor 0 (= ausencia de la magnitud q estamos midiendo). El 0 ya no es arbitrario, sino un valor absoluto, se puede afirmar q A tiene 2, 3 o 4 veces la magnitud de la propiedad presente en B. Altura y peso. *Si una rata de laboratorio pesa 350 gramos y otra 175, la 1ª pesa el doble q la 2ª.*

En muchas ocasiones el nivel de medida de una V va a depender de cómo se haya definido. *La V calificación obtenida en el examen de Lengua de la PAU puede suscitar dudas razonables sobre su nivel de medida. Si se define como el nivel de conocimientos de Lengua necesarios para ingresar en la universidad se trataría de una de intervalo pq con esta interpretación el 0 sería arbitrario (0 ≠ ausencia total de los conocimientos necesarios sino q se ha obtenido un rendimiento nulo en las preguntas en concreto). Si en lugar del nivel de conocimientos, nos interesa simplemente contar el nº de aciertos se trataría de razón, aquí el 0 sí es absoluto e indicaría ausencia absoluta de preguntas acertadas.*

Es muy importante la definición operativa de una V (cómo se define y se registra) pq puede det. su nivel de medida.
La mayoría de las V psicológicas se considera q están medidas en una escala de intervalo. *Perseverancia, rasgo de personalidad, se ha medido mediante una prueba psicológica o test. Sin embargo, si se define como el nº de intentos o ensayos q realiza una persona para conseguir un objetivo se trata de una escala de razón.*

Tipo de variable	Escala de Medida	Características básicas	Relaciones válidas	Ejemplos
Cualitativa: ▪ Dicotómica ▪ Politómica	→ Nominal	Los números identifican y clasifican objetos	Relaciones del tipo «igual que» o «distinto que»	Sexo, estado civil, raza, diagnóstico clínico.
Cuasicuantitativa	→ Ordinal	Además, los números indican las posiciones relativas de los objetos	Además, relaciones del tipo «mayor que» o «menor que»	Dureza, posición en el ranking de la ATP, grado de satisfacción.
Cuantitativa: ▪ Discreta ▪ Continua	Intervalo	Además, hay una unidad de medición común	Además, igualdad o desigualdad de diferencias	Temperatura en grados centígrados, inteligencia.
	Razón	Además, el punto cero es absoluto	Además, igualdad o desigualdad de razones	Longitud, peso, altura, tiempo de reacción.

Todas las puntuaciones de test psicológicos se consideran de intervalo (V cuantitativa)

Las V medidas en escala de intervalo y de razón son cuantitativas.
En función de los valores numéricos q pueden asignarse:

- **V continua**: para la q, dados 2 valores, siempre se puede encontrar un 3ero q esté incluido entre los 2 primeros.
  El peso, entre los valores 79 y 80 kg. se pueden considerar 1, 2, 3 o todos los decimales q se quiera.
- **V discreta**: adopta valores aislados. Fijados 2 valores consecutivos, no se puede tomar ninguno intermedio.
  Nº de hijos.

4. DESCRIPCIÓN DE V: DISTRIBUCIÓN DE FRECUENCIAS Y REPRESENTACIÓN GRÁFICA

Una vez q el investigador ha recabado la inf. a través del proceso de medida y recogido los datos correspondientes, dispone de un listado o base, comúnmente llamado **matriz de datos**. La generación de una base de datos supone la codificación previa de las observaciones, la introducción de estos en algún programa informático, depuración de los datos ya grabados (detección y tratamiento de los errores de grabación y valores perdidos), y eventualmente la realización de transformaciones de V q faciliten su posterior tratamiento estadístico. Hay muchos programas estadísticos para organizar y analizar los datos. En concreto, Excel para hacer distribuciones de frecuencia, gráficos y diversos análisis.

Codificar datos es asignar nº a las V cualitativas y cuasicuantitativas, y registrar los valores de las V cuantitativas q constituyen la BD, así como asignar un código (espacio en blanco o valor numérico) a los perdidos (no registrados u observados). En la matriz de datos, los casos se sitúan en las filas y las V en las columnas.

Una vez que los datos están codificados es preciso realizar una depuración de la BD, q conlleva el procesamiento de:

- Los datos perdidos; valores q no han sido registrados, habitualmente pq el participante no ha consignado ese dato. Existen procedimientos de imputación de datos, basados en los valores válidos de otros casos q se utilizan en ocasiones en V cuantitativas.
- Datos atípicos: valor muy dif. al resto de valores de la misma V. Suelen ser ocasionados por errores al introducir los datos o valores extremos. Distorsionan los resultados de los análisis, hay q identificarlos y tratarlos de manera adecuada, generalmente excluyéndolos.

Si los datos han sido registrados manualmente en un software es recomendable hacer un control de calidad de la grabación de los mismos, revisando la codificación de un porcentaje de los casos, habitualmente un 5%-10% del total.

Una vez depurada, se utiliza para extraer la inf. relevante. Si tenemos muy pocos datos es posible q la simple inspección visual sea suficiente para describir el fenómeno estudiado. Pero no es nada frecuente. Habitualmente el número de datos es elevado, por lo que se hace necesario organizar la inf. mediante una **distribución de frecuencias**: tabla en la q se resume la inf. disponible de una V. Se sitúan los valores de la V por filas y en las columnas se dispone el nº de ocurrencias por cada valor, porcentajes…

- Facilita la lectura de la inf. q contienen los datos, organización de estos.
- Ofrece la inf. necesaria para realizar representaciones gráficas.
- Facilita los cálculos para obtener los estadísticos q serán objeto de estudio en los próximos temas.

Descripción de variables cualitativas

V cualitativa: distribución de frecuencias y su representación gráfica mediante un diagrama de barras o de sectores.
En la 5ª columna aparece el Bachillerato elegido. Sin embargo, la simple inspección visual no es suficiente para hacerse una idea precisa de cuántos estudiantes han elegido cada una de las modalidades.

En la distribución de frecuencias de variables cualitativas habitualmente se muestran las **frecuencias absolutas**, **relativas y** los **porcentajes**.

1. Se inspeccionan los valores q toma la V. *En este caso cualitativo (nominal) q puede adoptar 3 valores.*
2. En la 1ª columna se especifican los valores q adopta la variable X o el nº asignado a ese valor.
3. **Frecuencia absoluta** (n_i) q es el nº de observaciones en cada categoría. La suma de todas representa el total de la muestra (n)
4. **Frecuencia relativa o proporción** (p_i): q se obtiene dividiendo la frecuencia absoluta, n_i, entre el nº total de observaciones, q se representa por n**.**
5. La frecuencia relativa tb se expresa en términos de **porcentaje** (P_i); multiplicar cada una de las proporciones x100.

X	n_i	p_i	P_i
1. Ciencias y Tecnología	13	0,325	32,5
2. Humanidades y CC Sociales	21	0,525	52,5
3. Artes	6	0,15	15
Σ	40	1	100

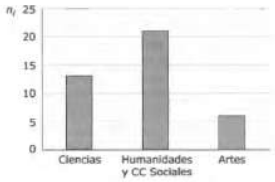
$p_i = n_i / n$
$P_i = p_i \times 100$

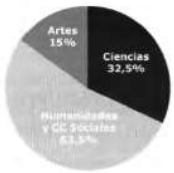
Ahora sí podemos hacernos una idea de la distribución de los estudiantes según el Bachillerato quhan elegido; el más demandado es el de Humanidades y CCSS (52,5%) y el menos demandado es el de Artes (15% del total).

Los 2 gráficos + habituales en la descripción de V cualitativas son los:

- **Gráficos de barra**: los distintos valores de la V se sitúan en el eje horizontal y las frecuencias/porcentajes en el eje de ordenadas. Cada barra representa una categoría, siendo su altura igual a su frecuencia (o porcentaje).
- **Gráficos de sectores**: cada sector representa una categoría de la V y su ángulo central debe ser proporcional a su frecuencia (o porcentaje).

El diagrama de barras se ha construido sobre las frecuencias absolutas de la variable y el diagrama de sectores sobre los porcentajes.





El único índice apropiado para V cualitativas es la moda.

Descripción de variables ordinales o cuasicuantitativas

En las V ordinales se procede de la misma manera, aunq con los valores situados en la tabla de acuerdo a un det. orden. *V nivel de estudios de la madre presenta los valores: Primarios, ESO, Bachillerato, Grado universitario y Posgrado universitario.*
En la distribución de frecuencias hay que preservar este orden, ya sea empezando por el valor + bajo o + alto de la V:

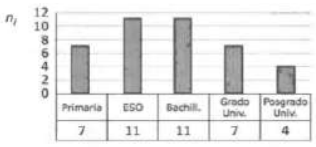
3 columnas +, para obtener estos valores, simplemente hay q ir acumulando (sumando), desde la cat. de menor valor a la de mayor. *Fecuencia absoluta acumulada en Bachillerato es 29, resultado de + las frecuencias de los valores anteriores (7 + 11 = 18) y la suya propia (18 + 11 = 29), 29 personas presentan un nivel de estudios de Bachillerato o inferior.* En las nominales carece de sentido el cálculo de las frecuencias acumuladas, ya q sus valores no establecen un orden det.

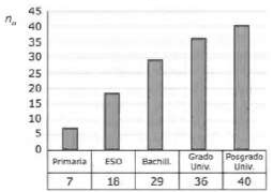
X	n_i	p_i	P_i	n_a	p_a	P_a
1. Primarios	7	0,175	17,5	7	0,175	17,5
2. ESO	11	0,275	27,5	18	0,450	45
3. Bachillerato	11	0,275	27,5	29	0,725	72,5
4. Grado universitario	7	0,175	17,5	36	0,900	90
5. Posgrado universitario	4	0,1	10	40	1	100
Σ	40	1	100			

- **Frecuencia absoluta acumulada** (n_a): nº de veces q se repite cada valor o cualquiera de los valores inferiores.
- **Proporción acumulada o frecuencia relativa acumulada** (p_a): cociente entre la frecuencia absoluta acumulada y el total de observaciones.
- **Porcentaje acumulado** (P_a) : valor de la frecuencia relativa acumulada multiplicado por cien.

$P_a = n_a / n$.
$P_a = P_a \times 100$

Generalmente tb representadas con un diagrama de barras o de sectores.
El diagrama de barras tb se puede realizar sobre las frecuencias, proporciones o porcentajes acumulados, respetando el orden.





Algunos índices apropiados son la mediana y la moda y la amplitud intercuartil.

Descripción de variables cuantitativas

Puede pasar q el nº de la variable sea:

1. Muy reducido; *nº de hijos*. Se ordenarán las frecuencias procediendo de forma indicada como para las ordinales.
2. Muy amplio; *V ansiedad ante los exámenes*. **Se agrupará la V en intervalos**: se forman grupos con valores consecutivos. Cada uno de estos grupos es situado en una fila y se calculan las frecuencias de cada uno o intervalo de valores y no de cada valor de la V.
   - Decidir el nº de intervalos. Siempre habrá muchas posibilidades; desde un nº pequeño a muchos intervalos de muy pequeña amplitud. Hay q tener presente q siempre se pierde inf. Dependerá del tratamiento q el investigador quiera dar, encontrando el equilibrio entre precisión del estudio y manejabilidad de datos.
   - Así, se formarán los **límites aparentes del intervalo**. *Entre 1 y 5, 5 y 10…* Para cada intervalo existe un límite superior e inferior.
     Tienen la misma unidad de medida q los valores de la V. Si los datos son enteros, los límites aparentes son enteros. Si contienen decimales, los tendrán el mismo nº de decimales q los datos recogidos.

Con los límites aparentes, existirá una discontinuidad entre un intervalo y el siguiente, ya q el límite superior de un intervalo no coincide con el inferior del siguiente intervalo. Con los límites exactos de una distribución no existe discontinuidad, ya q el superior exacto de un intervalo coincide con el inferior exacto del intervalo sig.

- **Límite Inferior Exacto** (LIE): restando al valor del límite inferior aparente media unidad de medida.
- **Límite Superior Exacto** (LSE): sumando al valor del límite superior aparente media unidad de medida.

Los límites exactos del intervalo 1-5 son 0,5-5,5, los del intervalo 6-10 son 5,5-10,5 y así sucesivamente.

A partir de los límites aparentes o exactos se calcula el **punto medio del intervalo**; la semi-suma del límite superior e inferior. Se utilizará para algunos cálculos de índices estadísticos con distribuciones agrupadas en intervalos.

$$PM = \frac{LIE + LSE}{2} \quad \text{ó} \quad PM = \frac{LIA + LSA}{2}$$

- Si los límites aparentes son enteros, la unidad de medida de la V es 1, y su mitad es 0,5, cantidad q habrá q restar al límite inferior y sumar al superior para calcular los límites exactos.
- Si los límites aparentes son nº con un decimal, la unidad de medida de la V es 0,1, por lo q la cantidad a sumar y restar para calcular los límites exactos será 0,05.
- Si los límites aparentes son nº con 2 decimales, la unidad de medida de la V es 0,01, por lo q la cantidad a sumar y restar para calcular los límites exactos será 0,005.
- Así sucesivamente...

x Límites aparentes	x Límites exactos	x Punto medio	n_i	p_i	n_a	p_a
1 – 5	0,5 – 5,5	3	13	0,325	13	0,325
6 – 10	5,5 – 10,5	8	12	0,3	25	0,625
11 – 15	10,5 – 15,5	13	8	0,2	33	0,825
16 – 20	15,5 – 20,5	18	4	0,1	37	0,925
21 – 25	20,5 – 25,5	23	2	0,05	39	0,975
26 – 30	25,5 – 30,5	28	1	0,025	40	1
Σ			40	1		

Amplitud: dif. entre el LSE y el LIE

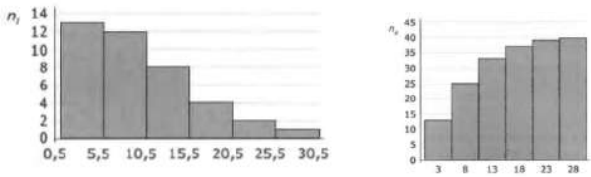
Si se mide el tiempo q se emplea en ejecutar una det. tarea, y los valores resultantes oscilan ente 3,01s y 3,30s, se podría establecer una distribución de frecuencias con 6 intervalos:
En ese caso nuestra unidad de medida es 0,01, ya q contienen 2 decimales. Para calcular los límites exactos hay q sumar y restar la mitad de esta unidad de medida que es 0,005. Así, los límites exactos serían:

x Límites aparentes	x Límites exactos
3,01 – 3,05	3,005 – 3,055
3,06 – 3,10	3,055 – 3,105
3,11 – 3,15	3,105 – 3,155
3,16 – 3,20	3,155 – 3,205
3,21 – 3,25	3,205 – 3,255
3,26 – 3,30	3,255 – 3,305

Intervalo abierto: q no tiene límite inferior o superior. *V ansiedad hubiera 2 sujetos con puntuación de 41 y 43, se puede optar por establecer el intervalo abierto «+ de 30», en lugar de añadir los 3 correspondientes, 2 con frecuencia nula.*

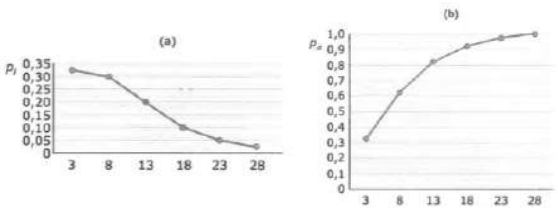
- **Histograma**: extensión del diagrama de barras. Dibuja los rectángulos unidos entre sí, indicando q existe continuidad en los valores. Es un gráfico de V continua dividida en intervalos en los q se eleva un rectángulo con área proporcional a su frecuencia. Puede construirse sobre frecuencias absolutas, relativas o porcentajes, ya sean o no acumulados. En el eje horizontal se sitúan los límites exactos de los intervalos o su punto medio.

(a) histograma acumulado sobre los límites exactos (b) histograma de la V ansiedad ante los exámenes sobre los puntos medios.



- **Diagrama de lineas**: sitúa un punto a una altura proporcional a la frecuencia en cada valor o en el punto medio de cada intervalo (si la V está agrupada en intervalos). Se unen los puntos para formar una línea. Tb polígono de frecuencias.

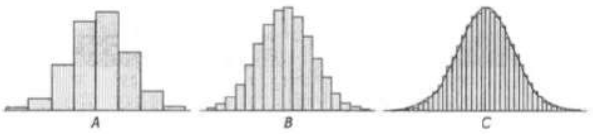
Diagrama de líneas de la V ansiedad ante los exámenes en proporciones (a) Diagrama de líneas acumulativo y en proporciones acumuladas (b)



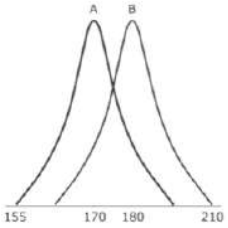
Los gráficos + habituales para representar a una V cuantitativa discreta son el diagrama de barras y de líneas.
Las V cuantitativas continuas agrupadas en intervalos en lugar del diagrama de barras se utiliza el histograma.
Para describir una V cuantitativa se utilizan algunos índices estadísticos, + frecuentes, la media y la desviación típica.

Tendencia central, variabilidad y forma de una V: aproximación gráfica

Cómo describir las V mediante los índices estadísticos adecuados, q se utilizan para medir la tendencia central, variabilidad y forma de la distribución. Pero, antes hay q ver gráficamente q carac. de la V pretenden evaluar utilizando curvas suavizadas, q son histogramas basados en un gran nº de observaciones, cuyos ángulos se han suavizado. Así, si disponemos de los datos de una muestra en una variable X (A) y hacemos esos intervalos más pequeños (B), y más pequeños aún (C), al trazar un diagrama de líneas sobre los puntos medios, la línea resultante será una curva.



Tendencia central

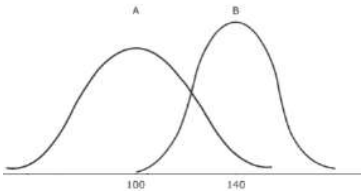


Se refiere al lugar donde se centra una distribución particular en la escala de valores.

Estatura medida en un grupo de 1000 hombres. La tendencia central de los 2 grupos es distinta (las curvas no se solapan completamente) y el B (nacido en 1990) tiene una estatura promedio mayor q el A (pq la curva del grupo B está situada a la derecha, en puntuaciones + altas de estatura).

Esta puede cuantificarse mediante unos índices conocidos como estadísticos de tendencia central.

Variabilidad

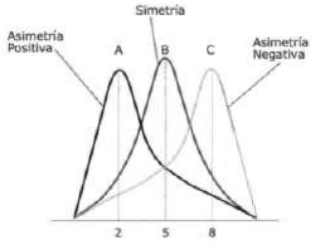


Grado de concentración de los valores entre sí o con respecto a un valor central de la distribución. Una distribución de frecuencias es **homogénea** (tiene poca variabilidad) si los valores están cercanos al promedio y es **heterogénea** (tiene mucha variabilidad) si los valores se dispersan mucho con respecto al promedio.

G A representa las puntuaciones en inteligencia medidas en un G de niños de distintos colegios de la geografía española; el B las de un G con altas capacidades. Además de una tendencia central distinta (GB nivel mayor de inteligencia) podemos apreciar q las puntuaciones en el G con altas capacidades están + próximas entre sí q las del otro grupo. G A presenta una mayor variabilidad en inteligencia.

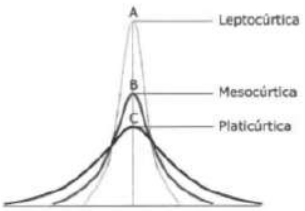
Forma

Para estudiar la forma de una variable se analiza su asimetría y su curtosis.



- **Asimetría**: grado en q los datos se reparten equilibradamente por encima y debajo de la tendencia central.
  - Asimetría positiva: mayor concentración de puntuaciones en la parte baja de la escala.
  - Asimetría negativa: mayor parte de las puntuaciones se sitúan en la parte alta de la escala.
  - Simétrica cuando al dividirla en 2 partes iguales, las 2 mitades se superponen.

Puntuación obtenida por un G de alumnos en un examen muy difícil (A), intermedio (B) y muy fácil (C). El conjunto de puntuaciones presenta una distribución asimétrica positiva si la mayoría de las puntuaciones obtenidas son bajas (A, difícil), simétrica cuando hay un número similar a ambos lados (B) y asimétrica negativa si la mayoría son altas (C, fácil).



- **Curtosis**: grado de apuntamiento de los datos. Si la distribución de frecuencias es muy apuntada se llama **leptocúrtica** (A), y si es muy aplastada se denomina **platicúrtica** (C) y si es intermedio, **mesocúrtica** (B).

Alumno	Xi (nº h)
1	7
2	11
3	16
4	5
5	14
6	10
7	12
8	10

T2: ÍNDICES DE TENDENCIA CENTRAL Y POSICIÓN

Como se ha mencionado en el tema anterior, una de las propiedades + importantes de una distribución de frecuencias es la **tendencia central**; valor o puntuación q represente al conjunto de observaciones. Con el fin de cuantificarla se han desarrollado una serie de medidas o estadísticos q indican sobre q puntuación se concentran las observaciones.

1. ÍNDICES DE TENDENCIA CENTRAL

En el análisis descriptivo de la distribución de frecuencias de una V, es habitual q el nº de observaciones sea grande y q nos planteemos resumir, mediante valores numéricos, sus principales propiedades. Nos interesa calcular un valor central q actúe como resumen numérico para representar al conjunto de datos; **medidas, índices o estadísticos** de tendencia central, q representan toda la distribución de frecuencias con un único valor y facilitan la comparación de dif. conjuntos de puntuaciones de una V. *Medimos el nivel de autoestima en 200 niños (100 ♂ y 100 ♀), además de estudiar la tendencia central en niños y niñas de forma conjunta, estos posibilitan la comparación de niños y niñas en su grado de autoestima. Para saber si es mayor en un sexo q en otro.*

3 medidas de tendencia central representativas de la distribució: media aritmética, mediana y moda.

Media aritmética / promedio

- El + conocido y usado. Por la sencillez de su cálculo y es el fundamento de un gran nº de técnicas estadísticas.
- Indica la tendencia general de una distribución de frecuencias de una V y es el valor central alrededor del cual están la mayoría de las observaciones.
- Desde una perspectiva geométrica: «centro de gravedad» de la distribución de frecuencias.
- **Sólo** puede calcularse para **V cuantitativas** (nivel de medida de intervalo o de razón).
- Suma de todos los valores observados de la V divididos por el nº total de observaciones.
- Utiliza toda la inf. disponible en los datos. Es necesario utilizar todas las puntuaciones de los participantes, no como los estadísticos.

$$\bar{X} = \frac{X_1 + X_2 + \ldots + X_n}{n} = \frac{\sum X_i}{n}$$

X_i = valor de la V en el sujeto
n = nº total de observaciones

Cálculo en una tabla de distribución de frecuencias

Como el nº de observaciones suele ser muy grande, es usual q los datos se presenten aquí agrupados o no en intervalos. A partir de las frecuencias absolutas (n_i) o frecuencias relativas o proporciones (p_i).

Absolutas

$$\bar{X} = \frac{\sum n_i X_i}{\sum n_i} = \frac{\sum n_i X_i}{n}$$

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

Esta es la expresión general de la media a partir de las frecuencias absolutas. La fórmula anterior, definida para pocas observaciones, no es más q un caso particular en el e las frecuencias absolutas de cada valor es igual a uno. En efecto, si n_i= 1 para todos los valores de X:

$$\bar{X} = \frac{\sum n_i X_i}{n} = \frac{\sum 1 \cdot X_i}{n} = \frac{\sum X_i}{n}$$

Como se ve, es la fórmula previa.

Relativas

$$\bar{X} = \sum p_i X_i$$

P_i = frecuencia relativa o proporción de observaciones

Como es de esperar, con ambas fórmulas se obtiene el mismo resultado.

! Si se tienen las frecuencias absolutas acumuladas (n_a) y no las absolutas (ni), se deben obtener para poder calcular la media.

Frecuencias de las calificaciones en el examen de Lengua PAU				
X_i (nota)	n_i	$n_i X_i$	$p_i = n_i / n$	$p_i X_i$
1	1	1	0,025	0,025
2	1	2	0,025	0,05
3	1	3	0,025	0,075
4	8	32	0,2	0,800
5	5	25	0,125	0,625
6	7	42	0,175	1,05
7	8	56	0,2	1,400
8	7	56	0,175	1,400
9	1	9	0,025	0,225
10	1	10	0,025	0,250
Σ	40	236	1	5,9

En una distribución de frecuencias agrupadas en intervalos se calcula igual teniendo en cuenta q los valores de X de la fórmula (Xi) serán los puntos medios de cada intervalo.

G1				
X	X_i	n_i	$p_i = n_i / n$	n_iX_i
1-5	3	5	0,25	15
6-10	8	4	0,2	32
11-15	13	6	0,3	78
16-20	18	3	0,15	54
21-25	23	1	0,05	23
26-30	28	1	0,05	28
Σ		20	1	230

G2				
X	X_i	n_i	$p_i = n_i / n$	n_iX_i
1-5	3	8	0,4	24
6-10	8	8	0,4	64
11-15	13	2	0,1	26
16-20	18	1	0,05	18
21-25	23	1	0,05	23
26-30	28	0	0	0
Σ		20	1	155

Puntuaciones agrupadas en intervalos de la V ansiedad antes de los exámenes. G1, sin tratamiento y G2 con tratamiento.

Con frecuencias absolutas (multiplicar columnas ni y Xi);
X= Σ n_iX_i / n
G1: 230/20 = 11,5
G2: 155/20 = 7,75

Con frecuencias relativas (multiplicar columnas pi y Xi):
X= Σ p_iX_i
G1: 11,5
G2: 7,75

Propiedades matemáticas de la media aritmética

1. En una distribución, la suma de las desviaciones de cada valor con respecto a su media es igual a 0.

$$(X_1 - \overline{X}) + (X_2 - \overline{X}) + \ldots + (X_n - \overline{X}) = \sum_{i=1}^{n}(X_i - \overline{X}) = 0 \qquad \sum_{i=1}^{n}(X_i - \overline{X}) = \sum_{i=1}^{n} X_i - \sum_{i=1}^{n}\overline{X} = n\overline{X} - n\overline{X} = 0$$

Si si se dispone de un nº mayor de observaciones en el q se repiten valores y éstos se presentan mediante una distribución de frecuencias agrupados o no en intervalos, la expresión para comprobar la propiedad es:

$$\sum_{i=1}^{n} n_i \left(X_i - \overline{X}\right) = 0$$

Se resta el X_i a la media de distribución y esa diferencia se multiplica x la frecuencia absoluta de cada intervalo (n_i) obteniendo una nueva columna; $n_i(X_i - X)$. Hay q tener en cuenta la frecuencia absoluta de cada valor (n_i), -nº de veces q aparece cada puntuación o intervalo-.

2. Si a cada puntuación X_i de la variable X le sumamos una constante a (elegida arbitrariamente), la media de las nuevas puntuaciones es igual a la media de X más la constante. Si $Y_i = X_i + a$, entonces $\overline{Y} = \overline{X} + a$. —> Cuando se suma una constante cualquiera (*a*) a las puntuaciones de una distribución (X_i), la media de esa distribución ($\overline{X}$) se ve afectada, y para obtener la nueva media ($\overline{Y}$) tb se debe sumar esa constante a la media original.

3. Si cada puntuación X_i de la variable X se multiplica por una constante *b* (elegida arbitrariamente), la media de las nuevas puntuaciones es igual a la media de $\overline{X}$ multiplicada por la constante. Y; = b • X;, entonces $Y_i = b \cdot X_i$. Cando multiplicamos una constante cualquiera (b) por cada una de las puntuaciones de una distribución (X_i), la media de esa distribución ($\overline{X}$) se ve afectada y para obtener la nueva media ($\overline{Y}$) tb debemos multiplicar esa constante a la media original.

4. <u>La media de *J* muestras o media ponderada</u>:
Hasta ahora se ha hablado de la media de una V en una muestra con n casos u observaciones. Si se cuenta con la media de varios grupos e interesa conocer la media de todas las observaciones juntas, disponemos de las puntuaciones en la variable X en *J* muestras distintas o grupos con n_1, n_2, …, n_J observaciones y con medias $\overline{X}_1$,$\overline{X}_2$, …, $\overline{X}_J$ respectivamente. La media total de los *J* grupos, q incluye las puntuaciones de todas las muestras:

J = nº de grupos o muestras

$$\overline{X}_p = \frac{n_1 \cdot \overline{X}_1 + n_2 \cdot \overline{X}_2 + \ldots + n_J \cdot \overline{X}_J}{n_1 + n_2 + \ldots + n_J} = \frac{n_1 \cdot \overline{X}_1 + n_2 \cdot \overline{X}_2 + \ldots + n_J \cdot \overline{X}_J}{n}$$

Como se puede apreciar, la media de los *J* grupos no es más q una ponderación de las medias de cada grupo en base al nº de observaciones de dicho grupo (n_1). Es decir, la media de cada grupo tiene un peso en la media total q está en función del tamaño de la muestra o nº de casos de cada grupo. Por ello, a la media total (de todas las puntuaciones) se le denomina **media ponderada ($\overline{X}_p$)**.
No podemos calcular la media de una V medida en distintos grupos como la media de las medias de los grupos. Es necesario tener en cuenta el peso de la media de cada grupo a través del nº de casos de cada grupo (n_J . $\overline{X}_J$).
En el caso particular en el q todos los grupos presentan el mismo nº de observaciones (k), es decir, n1 = n2 = ... = n1 = k, la fórmula de la media ponderada se simplifica pq la media total se corresponde con la media de las medias de los grupos.

$$\overline{X} = \frac{\overline{X}_1 + \overline{X}_2 + \ldots + \overline{X}_J}{J}$$

La media ponderada tb se utiliza para obtener la media global en una puntuación q se basa en distintas pruebas con pesos diferentes en su importancia para la puntuación final. *A las prácticas se le ha otorgado un peso del 60% y al examen 40% en la calificación final. Si la media de las prácticas = 6 y en el examen = 5, la nota media final:*

$$\bar{X}_p = \frac{p_1 \cdot \bar{X}_1 + p_2 \cdot \bar{X}_2}{p_1 + p_2} = \frac{60 \times 6 + 40 \times 5}{100} = 5,6$$

P1 y P2: pesos asignados a las prácticas

Limitaciones de la media aritmética

- Cuando los datos están agrupados en intervalos, la media no se puede calcular si el intervalo máximo no tiene límite superior y/o el mínimo no tiene límite inferior.
- Es sensible a la existencia de unas pocas observaciones con valores extremos en la distribución de frecuencias. Como en distribuciones marcadamente asimétricas, por lo q afecta a su representatividad como valor central de la distribución. Estos valores extremos pueden ser producto de errores en la recogida (entonces se eliminan) o grabación de los datos, o valores q aportan inf. relevante de la V (entonces se recomienda aplicar otros índices menos sensibles a los valores extremos como la mediana).

La mediana (Md)

Valor q divide la distribución de frecuencias de la V en 2 partes iguales, cada una el 50% de las observaciones.

- Útil cuando la distribución es asimétrica una buena alternativa a la media aritmética para resumir la tendencia central de las puntuaciones. No se ve afectada por los valores extremos q pueda adoptar la V debido a q en su cálculo no intervienen todos los valores de la distribución, únicamente los q ocupan las posiciones centrales.
- Es un valor + apropiado para representar la tendencia central de la distribución.
- Se puede obtener en todo tipo de V excepto cualitativas.
- Su valor no tiene por qué coincidir con un valor real de la V (especialmente en cuantitativas discretas). Se trata de un valor q cuantifica la tendencia central de la distribución.

1. Se ordenan las *n* puntuaciones de menor a mayor.
2. Se observa si el nº de observaciones *n* es impar o par.
   - Impar; el valor de la mediana es el de la observación q ocupa la posición central dentro de ese conjunto de observaciones ya ordenadas. Coincide con la posición (n+1)/2.
   - Par; es la media aritmética de los 2 valores centrales de la distribución, los q ocupan las posiciones $X_{n/2}$ y $X_{(n/2)+1}$:

V: nº de h de estudio semanales.
La 4ta y la 5ta posición son los valores centrales de la distribución (de 8 = n).

4ta: (n/2 = 8/2 = 4)
5ta: ((n/2)+1) = 5

Md = 10+11/2 = 10,5

$$Md = \frac{X_{n/2} + X_{(n/2)+1}}{2}$$

$X_{n/2}$ = valor de la V en la posición n/2
$X_{(n/2)+1}$ = valor en la posición (n/2)+1

Aquí tb es normal es q el nº de observaciones no sea tan pequeño y aparezcan valores repetidos = datos presentedos en tablas de distribución de frecuencias agrupados o no en intervalos. En este caso, el intervalo en el q se encuentra la mediana se denomina **intervalo crítico** y se corresponde con aquél en el q la frecuencia absoluta acumulada n_a es igual o superior a n/2 o la proporción acumulada (p_a) es igual o mayor a 0,50:

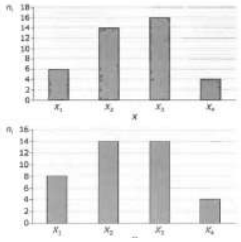
$$Md = L_i + \left(\frac{\frac{n}{2} - n_d}{n_c}\right) \cdot I$$

Ej pág 65

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

- Siempre se empieza desde el valor más bajo de la V hasta el más alto, pq se define como el nº de veces q se repite cada valor o cualquiera de los valores inferiores.
- El intervalo crítico es el 1er intervalo (empezando por el intervalo de valores de X más pequeño) cuya frecuencia acumulada sea igual o mayor a n/2 es decir, al 50% de n. Debe ser igual o superior.
- El origen de la fórmula planteada se basa en el **método de interpolación**, en el q se asume la distribución homogénea de las puntuaciones dentro de cada intervalo.
- Cuando los datos no están agrupados en intervalos, el cálculo a es un caso particular de la fórmula anterior en la q la amplitud (I)= 1 y los límites exactos se obtienen sumando y restando 0,5 unidades a cada valor de la V.
- Se puede calcular en cualquier distribución de frecuencias de V, excepto cuando cualitativas o agrupadas en intervalos en la q existe uno abierto y éste es el intervalo crítico en el q se encuentra la mediana -pq necesitamos conocer su amplitud-.

La moda (Mo)
Valor o categoría de la V con mayor frecuencia absoluta.



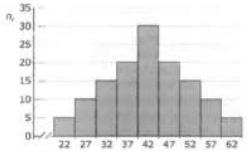
- Tanto en V cualitativas como en cuantitativas.
- Cuando en una V existe un único valor con la frecuencia absoluta máxima, la distribución presenta una única moda y es **unimodal**. Si son 2 los valores con la frecuencia más alta la distribución es **bimodal**, si son 3, **trimodal**...
- Puede ocurrir q una distribución no tenga moda, **distribución amodal**, todos los valores tienen la misma frecuencia absoluta.
- En V cualitativas: es la categoría con la máxima frecuencia (n_1)
  En V cuantitativas con datos no agrupados en intervalos, el valor con mayor frecuencia absoluta.
  En una distribución de una V cuantitativa con datos agrupados en intervalos, se localiza en el intervalo modal (intervalo con la frecuencia máxima) y es su punto medio (si el intervalo es 1-5, la moda = 3).

Principales carac. de la moda:
- Índice de cálculo e interpretación sencillos.
- Único índice q tb puede aplicarse a V cualitativas.
- Cuando los datos están agrupados en intervalos o existen abiertos, se puede calcular a no ser q el intervalo modal coincida con el abierto.

Elección de un índice de tendencia central
Como 1ª opción se recomienda la media aritmética, pq en ella están basadas un gran nº de técnicas estadísticas de gran importancia y uso frecuente (a no ser q tenga distribución asimétrica). Entonces, la siguiente opción es la mediana, q se puede obtener en V con nivel de medida ordinal, y se puede calcular en distribuciones con datos agrupados en intervalos con intervalos abiertos. Sin embargo, en ocasiones no se puede obtener
1. El nivel de medida de la V es nominal.
2. Con datos agrupados en intervalos, se encuentra en el intervalo abierto.

Entonces, la única alternativa es utilizar la moda, q no se puede calcular cuando la distribución sea amodal o el intervalo abierto coincide con el intervalo modal.

Hoy en día, con el uso de programas informáticos para el análisis estadístico de los datos, se recomienda, siempre y cuando sea pertinente, el cálculo de los 3. Es interesante resaltar q cuando la distribución de una V cuantitativa es simétrica y unimodal, coinciden los valores de la media, mediana y moda.



Nivel de medida	Tipo de variable	Índice estadístico de tendencia central aplicable
Nominal	Cualitativa	*Mo*
Ordinal	Cuasicuantitativa	*Mo, Md*
De intervalo De razón	Cuantitativa discreta Cuantitativa continua	*Mo, Md,* $\overline{X}$

2. ÍNDICES DE POSICIÓN
Hasta ahora hemos definido medidas q representaban al conjunto de datos para disponer de un indicador o resumen numérico de la tendencia central de todas las puntuaciones. Ahora la cuestión va dirigida a un sujeto o dato particular: en una distribución de frecuencias de una V, un sujeto «s» obtiene una puntuación X , ¿q posición ocupa este en la distribución con respecto al resto?, ¿q puntuación tendría q obtener para superar a un porcentaje det.?

Los índices estadísticos de posición informan acerca de la posición relativa de un sujeto con respecto a su grupo dentro de la distribución de frecuencias de la V. Debemos dividir la distribución en un nº de partes o secciones iguales entre sí en cuanto al nº de observaciones. Los 3 -percentiles, cuartiles y deciles- se usan con mucha frecuencia en la presentación de resultados estadísticos, especialmente los dos 1ºs. En Psicología, por ej para las normas de interpretación de las puntuaciones de los tests o baremos. Asimismo, su aplicación requiere al menos de un nivel de medida ordinal en la V objeto de estudio.

Percentiles (o centiles)
Los 99 valores q dividen en 100 partes iguales la distribución de frecuencias de la V. El **percentil k**, denotado por P_k, es deja por debajo de sí un porcentaje k de observaciones, donde k=1, 2...99 de la V de interés.
No son porcentajes, sino valores q dejan por debajo de sí un det. tanto por ciento o porcentaje de las observaciones o casos. Aunq el concepto es sencillo y fácil de entender, no hay una única manera de calcularlo. De hecho no es de extrañar q dependiendo del procedimiento aplicado y el software informático obtengamos percentiles ligeramente distintos, aunq todos válidos según la definición general.

El percentil 50, P_{50}, de una distribución deja por debajo de sí al 50% de las observaciones y por encima al otro 50% = la mediana o percentil 50. Por este motivo, el cálculo de los percentiles lo vamos a realizar utilizando una extensión del método de la mediana. La diferencia estriba en q, en la mediana se trataba de localizar la posición de n/2 en la columna

de las frecuencias absolutas acumuladas. En cambio, en los percentiles y de forma más general, se hace en base al nº n·k/100, el nº de casos q se correponden con el k% del percentil y se obtiene a partir de una sencilla regla de 3: si *n* observaciones son el 100% de los participantes, ¿cuántas observaciones serán el k%?

$$\left.\begin{array}{l} n \text{ observaciones} \rightarrow 100\% \text{ de los participantes} \\ x \rightarrow k\% \text{ de los participantes} \end{array}\right\} \Rightarrow x = \frac{n \times k}{100}$$

n = Nº de casos, suma de todos los n_i
K = Percentil a obtener

Este número, n·k/100, es igual a n/2 cuando calculamos el percentil 50.

Cálculo con datos agrupados en intervalos

1. Saber q nº de casos de todos (*n*) deja por debajo de sí el percentil k. Con n·k/100.
2. Localizar el intervalo en el q se encuentra el percentil *k*. Intervalo crítico, aquel donde la frecuencia absoluta acumulada (n_a) es igual o superior a n·k/100, es decir, al k% de *n*.
3. Obtenerlo utilizando la sig. fórmula:

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

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

- Cuando n·k/100 es exactamente igual a la frecuencia acumulada hasta un valor o intervalo el percentil se corresponde directamente con el límite superior exacto del intervalo crítico y no es necesario aplicar la fórmula.
- Cuando se tienen muy pocos datos no es habitual calcular percentiles pq tienen poca utilidad. Pero se les aplicaría tb la formula general asumiendo los intervalos con amplitud igual a 1.

Puede suceder q se tenga un valor o puntuación de la variable, X_i, y nos interese saber q percentil ocupa ese valor en la distribución. Realmente se está pidiendo el valor de k dado el valor de X_i, despejándo la k de la ecuación anterior.

Cálculo de K para X_i

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

Podemos obtener un valor con decimales, y dado q los percentiles son 99 valores enteros, tomamos la cantidad entera + próxima. Si el primer decimal es =/mayor a 5, tomamos el nº entero superior; si es >5, el entero inferior.

Otra situación es q se pida el percentil de una puntuación q es, al mismo tiempo, el límite exacto superior de un intervalo y el inferior del siguiente. Se puede elegir cualquiera de los 2 como crítico y obtendríamos el mismo resultado.

Cuartiles y deciles

Las secciones o partes en las q se divide la distribución de frecuencias son muchas menos q en los percentiles.

Los **cuartiles** son 3 valores de la distribución q dividen en 4 partes de igual frecuencia a la distribución.

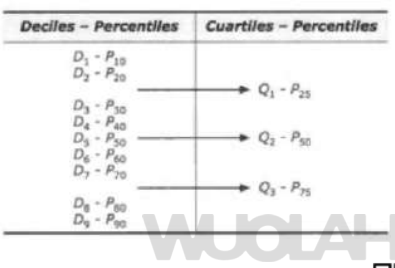
- **1er cuartil,** Q_1, deja por debajo al 25% de las observaciones y r encima al 75%. Se corresponde con el percentil 25 de la distribución; $Q_1 = P_{25}$.
- **2º cuartil**, Q_2, deja por debajo de sí al 50% y por encima al otro 50%. Equivalente al percentil 50 y a la mediana de la distribución, $Q_2 = P_{50} = Md$.
- **3er cuartil,** Q_3, deja por debajo de sí al 75% y por encima al 25%. Se corresponde con el percentil 75 de la distribución, $Q_3 = P_{75}$.

Debido a la equivalencia con los percentiles, para el cálculo se utilizan los métodos para estos. Q_1 se calcula mediante P_{25}, Q_2 con P_{50}, y Q_3 con P_{75}. Se utilizan para construir índices para el estudio de la variabilidad de una distribución de frecuencias.

Los **deciles** son 9 valores q dividen en 10 partes iguales a la distribución. Se representan por D_i, donde i = 1,2,...,9.
El primer decil, D_1 deja por debajo de sí al 10% de las observaciones, el D_2 al 20%,
el D_3 al 30% y así hasta el D_9.

D_1=P10, D_2=P20 ... D_5=P50=Md … D_9=P90.

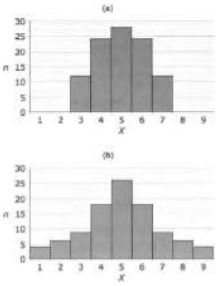
Por lo tanto, tb se pueden calcular a partir de los percentiles correspondientes.



T3: MEDIDAS DE VARIABILIDAD Y FORMA

2 nuevas propiedades de una distribución de puntuaciones: variabilidad o dispersión y forma de la distribución.

1. MEDIDAS DE VARIABILIDAD



El estudio de una distribución resultaría incompleto sin el análisis de la tendencia central; la variabilidad de los datos. Es el grado de variación en un conjunto de puntuaciones; en q se asemejan o diferencian entre sí o se aproximan o alejan de una medida de tendencia central.

Se han propuesto numerosos índices para medirla; la amplitud total, varianza y desviación típica, y amplitud intercuartil. Además, se presenta un índice, el coeficiente de variación, útil para comparar distintas distribuciones de frecuencias en términos de su variabilidad.

En la imagen, el gráfico a) presenta menor variabilidad = muestras más homogéneas. En el caso extremo y poco habitual de máxima homogeneidad, todos los valores de la V serían iguales entre sí y a la media, y no habría variabilidad. Si existe cierta dispersión, la muestra es más o menos heterogénea y las puntuaciones difieren entre sí.

2 tipos de índices:
- Miden el grado en el q las puntuaciones se asemejan o diferencian entre sí.
  - Amplitud total o rango.
  - Amplitud intercuartil.
- Se mide con respecto a alguna medida de tendencia central como la media aritmética.
  - Varianza.
  - Desviación típica.

Tanto unos como otros son útiles para el estudio de la variabilidad pero poco adecuados al comparar la dispersión de 2/+ distribuciones. Para realizar dicho análisis, un índice apropiado es el **coeficiente de variación**, q se basa en la relación entre la desviación típica y la media de cada distribución de frecuencias.

Amplitud total o rango / recorrido de las observaciones

Amplitud total = A_T.

$$A_T = X_{max} - X_{min}$$

Distancia en la escala numérica de un conjunto de puntuaciones entre los valores q representan la puntuación máxima y mínima. En V agrupadas en intervalos, la máx es el límite superior exacto del intervalo máximo y la mínima, el límite inferior exacto del intervalo mínimo.

Su principal inconveniente, es que por su facilidad y utilización de poca inf., es sensible únicamente a los valores extremos de la distribucion y no captura la poca o mucha dispersión entre el resto de valores.

Varianza y desviación típica

La medida de variabilidad tb se puede basar en la distancia entre las puntuaciones y un valor central de la distribución como la media aritmética. Un 1er índice q se puede plantear de forma lógica es el **promedio de las desviaciones o diferencias de cada puntuación** con respecto a su media.

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

El problema de este índice es q el sumatorio del numerador siempre = 0, por lo q carece de sentido como índice. Para utilizar un índice con estas desviaciones, evitando q sea =0, 2 soluciones:

1. Calcular el valor absoluto de cada desviación antes de realizar la suma, obteniendo la **desviación media**:

$$DM = \frac{|X_1 - \bar{X}| + |X_2 - \bar{X}| + \cdots + |X_n - \bar{X}|}{n} = \frac{\sum |X_i - \bar{X}|}{n}$$

   Se emplea muy poco en la actualidad pq es poco manejable matemáticamente por el uso del valor absoluto.

2. Basarnos en el cuadrado de las diferencias y así obtenemos la **varianza**, de un conjunto de *n* puntuaciones en una variable X, denotada por S^2_X se define como el promedio de los cuadrados de las desviaciones de las puntuaciones con respecto a la media:

$$S_X^2 = \frac{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \cdots + (X_n - \bar{X})^2}{n} = \frac{\sum (X_i - \bar{X})^2}{n}$$

   Primero se elevan al cuadrado las diferencias y después se obtiene el promedio de esas desviaciones al cuadrado. Se pueden simplificar los cálculos con la siguiente fórmula:

$$S_X^2 = \frac{\sum X_i^2}{n} - \bar{X}^2$$

Cuando los datos se presentan en tablas de distribución de frecuencias es necesario tener en cuenta la frecuencia de cada intervalo:

$$S_X^2 = \frac{\sum n_i (x_i - \bar{x})^2}{\sum n_i} = \frac{\sum n_i (x_i - \bar{x})^2}{n}$$

$$S_X^2 = \frac{\sum n_i X_i^2}{\sum n_i} - \bar{X}^2 = \frac{\sum n_i X_i^2}{n} - \bar{X}^2$$

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

Varianza a partir de una distribución de frecuencias relativas:

$$S_X^2 = \sum p_i X_i^2 - \bar{X}^2$$

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

La varianza, al basarse en diferencias al cuadrado, es un nº positivo q se expresa en las unidades de la V al cuadrado. Si la variable X se mide en metros, las desviaciones con respecto a la media (xi - ẋ), tb vendrán expresadas en metros, mientras q al elevarlas al cuadrado, son + fácilmente interpretables. Se calcula la raíz cuadrada de la varianza y se obtiene un índice q se denomina **desviación típica**.

La desviación típica de un conjunto de *n* puntuaciones, S_x, es la raíz cuadrada positiva de la varianza:

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

Tanto la varianza como la desviación típica son índices de dispersión muy útiles en el desarrollo posterior de la estadística inferencial, estando en la base de numerosas técnicas estadísticas. A la hora de cuantificar la variabilidad, la desviación típica se suele utilizar + debido a q se expresa en las mismas unidades de medida q la V objeto de estudio. Propiedades de ambos índices:

1. Su cálculo requiere el uso de todas las puntuaciones observadas en la distribución.
2. Miden la variabilidad de los datos con respecto a la media aritmética; únicamente deben aplicarse si es apropiado utilizar la media como medida de tendencia central.
3. Siempre son no negativas, pueden ser = o > q 0. Son iguales a cero únicamente si todas las puntuaciones son iguales entre sí = no variabilidad o dispersión. En el resto de los casos son positivas, siendo sus valores mayores a medida q aumenta la variabilidad de las puntuaciones.
4. Si a las puntuaciones de la variable X les aplicamos una transformación lineal: $Y_i = bX_i + a$ la varianza de las nuevas puntuaciones Y será $S^2_Y = b^2S^2_X$; y la desviación típica será $S_Y = |b|S_X$· Es decir, si a una variable X se le suma o resta una constante *a*, la varianza y desviación típica de la V original no se ven afectadas. En cambio, cuando multiplicamos los valores de X por una constante *b*, la varianza queda multiplicada por la constante al cuadrado y la desviación típica por el valor absoluto de dicha constante.

Por último, otro índice de variabilidad relacionado es la **cuasivarianza**, donde se divide por *n*-1. Y su **cuasidesviación típica** = su raíz cuadrada. Son medidas de dispersión para inferencia estadística. T10.

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

Coeficiente de variación

Es frecuente q uno de los objetivos del análisis descriptivo de los datos sea la comparación del grado de variabilidad o dispersión entre 2 conjuntos de puntuaciones en una misma o distintas V. Como por lo general se miden en unidades distintas no tiene sentido compararlas en base a los valores de sus varianzas o desviaciones típicas. La solución es definir un índice de variabilidad relativa q no dependa de las unidades de medida; el **coeficiente de variación**.

Está definido para variables con X > O y es recomendable q su resultado se acompañe de la media y desviación típica de la distribución a partir de las cuales ha sido calculado. Cuando comparamos 2 conjuntos de puntuaciones de la misma V, tb es necesario para comparar la dispersión de ambas distribuciones. Únicamente es posible utilizar la desviación típica cuando la media de ambos grupos es la misma = mismas conclusiones con ambos índices.

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

Amplitud intercuartil

La varianza y desviación típica + media aritmética, son los estadísticos recomendados para estudiar la variabilidad y la tendencia central de una distribución de frecuencias. En ocasiones, por la asimetría de la distribución, no son aconsejables; necesitamos un índice resistente de dispersión adecuado, q se utilizaría junto con la mediana como medida de tendencia central, sería la **amplitud intercuartil, A_{IQ}**, o rango intercuartil = la diferencia entre el 3er y el 1er cuartil.

$$A_{IQ} = Q_3 - Q_1 = P_{75} - P_{25}$$

Como se puede observar, este índice no informa de la variabilidad del conjunto de puntuaciones, sino del 50% de las mismas comprendidas entre el percentil 25 y el 75 de la distribución.

2. MEDIDAS DE FORMA

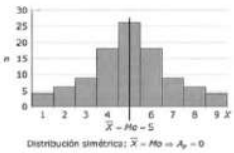
Otro aspecto importante es la forma q presenta la distribución, q se estudia a través de 2 propiedades, la asimetría (visto en el T1) y la curtosis. A continuación se van a describir 2 índices de asimetría y un coeficiente de curtosis, q aportan datos numéricos a ambas propiedades de la forma de la distribución.

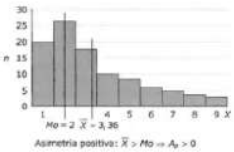
Asimetría de una distribución

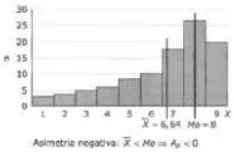
La asimetría indica el grado en el q las puntuaciones se reparten por debajo y por encima de la medida de tendencia central. Existen dif. índices para cuantificarla. Vamos a describir los 2 q se utilizan con + frecuencia:

Índice de asimetría de Pearson: se basa en la relación entre la media y la moda, Es un índice **adimensional** (no tiene unidades de medida) q se aplica a **distribuciones unimodales** (con una única moda).

- Con distribución simétrica, media y moda coinciden = el numerador se anula y el valor de $A_p = 0$.
- Con asimetría positiva, la media es mayor q la moda, $A_p > 0$.
- Con asimetría negativa, el valor de la moda es superior al de la media, $A_p < 0$.







$$A_p = \frac{\bar{X} - Mo}{S_X}$$

Relación entre la representación gráfica de la asimetría de una distribución y el índice de asimetría de Pearson:

Índice de asimetría de Fisher: Se basa en las distancias de las puntuaciones respecto a su media elevadas al cubo, por lo q su valor puede ser positivo, negativo o 0.

$$A_F = \frac{\frac{\sum (X_i - \bar{X})^3}{n}}{S_X^3} = \frac{\sum (X_i - \bar{X})^3}{n \cdot S_X^3}$$

Si los datos se presentan en tablas de distribución de frecuencias —>

$$A_F = \frac{\sum n_i (X_i - \bar{X})^3}{n \cdot S_X^3}$$

Este tiene en cuenta todas y cada una de las puntuaciones de la muestra por lo q puede considerarse el mejor. Al igual q el de Pearson su valor es 0 si la distribución es simétrica ($A_F = 0$); menor q 0 si la distribución es asimétrica negativa ($A_F < 0$); y mayor q 0 si es asimétrica positiva ($A_F > 0$) .

Apuntamiento o curtosis de una distribución

La curtosis se refiere al grado de apuntamiento de los datos en la distribución de frecuencias. Tomando como referencia la curva normal, puede adoptar 3 formas:

A. **Leptocúrtica**: distribución muy apuntada, mayor q en la ‘’normal”. Positiva ($C_r > 0$).
B. **Platicúrtica**: muy aplastada, menor q lo ‘’normal”. índice negativo ($Cr < 0$).
C. **Mesocúrtica**: grado de apuntamiento intermedio.”normal”. índice 0 ($C_r = 0$).

Se basa en las distancias de cada puntuación respecto a la media elevadas a la 4ª potencia:

$$C_r = \frac{\sum (X_i - \bar{X})^4}{n \cdot S_X^4} - 3$$

Si los datos se presentan en tablas de distribución de frecuencias —>

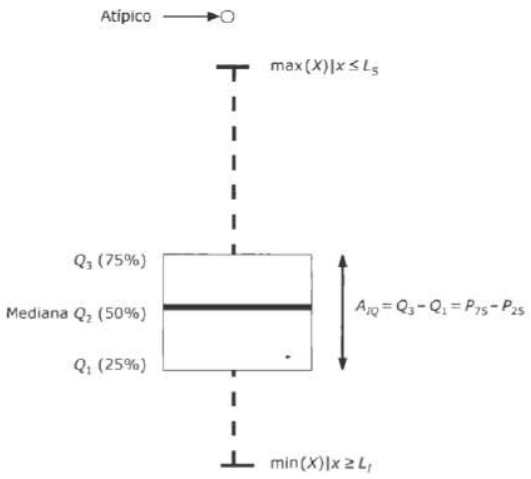
$$C_r = \frac{\sum n_i (X_i - \bar{X})^4}{n \cdot S_X^4} - 3$$

3. DIAGRAMA DE CAJA

O gráfico de caja y bigotes *(boxplots o box and whiskers)* fue propuesto por Tukey (1977). Es una presentación visual útil para estudiar la asimetría de una V cuantitativa, y detectar si hay valores extremos o atípicos (outliers) en la distribución de frecuencias (sin agrupar en intervalos).

Se representa mediante una caja rectangular cuya altura = la amplitud o rango intercuartil $A_{IQ} = Q_3 - Q_1 = P_{75} - P_{25}$. Dentro se dibuja una línea para indicar dónde se sitúa la mediana -q coincide con el 2º cuartil-. Es atravesada por una línea vertical llamada **bigote**, en cuyos extremos se sitúan los valores mínimos y máximos de la V (sin considerar los valores atípicos en caso de q existan). Los límites q determinan si un valor es atípico se calculan multiplicando la amplitud intercuartil por 1,5 y restando este resultado al Q_1 (cálculo del límite inferior) o sumándolo al Q_3 (cálculo del límite superior).

$$L_S = Q_3 + A_{IQ} \times 1,5 \qquad L_I = Q_1 - A_{IQ} \times 1,5$$



Cuando existen casos extremos o atípicos, aparecen como un círculo pequeño por encima o debajo de los bigotes del diagrama de caja. Para estudiar la asimetría se va a tener en cuenta la longitud de los bigotes y el nº de casos atípicos en ambas colas:

- Bigotes de la misma longitud y mismo nº de casos atípicos en ambos lados: aproximadamente simétrica.
- Bigotes de igual longitud pero + casos atípicos en un extremo = distribución con asimetría.
- Bigotes de dif. longitud = distribución asimétrica, como es el caso q se ha representado. La longitud del bigote superior es mayor que la del bigote inferior.

4. PUNTUACIONES TÍPICAS

Hasta ahora hemos tratado fundamentalmente con **puntuaciones directas** *(de un sujeto en un test)*. Estos son los 1ºs datos de los q habitualmente disponemos, pero la comparación de las puntuaciones directas de un mismo sujeto en 2 V distintas puede llevarnos a confusión pq nos ofrecen muy poca inf. De hecho, conocida una puntuación directa no sabemos si se trata de un valor alto o bajo pq esto depende del promedio del grupo. Una solución es trabajar con puntuaciones diferenciales.

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

Si a una puntuación directa x_i le restamos la media de su grupo obtenemos una **puntuación diferencial** o de diferencia, q representamos por x_i (minúscula):
Estas aportan más inf.; si la puntuación coincide con la media de su grupo, es inferior o es superior a ella.
Estas puntuaciones presentan las siguientes propiedades:

A. su media es cero: $\bar{x} = 0$

$$\overline{x} = \frac{\Sigma x_i}{n} = \frac{\Sigma\left(X_i - \overline{X}\right)}{n} = \frac{\Sigma X_i - \Sigma\overline{X}}{n} = \frac{\Sigma X_i}{n} - \frac{n\overline{X}}{n} = \overline{X} - \overline{X} = 0$$

B. La varianza de las puntuaciones diferenciales es igual a la varianza de las puntuaciones directas:

$$S_x^2 = \frac{\Sigma\left(x_i - \overline{x}\right)^2}{n} = \frac{\Sigma x_i^2}{n} = \frac{\Sigma\left(X_i - \overline{X}\right)^2}{n} = S_X^2$$

$$z_x = \frac{x_i}{S_x} = \frac{X_i - \overline{X}}{S_x}$$

Por tanto, al restar a las puntuaciones directas su media hemos obtenido una nueva escala con media 0 y con idéntica varianza a las puntuaciones directas. Sin embargo, 2 puntuaciones diferenciales idénticas pueden tener un significado muy dif. en función de la media y de la varianza de las distribuciones de las q proceden; entonces se utilizan las **puntuaciones típicas**, q van + allá y nos permiten no sólo comparar las puntuaciones de un sujeto en 2 V distintas, tb 2 sujetos distintos en 2 pruebas o V distintas. Reflejan las relaciones entre las puntuaciones con indep. de la unidad de medida = permiten hacer comparaciones entre distintos grupos e incluso entre distintas V.
Al proceso de obtener puntuaciones típicas se llama **tipificación**. Indican el nº de desviaciones típicas q se aparta de la media una det. puntuación. Tienen las siguientes propiedades:

A. Su media es 0

$$\bar{z}_x = \frac{\sum z_x}{n} = \frac{\Sigma\left(\frac{x_i}{S_x}\right)}{n} = \frac{\frac{1}{S_x}\Sigma x_i}{n} = \frac{\sum x_i}{nS_x} = \frac{0}{nS_x} = 0$$

B. Su varianza es igual a 1

$$S_{z_x}^2 = \frac{\Sigma\left(z_x - \bar{z}_x\right)^2}{n} = \frac{\Sigma\left(z_x\right)^2}{n} = \frac{\Sigma\left(\frac{x_i}{S_x}\right)^2}{n} = \frac{\frac{1}{S_x^2}\Sigma x_i^2}{n} = \frac{1}{S_x^2}\frac{\sum x_i^2}{n} = \frac{1}{S_x^2}S_x^2 = 1$$

T4: RELACIÓN ENTRE VARIABLES I

Estudiar conjuntamente +1 V nos va a permitir responder a preguntas. Al igual q en la descripción de 1 única V, el procedimiento a utilizar es dif. en función del tipo de V q se trate:

- **Ambas cualitativas**
  ¿Hay relación entre el sexo de los pacientes y el trastorno psicológico que padecen?
  Si existe o no cierta relación o asociación q hace que el valor q adopte una se asocie en alguna medida con det. valores de la otra. El estadístico q se utiliza para comprobar si existe esa relación o son Indep. es x^2, junto a algunos coeficientes q permiten valorar la fuerza de la asociación entre dichas V.
- **Ambas ordinales**
  Su tratamiento dependerá de la cantidad de valores q adopten ambas variables: si es un nº muy reducido, entonces se utilizarán los mismos procedimientos q para V cualitativas y si es amplio se utilizarán índices estadísticos adaptados como el coeficiente de correlación por rangos de Spearman.
- **Ambas cuantitativas**
  ¿Hay relación entre la motivación y el rendimiento académico de los estudiantes en una det. asignatura?
  Lo q se realmente se cuestiona es si al aumentar la motivación se incrementa tb su rendimiento y a la inversa. El coeficiente de correlación momento-producto de Pearson es el índice + utilizado.
  En el caso de encontrar relación se puede ir un paso más allá y predecir una V en función de la otra. Así, se podría pronosticar el rendimiento en función de la motivación, o mejor aún, en función de su motivación y Cociente Intelectual (CI) mediante el análisis de la regresión.
- **Una cualitativa y otra cuantitativa**
  ¿Hay relación entre la puntuación obtenida en una pregunta del examen y la obtenida en todo el examen?
  En q medida una predice a la otra. Se puede utilizar el coeficiente de correlación biserial puntual, muy directamente relacionado con el coeficiente q cuantifica la relación entre 2 V cuantitativas (el coeficiente de correlación momento-producto de Pearson) .

Para estudiar 2 V cualitativas, se utilizarán 3 estrategias relacionadas:

- **Tabla de contingencia**: forma resumida de representar los datos de las 2 V. informa sobre las frecuencias conjuntas (de ambas V) y marginales (de cada una de las 2 por separado), permite valorar la relación de ambas mediante el estudio de las distribuciones condicionadas de 1, agrupadas en función de los valores de la otra.
- **Diagramas de barras conjuntos**: representación gráfica apropiada. diagrama de barras adosadas y de barras apiladas.
- **Estadístico x2**: para comprobar la indep. entre 2 V cualitativas. Se proponen varias medidas globales para valorar la fuerza de la asociación: los coeficientes *C* de Contingencia, *V* de Cramer, y ϕ; índices, basados en el estadístico x^2, q tratan de superar algunas de sus limitaciones.

En V ordinales: el coeficiente de correlación por rangos de Spearman.

1. ASOCIACIÓN ENTRE 2 VARIABLES CUALITATIVAS

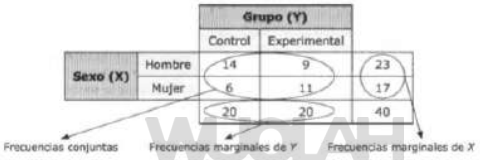
- A lo largo de ellas sólo es posible establecer categorías no ordenadas; q pueden ser intercambiadas.
- Pueden ser dicótomicas o politómicas.
- Tb se considerarán aquellas q, en un principio, tienen un mayor nivel de medida (ordinal, intervalos o razón) pero a posteriori, han sido categorizadas.
- Se dice q hay asociación si existe algún tipo de tendencia o patrón de emparejamiento entre sus distintos valores.

Tabla de contingencia

Forma de ordenar los datos para estudiar la relación entre V con pocas categorías. Es una distribución de frecuencias clasificada de acuerdo a los valores q pueden tomar las 2 V. Por eso, se sitúan los valores de una en las filas y los de la otra en las columnas.

Frecuencias conjuntas (n_{ij}): nº de individuos q toman el valor X_i en la variable X, e Y_j en la variable Y. La suma de todas representa el total de la muestra (*n*). Toma en consideración 1 de los valores de las 2 V.
Frecuencias marginales: Totales de cada valor de una única V. Hay de la variable X y de la variable Y. La suma de estas de cada V tiene q ser igual al total de la muestra.
Es muy frecuente tener +2 categorías en alguna de las V. El formato general es el mismo, añadiendo filas o columnas, y calculándose las distintas frecuencias de la forma indicada.

Distribución marginal: distribución de frecuencias unidimensional (marginal) q nos informan del nº de observaciones para cada valor de una de las V prescindiendo de la inf. sobre los valores de las demás. Hay una de la variable X (q contiene todas las frecuencias marginales de X) y una de la variable Y (q contiene todas las frecuencias marginales de Y).
Distribución condicionada: especifica las observaciones q hay de cada valor de una de las V al imponer la condición de q la otra tome un valor det. Hay una **de Y condicionada a un valor de X_i** q considera únicamente una fila y una **de X condicionada a un valor de Y_i,** q únicamente tiene en cuenta una columna.



		Variable Y				Total
		y_1	y_2	...	y_j	
Variable X	x_1	n_{11}	n_{12}	...	n_{1j}	n_{1+}
	x_2	n_{21}	n_{22}	...	n_{2j}	n_{2+}
	⋮	⋮	⋮	...	⋮	⋮
	x_i	n_{i1}	n_{i2}	...	n_{ij}	n_{i+}
Total		n_{+1}	n_{+2}	...	n_{+j}	n

Siendo $x_1,...,x_i$, los distintos valores de la variable X;
$y_1,...,y_j$, los distintos valores de la variable Y

			Grupo (Y)		Total
			Control	Experimental	
Sexo (X)	Hombre	n_{ij}	14	9	23
		P_{ij} del total	35%	22,5%	57,5%
		P_i por Sexo	60,87%	39,13%	100%
		P_j por Grupo	70%	45%	57,5%
	Mujer	n_{ij}	6	11	17
		P_{ij} del total	15%	27,5%	42,5%
		P_i por Sexo	35,3%	64,7%	100%
		P_j por Grupo	30%	55%	42,5%
Total		n_{ij}	20	20	40
		P_{ij} del total	50%	50%	100%
		P_i por Sexo	50%	50%	100%
		P_j por Grupo	100%	100%	100%

Aquí aparecen varios tipos de frecuencias absolutas:

- La de cada casilla o celda: surge de la distribución conjunta por combinación de 2 valores: nº de casos q comparten 2 características a la vez (n_{ij})·
- La total de cada valor o categoría de la V. El conjunto de estos valores = **distribución marginal absoluta**: nº de casos q tienen 1 carac., de la variable X o fila (n_{j+}) o de la variable Y o columna (n_{+j})·
- El total de casos analizados es *n* o n_{++}, de una muestra o del total.

Todas las de esta, son frecuencias absolutas (n_{ij}), pero de manera habitual, se presentan además con inf. de los porcentajes. Hay 3 tipos de porcentajes conjuntos q se pueden utilizar en ellas:

- **Porcentaje del total (P_{ij})**: nº de casos de cada celdilla dividido por el total de casos (*n*) y multiplicado por 100.
- **Porcentaje condicionado a X_i o porcentaje por fila**: nº de casos de cada celdilla dividido por el total de casos por fila y multiplicado por 100. El conjunto de estos valores = **distribución condicional de filas**.
  Hay q establecer la condición previa de pertenecer a uno de los 2 valores de X_i, *En la Tabla hay q fijarse en la 1ª fila, 23 hombres. Calculando solamente dentro de esta, la distribución de porcentajes para los 2 grupos son 60,87% hombres asignados al grupo control y 39,13% al experimental. Estos 2 porcentajes suman 100, ya q para su cálculo estamos considerando únicamente hombres.*
- **Porcentaje condicionado a Y_j** o por columna: nº de casos de cada celdilla dividido por el total de casos por columna y multiplicado por 100. El conjunto = **distribución condicional de columnas**.
  Hay q establecer la condición previa de pertenecer a uno de los 2 valores de Y_j. *En la 1ª columna hay 20 personas asignadas al grupo control. Calculando los porcentajes dentro de esta, el 70% son hombres y el 30% mujeres; Q suman 100.*

Al considerar los porcentajes para interpretar la relación hay q tener en cuenta si es simétrica o asimétrica.

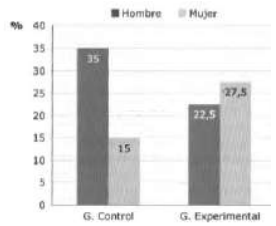
- **Relación asimétrica**: una de las 2 se considera como factor explicativo de la distribución de la otra. Los porcentajes se calculan en el sentido de la V explicativa, por lo q la suma de los porcentajes en cada categoría referidos al total marginal de esa categoría será el 100%. Dicho de otra forma, si la se sitúa en las columnas de la tabla de contingencia, para hacer las comparaciones se calcularán los porcentajes por columna.
- **Simétrica**, no existe esa distinción. Se puede utilizar cualquiera de los porcentajes.

Cada uno de los 3 tipos de porcentajes pone el énfasis en una distribución diferente y ofrece comparaciones distintas, según el sentido de la predicción. Su utilización permite eliminar la influencia del tamaño de la muestra y de los marginales, por lo q se pueden realizar comparaciones entre valores de las distribuciones condicionadas, q indican la existencia de relación o no entre las V, así como la naturaleza de la relación.

Representación gráfica: diagrama de barras conjunto

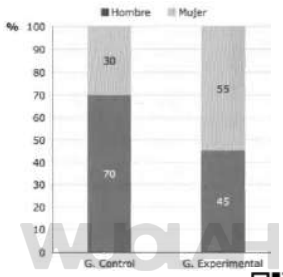
Apropiado cuando al menos una de las 2 V es cualitativa. Se construye sobre los datos de la tabla de contingencia, situando una en el eje horizontal, y para identificar la otra se utilizan barras de distinto color o trama. Básicamente hay 2 formas de representarlo; barras adosadas y barras apiladas.

Diagrama de barras adosadas



Barras colocadas horizontalmente o verticalmente la frecuencia de cada casilla del interior de la tabla de contingencia. Para cada valor de la variable X se representa, una al lado de otra, la frecuencia con q se presenta cada valor de Y dentro de ese valor de X. Al estar situadas unas junto a otras permite la comparación rápida entre y dentro de cada V. Es habitual mostrarlo en porcentajes del total o condicionales (por fila o por columna).

Diagrama de barras apiladas



Muestra una barra por cada valor q toma la variable Y, las cuales se dividen en distintos colores q representa a cada valor de la variable X. Indica la frecuencia con la q aparece cada valor de X en cada valor de Y, comparando entre categorías, la aportación de cada valor al total. Es la + adecuada para visualizar porcentajes condicionados.

Medidas globales de asociación entre variables cualitativas

Independencia: x^2

Asociación = la distribución de las frecuencias de los valores de 1 de las 2 V difiere en función de los valores de la otra.
Independencia = no existe tal patrón de relación entre sus valores.

Para saber si existe o no independencia se utiliza el estadístico x^2, q se basa en la comparación de las frecuencias conjuntas. Así, se comparan las frecuencias empíricas (u observadas) con las frecuencias teóricas (o esperadas) suponiendo q no hubiera asociación. Al compararlas, si no existen diferencias se concluye la ausencia de asociación o de relación de interdependencia = son independientes entre sí.

$$\chi^2 = \sum\sum \frac{(n_e - n_t)^2}{n_t}$$

n_e = frecuencia empírica (u observada)
n_t = frecuencia teórica (o esperada)

- **Frecuencia teórica**: aparece en cada casilla en caso de independencia; se calcula multiplicando las 2 frecuencias marginales y dividiendo por la frecuencia total *n*.
- **Frecuencia empírica**: la q se corresponde con los datos de la casilla.

$$n_t = \frac{Total\ fila \times Total\ columna}{n}$$

Una vez conocidas todas las frecuencias se puede calcular x_2. El sumatorio engloba toda la fracción, por lo q se van a sumar 4 fracciones, una por cada celdilla. En cada una, la frecuencia empírica - la teórica; el resultado se eleva al cuadrado y se divide entre la frecuencia teórica. El índice x^2 toma el valor 0 cuando 2 V son independientes, siendo mayor q 0 cuando exista asociación entre ellas, tanto mayor cuanto más intensa. Ahora bien, no tiene un límite máximo, una dificultad a nivel interpretativo. A veces no podemos saber (sin aplicar técnicas de estadística inferencial) si el valor está lo bastante próximo a 0 como para considerar q la relación es mínima o considerable.

Otro inconveniente es q al multiplicar las frecuencias de todas las casillas por una constante, aumenta, a pesar de q las proporciones de todas las casillas sean las mismas antes y después de dicha multiplicación. Esto hace q su valor solo pueda compararse para V en tablas de contingencia del mismo tamaño (I x J) y con el mismo n.

Características del estadístico x^2 :

- **Adopta valores entre 0 y +∞**. Dado q está definido por valores elevado al cuadrado y las frecuencias nunca son negativas, no puede tomar valores negativos.
- Únicamente adopta el valor 0 si la frecuencia empírica de la celdilla es igual a la teórica q le corresponde, en todas las celdillas de la tabla de contingencia.
- El tamaño de la muestra, n, debe ser relativamente grande. El criterio q se utiliza habitualmente es q la frecuencia esperada mínima por casilla sea al menos de 5 en aprox. el 80% de las casillas, considerando además q la frecuencia mínima esperada en cada casilla sea 1.
- Sirve para valorar la existencia o no de indep., pero no resulta apropiado para medir la intensidad de la relación, pues el tamaño de la muestra y el nº de categorías de las V influyen sobre sus valores.

La asociación entre variables no debe entenderse como una cuestión de todo o nada, sino como un continuo, que iría desde la ausencia de relación (independencia) al nivel máximo de relación entre las variables, que sería una relación determinista. Dado que x2 no resulta apropiado para evaluar el grado de relación entre variables, se han desarrollado va- rios índices que tratan de superar sus limitaciones. Aquí se verán algunos de ellos, que están basados en x2 y no tienen en cuenta si la relación es simétrica o no.
Estos coeficientes son índices globales del grado de intensidad de la relación, que si bien, tienen la ventaja de simplificar la información que proporcionan al resumir la tabla de contingencia en un único valor numé- rico, tienen la desventaja de no permitir ver el detalle de la relación entre las categorías de las variables (lo que sí se puede apreciar con el estudio de las distribuciones condicionadas ya vistas anteriormente).

Coeficiente C de Contingencia

Medida de asociación derivada de x^2 aplicable a tablas de contingencia de cualquier dimensionalidad (con indep. del nº de filas y columnas).

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

Puede asumir valores mayores o iguales a 0 y menores q 1.
Cuanto mayor es el valor de C, mayor es la relación; valores cercanos a 0 indican ausencia de relación.
C adopta el valor 0 cuando $x^2 = 0$ (si todas las frecuencias teóricas coinciden con las empíricas). Para adoptar el valor 1 el nº de observaciones (n) tendría q ser = 0, motivo por el q nunca llega a ese valor.

Especialmente útil cuando el nº de filas y de columnas coinciden pq se puede precisar más su valor máximo, lo q permite una interpretación mejor con la siguiente fórmula:

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

k = nº de filas = nº de columnas

Coeficiente V de Cramer

Modificación de x^2 q alcanza un valor máximo de 1 en caso de máxima asociación o asociación perfecta y un valor mínimo de 0 en una situación de independencia perfecta.

$$V = \sqrt{\frac{\chi^2}{n(m-1)}}$$

Es poco frecuente encontrar valores próximos a 1, de hecho pocas veces se alcanza 0,6. En términos empíricos se puede considerar al 0,6 prácticamente como un valor máximo habitual, por lo q 0,3, antes q considerarlo como bajo, es + bien un valor intermedio.

m = valor más pequeño entre el nº de filas y el nº de columnas

Coeficiente ϕ

Medida de asociación derivada de x^2 q se aplica a V dicotómicas (con tablas de contingencia 2x2). Al utilizarse en V q solo pueden adoptar 2 valores, la tabla de contingencia general queda reducida.

		Y		
		0	1	
X	0	n_{11}	n_{12}	n_{1+}
	1	n_{21}	n_{22}	n_{2+}
		n_{+1}	n_{+2}	n

ϕ puede adoptar valores entre -1 y 1:
- Positivo: el producto de n_{11} x n_{22} es mayor q el de n_{12} x n_{21}.
- Negativo: caso contrario.

$$\varphi = \frac{n_{11} \times n_{22} - n_{12} \times n_{21}}{\sqrt{n_{1+} \times n_{2+} \times n_{+1} \times n_{+2}}}$$

Así, para 2 V dicotómicas codificadas con 0 y 1, un valor positivo de ϕ indicará q los sujetos tienden a estar clasificados en 1 o en 0 en las dos variables; un coeficiente negativo quiere decir q la tendencia es a estar clasificado en 1 en una variable y en 0 en la otra. Esta fórmula es equivalente al coeficiente de correlación de Pearson aplicado a V dicotómicas. Se obtiene el mismo resultado (en valor absoluto) con la aplicación de V de Cramer.

Existe una variación de esta fórmula q puede aplicarse a V politómicas pero no tiene valor máximo, lo q dificulta su interpretación, motivo por el cual se desaconseja.

2. RELACIÓN ENTRE VARIABLES ORDINALES

Con estas V se pueden establecer relaciones de tipo mayor, menor, o igual. Sin embargo, no se pueden evaluar las distancias entre los distintos valores de la V *-en la V nivel socioeconómico, evaluada teniendo en cuenta 3 niveles (bajo, medio y alto), se puede afirmar q una persona con un nivel bajo tiene un nivel menor q una persona con un nivel medio, pero no se puede evaluar cuánto menor es-*.

Habitualmente, cuando se estudia la relación entre una V cualitativa y una ordinal se utilizan las mismas estrategias q en el estudio de 2 cualitativas. En el caso del estudio de 2 ordinales, la estrategia dependerá del nº de valores distintos q puedan adoptar:

- Si ambas adoptan un **nº reducido** de valores, se suelen utilizar tablas de contingencia, de manera similar a lo visto en V cualitativas. Cuando interesa estudiar la fuerza de la asociación, teniendo en cuenta el carácter ordinal, en lugar de los índices globales vistos, se utilizan otros desarrollados específicamente como la *d* de Sommers, o el coeficiente Gamma.
- Si alguna de las 2 V (o ambas) adoptan un **nº amplio** de valores, el estudio en tablas de contingencia deja de ser práctico por el elevado nº de filas y columnas. Se suele utilizar el coeficiente de correlación de Spearman o el coeficiente *tau-b* de Kendall.

Coeficiente de correlación por rangos de Spearman

Se basa en los rangos de los datos en lugar de hacerlo en los valores reales. Resulta apropiado en el caso de V ordinales o cuantitativas q no tengan una distribución normal.

Para calcularlo, 1º hay q ordenar todos los casos para cada una de las V y asignar un rango consecutivo a cada observación de cada por separado. Frecuentemente se producen empates o puntuaciones iguales = rangos empatados. En estos casos se asigna a las puntuaciones el rango promedio q ocuparían las observaciones empatadas.

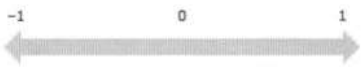
Si la asociación fuera perfecta, esperaríamos q el rango q corresponde a cada caso de la variable X fuera exactamente igual al de la variable Y, por lo tanto el coeficiente se calcula en base a las diferencias registradas en los rangos entre ambas V, esperando q estas diferencias fueran 0. Conforme mayores son las dif. en las ordenaciones de ambas, + se aleja la relación de ser perfecta. Para evitar q las diferencias positivas anulen las negativas, el estadístico se calcula en función de la suma de las diferencias elevadas al cuadrado.

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

Los valores oscilan de -1 a +1. El signo del coeficiente indica la dirección de la relación y el valor absoluto del coeficiente de correlación indica la fuerza de la relación.

- $r_s > 0$, existe relación directa. Cuánto + cercano a 1, más fuerte será la relación.
- $r_s < O$, relación inversa entre las V. Cuánto + se acerque a -1, más fuerte la relación.
- $r_s \approx 0$, apenas hay relación.

d_i = Rango(X_i) - Rango(Y_i)
n = n° de sujetos



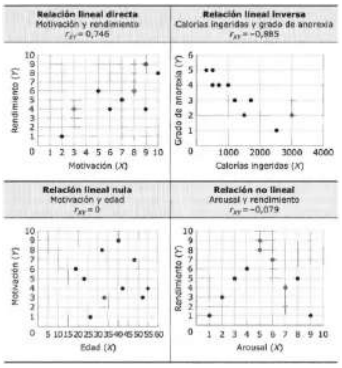
T5: RELACIÓN ENTRE VARIABLES II

1. RELACIÓN ENTRE VARIABLES CUANTITATIVAS

Las q están en un nivel de medida de intervalo o de razón. Poseen una unidad de medición común y constante. Su relación se estudiará mediante métodos gráficos y estadísticos.

Representación gráfica de la relación: el diagrama de dispersión

- O nube de puntos.
- Se utiliza con 2 V cuantitativas, ofreciendo una 1ª aproximación de la relación entre ambas.
- Para realizarlo se sitúa una de las V en el eje de abscisas y la otra en el de ordenadas. Para cada par de datos, se localiza la intersección de ambas V y se marca con un punto.
- Atendiendo, se puede observar q existe cierta relación lineal entre las V, correspondiendo, en mayor medida, calificaciones altas a mayor n° de h de estudio y viceversa. Hay, sin embargo, algunas excepciones *el estudiante 8, ha estudiado un nº de horas más bien alto (10) y ha obtenido un 1.*
- **Relación lineal directa**: los valores altos en Y tienden a emparejarse con valores altos en X, los valores intermedios en Y tienden a emparejarse con valores intermedios en X, y los valores bajos en Y tienden a emparejarse con valores bajos en X.
  Relación lineal inversa: los valores altos en Y tienden a emparejarse con valores bajos en X, los valores intermedios en Y tienden a emparejarse con valores intermedios en X, y los valores bajos en Y tienden a emparejarse con valores altos en X.
  Relación lineal nula: no hay un emparejamiento sistemático entre ellas en función de sus valores.
- Según la **Ley de Yerkes-Dodson**, la relación activación-rendimiento toma la forma de una U invertida. Para cada tipo de tarea se define un grado óptimo de activación en el cual el rendimiento para esa tarea es máximo. Por encima y por debajo de ese nivel óptimo, el rendimiento decrecerá tanto más cuanto más lejos se encuentre el nivel actual de activación del óptimo.



índices estadísticos para cuantificar la relación:

Covarianza

- Detecta la relación lineal entre X e Y; la variación conjunta de 2 V.
- Valor positivo = relación es directa.
  Valor negativo = relación inversa.
  Valor en torno a 0 = relación nula.
- Su valor absoluto será mayor cuanto + acusada sea la tendencia a la linealidad en el diagrama de dispersión. Se designa por S_{XY}, o Cov*(X, Y).*
- Su limitación, al igual q el coeficiente x^2, se desconocen los valores mín. y máx. q puede adoptar, lo q merma su capacidad para interpretar el grado de relación.

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

X_i = valor de la variable X en el caso i.
Y_i = valor de la variable Y en el caso i.
$\bar{X}$ =media de la variable X.
$\bar{Y}$ = media de la variable Y.

Coeficiente de correlación lineal de Pearson

- Índice q detecta la relación lineal entre X e Y, superando los límites de interpretación de la covarianza, al tener establecido un valor máximo (1) y mínimo (-1).
- Solo es apropiado para el estudio de las relaciones lineales.

$$r_{XY} = \frac{Cov(X,Y)}{S_X S_Y} = \frac{S_{XY}}{S_X S_Y}$$

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

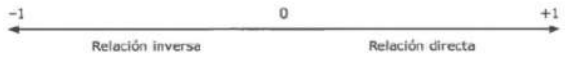
S_x = desviación típica de la variable X.
S_y = desviación típica de la variable Y.
S_{xy} = covarianza entre X e Y.

Atendiendo a la **1ª fórmula**;
El coeficiente es el cociente entre la covarianza entre X e Y y el producto de la desviación típica de X y la desviación típica de Y. Por tanto, será preferible, si ya se tienen alguno de los cálculos previos (la covarianza o las desviaciones típicas de X e Y). Cuando no haya ninguno, la fórmula 2 será + rápida.
En caso de utilizarla, hay q empezar calculando las medias y desviaciones típicas de X e Y, así como la covarianza. Una vez calculados ya se puede sustituir directamente.
Al utilizar la **fórmula 2ª**, se puede sustituir directamente, nada más añadir las columnas relativas a *XY*, X^2 e Y^2.

Obviamente el resultado es idéntico utilizando las 2.
Para interpretar los resultados hay q tener en cuenta :
- El valor absoluto. Cuanto mayor es, la relación tb es + fuerte.
- El signo del coeficiente de correlación de Pearson.
  + = valores mayores de la variable X tienden a corresponder a valores mayores de la variable Y, y a valores menores, a valores menores de la variable Y. Relación lineal directa.

- = valores mayores de la variable X tienden a corresponder a valores menores de la variable Y, y a valores menores a valores mayores de la variable Y. Relación lineal inversa.

- El valor oscila siempre entre -1 y +1:
  $r_{xy} > 0$, relación lineal directa. Cuanto más se acerque a 1, más fuerte será. A mayores valores de la variable X nos encontraremos con valores altos de la variable Y y, a la inversa (bajos = bajos).
- $r_{xy} < 0$, relación lineal inversa. Cuanto más se acerque a -1, más fuerte será la relación, de forma q los valores altos de X se corresponderán con valores bajos de Y y viceversa.
- $r_{xy} \approx 0$, apenas hay relación lineal; una poco o nada tiene q ver con la otra.



Carac. del coeficiente de correlación lineal de Pearson:

- Índice simétrico, es igual la correlación de X con Y que la de Y con X ($r_{xy} = r_{yx}$).
- Los valores - 1 y 1 indican una correlación lineal perfecta y el valor 0, ausencia de correlación lineal.
- El valor absoluto no se ve afectado por transformaciones lineales de las V. Asimismo, $r_{xy} = \pm 1$, si una v es una transformación lineal de la otra.
- La correlación (por alta que sea) no implica q X sea la causa de Y, ni q Y sea la causa de X. Para poder hablar de causalidad se tienen que cumplir unos requisitos relativos al diseño de la investigación (q debe ser experimental).
- Puede verse afectado por 3ªs V *Si se mide la estatura y el razonamiento abstracto de los niños de Primaria, habrá una alta correlación; los + altos tendrán + edad. Si se limita a los niños de la misma edad, posiblemente desaparezca la correlación.*

Casos particulares del coeficiente de correlación lineal de Pearson

Hay varias fórmulas q se derivan y utilizan en algunos casos particulares, como el estudio de la relación entre 2 V ordinales, entre 2 V dicotómicas y entre 1 dicotómica y otra cuantitativa. Estas se desarrollaron pq su cálculo es más rápido, aunq con los programas informáticos existentes, no es un problema.

Relación entre variables ordinales

En el tema anterior, el coeficiente de correlación lineal de Spearman 2 V ordinales. Esta fórmula se deriva matemática-mente del coeficiente de correlación lineal de Pearson aplicado a rangos, por lo q su resultado es idéntico.
El único caso en el q no coinciden es en el de empates en los rangos, en cuyo caso hay q utilizar el coeficiente de correlación lineal de Pearson entre los rangos de las V. En caso de no haber empates se puede utilizar cualquiera teniendo en cuenta q el de Spearman simplifica bastante los cálculos.

Relación entre variables dicotómicas

La fórmula del coeficiente ϕ se deriva del coeficiente de correlación lineal de Pearson, por lo q el resultado de ambas es igual. Eso sí, el cálculo de ϕ se basa en la tabla de contingencia, por lo q es bastante + rápido q el de r_{xy} q precisa de las puntuaciones de cada sujeto en ambas V.

Relación entre una variable dicotómica y otra cuantitativa

En 1 V dicotómica los 2 valores se suelen representan por 0 y 1. El **coeficiente de correlación biserial puntual** se utiliza cuando una es dicotómica y la otra cuantitativa. Es muy utilizado en Psicometría . Se denota como r_{bp}.

$$r_{bp} = \frac{\overline{X}_p - \overline{X}_q}{S_x} \cdot \sqrt{p \cdot q}$$

$\overline{X}_p$: la media de las puntuaciones de la V cuantitativa X obtenidas por el grupo al q se le ha asignado un 1 en la V dicotómica.
$\overline{X}_q$: media de las puntuaciones de la V cuantitativa X obtenidas por el grupo al q se le ha asignado un 0 en la V dicotómica.
S_X: desviación típica de la V cuantitativa X.
***p*: proporción de casos del grupo al q se le ha asignado un 1 en la V dicotómica.**
***q*: proporción de casos del grupo al q se le ha asignado un 0 en la V dicotómica.**

Si la V en lugar de ser dicotómica es dicotomizada, el coeficiente a utilizar es el **coeficiente de correlación biserial**. 1 V se ha dicotomizado si hay un continuo latente entre las 2 categorías, pero se han establecido 2 únicos valores.

		Variable Y			
		Cualitativa dicotómica	**Cualitativa politómica**	**Ordinal**	**Cuantitativa**
Variable X	**Cualitativa dicotómica**	Estadístico x^2 Coef. C de Contingencia Coef. V de Cramer Coeficiente ϕ	Estadístico x^2 Coef. C de Contingencia Coef. V de Cramer	Estadístico x^2 Coef. C de Contingencia Coef. V de Cramer	Correlación biserial puntual
	Cualitativa politómica		Estadístico x^2 Coef. C de Contingencia Coef. V de Cramer	Estadístico x^2 Coef. C de Contingencia Coef. V de Cramer	
	Ordinal			Coef. de correlación de Spearman	Coef. de correlación de Spearman
	Cuantitativa				Covarianza Coef. de correlación de Pearson

2. REGRESIÓN LINEAL SIMPLE

Regresión - psicólogo inglés Sir Francis Galton (1822-1911): analizando la estatura de una muestra numerosa de padres e hijos, advirtió q los hijos de padres altos eran tb, en general, superiores en estatura al promedio, pero no tan altos como sus progenitores. Algo parecido ocurría con los hijos de padres bajos q, aún siendo + bajos q el promedio, no eran por término general, tan bajos como ellos. Dedujo q se producía en los hijos una **regresión a la media de la distribución**, q los valores tendían a regresar a la media de la V. Buscó una ecuación matemática para estimar los valores q adoptarían en una V sujetos para los q se conoce sus puntuaciones en otra V y la relación entre ambas. Esta ecuación resultaría ser la ecuación de una recta.

El modelo de regresión utiliza la inf. contenida en las relaciones lineales observadas entre las V. Si 2 variables X e Y se relacionan linealmente, la representación gráfica de su distribución conjunta se aproximará visualmente bastante a una línea recta = poder escribir una V en función de la otra con la ecuación de una recta: Y= *a* + bX.

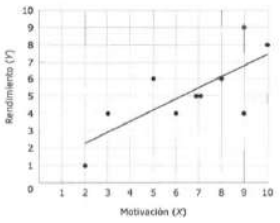
Correlación y regresión son 2 conceptos muy cercanos:

- **Regresión**, el interés se centra en predecir los valores de una V (Y) a partir de los valores conocidos en otra (X), más q en la variación conjunta de las 2. Relación asimétrica o direccional, los resultados serán distintos según se trate de la ecuación de regresión de Y sobre X, o de la de X sobre Y (q predice los valores de X a partir de Y).
- **Correlación**: la relación entre es simétrica. La forma de proceder implica 3 fases:
  1. Identificación del modelo de regresión = obtener los coeficientes de regresión q le caracterizan;
  2. Valoración del modelo = estudio de la capacidad predictiva del mismo;
  3. Aplicación del modelo para predecir V.

Cálculo de los coeficientes de regresión

Desde el punto de vista geométrico, la recta de regresión tiene la misma interpretación q cualquier otra recta, pero desde el estadístico, se ajusta a la nube de puntos del diagrama de dispersión con menos error.

- *a*: constante q se denomina **origen**, corresponde al valor q adopta la variable Y cuando la X vale 0. Indica el origen, el punto en el q la recta corta al eje de ordenadas.
- *b*: constante q se denomina **pendiente**, de ella depende la inclinación de la recta. Indica en q medida cambian los valores de Y por cada incremento de una unidad en los valores de X.
- *Y*: V cuyo valor se desea conocer y va a ser pronosticado a partir del valor de la variable X. Se suele denominar **V pronosticada o criterio** y denotarla como Y'.
- *X*: V cuyo valor se conoce y va a ser utilizado para pronosticar el valor del criterio. Se suele denominar **V predictora** o simplemente **predictor**.



Para calcular los valores de a y de b, conociendo los valores de X e Y, se utilizan:

$$b = \frac{n\sum(XY) - \sum X \sum Y}{n\sum X^2 - (\sum X)^2} \qquad b = r_{xy}\frac{S_Y}{S_X} \qquad a = \overline{Y} - b\overline{X}$$

Se han presentado 2 fórmulas para calcular b. La 1ª es más rápida cuando tenemos los datos directos y la 2ª cuando ya tenemos realizados algunos cálculos previos. Conocido el valor de las dos constantes a y b de la ecuación anterior, tenemos ya completamente formulado el modelo de regresión. Habitualmente se suele poner Y' en lugar de Y para denotar q nos estamos refiriendo a los valores pronosticados en el criterio, no a los valores reales obtenidos por los sujetos.

$$\mathbf{Y'_i = a + bX_i}$$

Además, hay q tener en cuenta q, dado q es una predicción hay cierto nivel de error. De no haberlo, todos los puntos del diagrama de dispersión se encontrarían sobre la recta. Para cada uno de los sujetos se comete cierta cantidad de error al asignarle la puntuación pronosticada Y' en lugar de la Y. En este sentido, a la ecuación anterior habría q añadirle un término q reflejase este error:

$$Y_i = a + bX_i + E_i$$

Donde: $E_i = Y_i - Y_i'$

E_i es una medida del error individual cometido para cada una de las observaciones. Al utilizar un modelo de regresión se utiliza el modelo lineal con el q se comete un error lo más pequeño posible para todos los sujetos. Con el **criterio de mínimos cuadrados**, procedimiento q proporciona valores tales q la suma de los errores al cuadrado (SCE) para los *n* participantes sea mínimo. Formalmente, se establece obteniendo los valores *a* y *b* q minimizan la siguiente expresión:

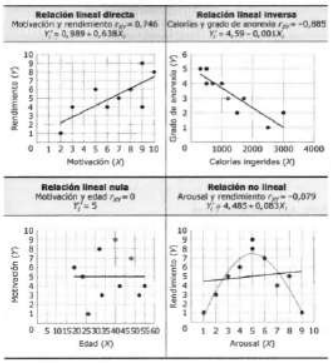
$$SCE = \sum (Y_i - Y_i')^2 = \sum [Y_i - (a + bX_i)]^2$$

Se puede demostrar (derivando parcialmente la función a minimizar respecto de cada uno de los parámetros, igualando a 0 y despejando) q este criterio proporciona las ecuaciones para estimar a y b con el menor error posible, consiguiendo la recta q mejor ajusta a la nube de puntos.

Valoración del modelo

¿Hasta q punto es un buen modelo para predecir la V criterio?

- Aunq se consiga el mejor ajuste posible a los datos disponibles, no es garantía de q ese sea óptimo para predecir la V criterio. Como es obvio, cuanta mayor relación, mejor pronóstico.
- Antes de realizar cálculos numéricos, lo ideal para valorar el ajuste es representar los datos mediante un diagrama de dispersión, para tener una 1ª aproximación de la posible relación.
- La línea continua representa la ecuación de regresión lineal q mejor ajusta a la nube de puntos. En algunos casos, este modelo lineal para predecir Y puede ser una mala opción.
- Para examinar la utilidad predictiva, además de la aproximación gráfica, se pueden utilizar 2 índices: varianza error y coeficiente de determinación.



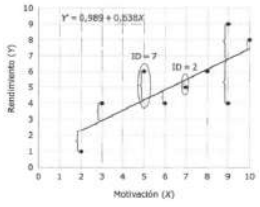
La varianza error

La varianza de los errores (o residuos) cometidos al pronosticar la variable Y a partir de X; la diferencia entre la puntuación q realmente obtendría y la q se le ha pronosticado .

$$E_i = Y_i + Y'_i$$

Se calcula aplicando la fórmula de la varianza a estas puntuaciones error, **error cuadrático medio**. Se puede denotar como S^2_E o S^2_{yx} y se interpreta como la varianza de los errores cometidos al pronosticar la variable Y a partir de X (o mediante la recta de regresión de Y sobre X).

$$S_E^2 = S_{Y.X}^2 = \frac{\sum E_i^2}{n} - \bar{E}^2$$



Se ha marcado con una llave la distancia entre cada uno de los valores asumidos por la variable Y' representada en la recta de regresión (pronosticada) e Y (observada). Estas «distancias» son los errores cometidos al pronosticar el rendimiento en la asignatura (Y) a partir de los valores en motivación (X).
Se aprecian los errores cometidos. El estudiante n° 7 obtuvo en rendimiento académico (Y) 6, mientras la línea de la recta de regresión le pronostica una puntuación menor. El 2° estudiante el error es menor, hay menor distancia entre la puntuación obtenida en Y (5) y la q le pronosticaríamos, cercana a 5,5.

Cuanto menor valor de la varianza error, + similares serán las puntuaciones pronosticadas por el modelo y las q realmente obtendrían los sujetos; mejores predicciones por el modelo de regresión.

El coeficiente de determinación

Igual al coeficiente de correlación de Pearson elevado al cuadrado. Indica la proporción de varianza de la V pronosticada o criterio (Y) explicada por el modelo lineal; por la V predictora X.

r^2_{XY} → Coeficiente de determinación

No depende de las unidades en q se expresan los datos y toma valores entre 0 y 1. ↑ valor = + similares las puntuaciones pronosticadas y las q realmente obtendrían.

- 0 = la V predictora tiene nula capacidad predictiva.
- 1 = la V predictora explicaría toda la variación de Y, y las predicciones no tendrían error.

Características del modelo de regresión

- La pendiente siempre será del mismo signo q el coeficiente de correlación lineal de Pearson, informará sobre el tipo de relación lineal entre las V (directa o inversa). Dado q las desviaciones típicas siempre son positivas, *b* adopta el signo del coeficiente de correlación lineal de Pearson.

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

- La media de los errores de predicción o residuos (E= Y- Y') es O. $\bar{E} = 0$
- La media de las puntuaciones pronosticadas coincide con la media de las verdaderas puntuaciones en Y. $\bar{Y}' = \bar{Y}$
- La varianza de las puntuaciones en Y = suma de la varianza de los pronósticos (hechos mediante la recta de regresión) + la varianza de los errores (o error cuadrático medio).

$$S_Y^2 = S_{Y'}^2 + S_{Y \cdot X}^2$$

- El coeficiente de determinación = cociente entre la varianza de las puntuaciones pronosticadas y la varianza de las puntuaciones en Y. De ahí, q sea un indicador de la proporción de varianza del criterio q queda explicada con el modelo de regresión lineal.

$$r_{XY}^2 = \frac{S_{Y'}^2}{S_Y^2}$$

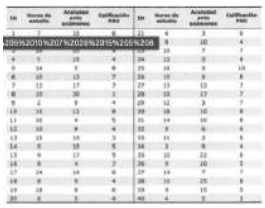
- El complementario del coeficiente de determinación = cociente entre la varianza de los errores y la varianza de las puntuaciones en Y, e indica la proporción de la varianza del criterio q NO queda explicada por el modelo de regresión lineal.

$$1 - r_{XY}^2 = \frac{S_{Y \cdot X}^2}{S_Y^2}$$

3. REGRESIÓN LINEAL MÚLTIPLE

Si se utiliza más de una V predictora, la capacidad predictiva del modelo puede mejorar. Suelen ser + realistas, es raro encontrar criterios q se puedan predecir a partir de una única V predictora.

Dado q se suele realizar mediante software estadístico, no se expondrán las fórmulas. Simplemente se tratará de explicar con un ej., el cambio q se produce en la valoración del modelo a partir del coeficiente de determinación, cuando se introduce una 2ª V predictora.



En la tabla se muestran 3 V, cuantitativas, y tiene sentido pensar q tanto nº de horas de estudio semanales como el nivel de ansiedad presentado en un examen pueden influir (o predecir) la calificación obtenida.
Utilizando la regresión lineal simple, hay q considerar 2 rectas de regresión, 1 para la V predictora n° de horas de estudio y otra para el nivel de ansiedad:
1- la ecuación de la recta de regresión de la Calificación obtenida sobre el n° de h de estudio: Y'_i = 2,82 + 0,292X_i 0,584, con lo q el coeficiente de determinación = ese valor al cuadrado, 0,341. El 34,1% de la variabilidad se explica por el tiempo empleado para estudiar.
2- la ecuación de la recta de regresión de la calificación obtenida sobre el nivel de ansiedad es: Y'_i =8,36- O,239X Y_i = -0,793, con lo q el coeficiente de determinación será 0,629. 62,9% de la variabilidad de la calificación se explica por el nivel de ansiedad.

Al utilizar las 2 V predictoras (X_1 al n° de ho- ras estudiadas y X2 al nivel de ansiedad ante los exámenes), la ecuación de regresión sería: Yi = 5,714 + O, 226X1 - O, 214X2
En este caso, al haber 2 V predictoras implicadas, el coeficiente de determinación varía. Se denota como $R^2_{Y \cdot X1 \cdot X2}$:

$$R^2_{Y \cdot X_1 X_2} = \frac{r^2_{YX_1} + r^2_{YX_2} - 2 \cdot r_{YX_1} \cdot r_{YX_2} \cdot r_{X_1X_2}}{1 - r^2_{X_1X_2}}$$

Para facilitar la utilización se presentarán los datos en una **matriz de correlaciones**; una tabla con el mismo nº de filas y columnas q de V, en la q en cada casilla aparece las correlaciónes correspondientes a la fila y a la columna.

El valor predictivo es muy superior al de ambos modelos de regresión simples. Teniendo en cuenta ambas predictoras simultáneamente se explica el 82,7% de la variabilidad de la calificación, a partir del tiempo de estudio y del nivel de ansiedad ante los exámenes.

	Horas de estudio (X_1)	Ansiedad ante los exámenes (X_2)	Calificación PAU (Y)
Horas de estudio (X_1)	1	-0.185	0,584
Ansiedad ante los exámenes (X_2)		1	-0,793
Calificación PAU (Y)			1

T6: NOCIONES BÁSICAS DE PROBABILIDAD

En las CCSS en general, y en la Psicología y Ciencias de la Salud en particular, es habitual la imposibilidad de prever el resultado de un fenómeno. Según los casos, esto puede ser debido a diversas causas.

- Q una persona tenga fiebre o le duela la cabeza (variaciones de la homeostasis interna) el día en el q participa en un exp. sobre estrés. Su situación y R serán diferentes a las q daría.
- Q estemos empleando un instrumento de medida afectado por condiciones medioambientales, dando lugar a resultados de medida dif.

Estamos expuestos continuamente a sucesos sobre los q no tenemos la certeza de que vayan a ocurrir; pueden tener una mayor o menor probabilidad. Por tanto, puede haber un rango amplio de probabilidades. El resultado está influenciado por el azar o, de forma más correcta, estamos ante un fenómeno aleatorio. Ante estas variaciones = gran incertidumbre en los resultados, ¿cómo se puede actuar desde la perspectiva metodológica? Existen herramientas metodológicas q pueden trabajar con este tipo de datos: «La Estadística permite esbozar conclusiones válidas en situaciones de incertidumbre y variabilidad» y de la **Probabilidad**, ya q ésta es la teoría matemática q desarrolla modelos matemáticos adaptados al estudio de estas situaciones, mediante la asignación de probabilidades.

Estadística Inferencial: conjunto de métodos y técnicas q permiten inducir, a partir de la inf. empírica proporcionada por una muestra, el comportamiento de una det. población, con un riesgo de error medible en términos de probabilidad. Probabilidad y Estadística son ramas de las matemáticas q se complementan. El objeto de la Probabilidad es el estudio de **V aleatorias**, valores q dependen básicamente del azar o de la posibilidad de q puedan o no ocurrir. La Estadística, por su parte, es otra rama de las matemáticas cuyo objeto de estudio son los datos, entendidos como valores o atributos de los objetos de estudio de interés. Son disciplinas íntimamente relacionadas, ambas se refieren al estudio de un mismo tipo de situaciones, en las q hay incertidumbre. La Probabilidad aporta los modelos matemáticos (las distribuciones) para el estudio de la incertidumbre, y la Estadística los adapta a los datos reales (con incertidumbre).

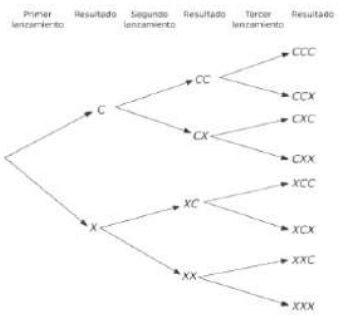
1. CONCEPTOS PREVIOS

Experimento aleatorio: proceso q se puede repetir indefinidamente en las mismas condiciones y cuyo resultado no se puede predecir con certeza. Conduce a la obtención de un resultado donde interviene el azar. 3 carac.:

- Todos los resultados posibles son conocidos con anterioridad a su realización. Tiene un conjunto definido.
- No se puede predecir con certeza el resultado concreto del exp.
- El exp. puede repetirse teóricamente un nº infinito de veces en idénticas condiciones.

Espacio muestral / espacio de resultados: conjunto de todos los resultados posibles de un exp. aleatorio. Habitualmente se representa por la letra mayúscula E.

- *Lanzar una moneda; C = cara y X = cruz. Espacio muestral E= {C; X}*
- *Lanzar un dado; E={1·2·3·4·5·6} • IIIII*
- *Lanzar 2 veces un dado: E= {(1,1); (1,2); (1,3); (1 ,4); (1,5); (1,6); (2,1); (2,2); (2,3); (2,4); (2,5); (2,6); (3,1); (3,2); (3,3); (3,4); (3,5); (3,6); (4,1); (4,2); (4,3); (4,4); (4,5); (4,6); (5,1); (5,2); (5,3); (5,4); (5,5); (5,6); (6,1); (6,2); (6,3); (6,4); (6,5); (6,6)}*
- *Introducir 3 ratas en un laberinto en forma de T (I = va hacia la izquierda; D = hacia la derecha): E = {(I,I,I); (I,I,D); (I,D,I); (D,I,I); (D,D,I); (D,I,D); (I,D,D); (D,D,D)}*

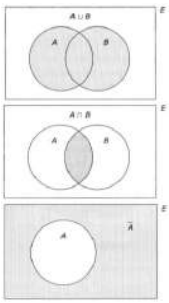


Una forma sistemática y didáctica de construirlos es mediante el **diagrama de árbol**, representación gráfica q muestra los resultados posibles de un exp. aleatorio.
En el caso de q tuviera solo 1 elemento no podríamos hablar de experimento aleatorio pq se puede predecir con certeza el resultado obtenido.

Sucesos: los resultados de un exp. aleatorio, o subconjuntos del espacio muestral. Se representan por letras mayúsculas: *A, B....* A su vez, pueden ser *elementales* o com*puestos.*

- **Suceso elemental**, suceso simple o punto muestra: cada uno de los resultados posibles del exp. aleatorio. 1 solo elemento del espacio muestral E.
- **Suceso compuesto:** 2 o + sucesos elementales.
- **Suceso seguro**: ocurre siempre. Consta de todos los sucesos elementales del espacio muestral y se identifica con el espacio muestral total E.
- **Suceso posible**: puede contener algún elemento del espacio muestral E = tiene posibilidad de q ocurra.
- **Suceso imposible**: no ocurre nunca. No contiene ningún elemento del espacio muestral y se identifica con 0 o conjunto vacío.

Operaciones con sucesos:
Entre los sucesos se establecen las mismas operaciones q en la **teoría de conjuntos**, ya que las operaciones son, en sí mismas, entre conjuntos teniendo en cuenta su definición.
Los **diagramas de Venn** se emplean para representarlos y estudiar visualmente propiedades y operaciones entre ellos. El espacio muestral o de resultados E se representa mediante un rectángulo, y dentro de él se incluyen los sucesos mediante círculos.



• **Unión**: Subconjunto de E formado por los sucesos elementales que pertenecen a A, a B, o a ambos. Representado por A *u* B.
• **Intersección**: Representado por por A *n* B, al subconjunto de E formado solamente por los sucesos elementales q pertenecen a ambos simultáneamente. Cuando la intersección de 2 sucesos no contiene ningún elemento son **sucesos incompatibles o excluyentes**, q no pueden verificarse a la vez.
• **Complementario**: Representado por A, al subconjunto de E formado por los sucesos elementales q no pertenecen a A.

Las operaciones de unión e intersección pueden extenderse al caso de 2 o + sucesos *A u B u C*, y puede hablarse de complementario de la unión o de la intersección de 2 sucesos.

2. DEFINICIÓN DE PROBABILIDAD

En teoría de la probabilidad se toman todos los posibles resultados de un exp. aleatorio como elementos del espacio muestral *E* (espacio de resultados). Si E contiene un número finito de elementos, a cada uno se le puede asociar un nº no negativo, q es su probabilidad de ocurrencia, tal q la suma de todos los nº correspondientes a todos los elementos de *E* sea 1.

La **probabilidad de un suceso** es una medida numérica e cuantifica la posibilidad de q este ocurra. Los valores se encuentran comprendidos entre 0 y 1.
- Muy probables estarán próximos a 1 y menos, al 0.
- El valor 0 se asigna a los sucesos imposibles y 1 para los seguros.

En función del enfoque desde el q se estudie se ha propuesto un tipo dif. de definición. Cada una de ellas tiene una operativa distinta, pero todas un mismo objetivo: calcular la posibilidad de ocurrencia.

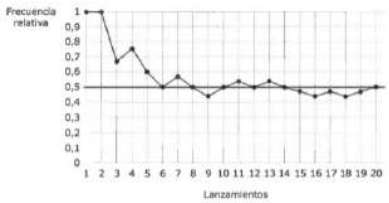
1. La **definición clásica** o a **priori**; ***Regla de Laplace***; la probabilidad de un suceso *A* es igual al cociente entre el nº de casos favorables de q ocurra y el nº de casos posibles en el supuesto de q todos los casos tengan la misma oportunidad de ocurrir (sean equiprobables).

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

P(A) = Probabilidad de un suceso *A*
n_A = Nº de casos favorables
***n* = Nº de casos posibles**

A priori pq antes de realizar el exp. aleatorio se conocen de antemano los posibles resultados del espacio muestral E y sus probabilidades. Esta plantea algunos problemas a la h de aplicación; parte de la base de q los sucesos son **equiprobables** (tienen la misma probabilidad de ocurrencia).

Nº de lanzamientos	Resultado	Nº de caras	Frecuencia relativa
1	C	1	1/1 = 1
2	C	2	2/2 = 1
3	X	2	2/3 = 0,6666
4	C	3	3/4 = 0,7500
5	X	3	3/5 = 0,6000
6	X	3	3/6 = 0,5000
7	C	4	4/7 = 0,5714
8	X	4	4/8 = 0,5000
9	X	4	4/9 = 0,4444
10	C	5	5/10 = 0,5000
11	C	6	6/11 = 0,5454
12	X	6	6/12 = 0,5000
13	C	7	7/13 = 0,5384
14	X	7	7/14 = 0,5000
15	X	7	7/15 = 0,4666
16	X	7	7/16 = 0,4375
17	C	8	8/17 = 0,4706
18	X	8	8/18 = 0,4444
19	C	9	9/19 = 0,4736
20	C	10	10/20 = 0,5000



La probabilidad de salir cara en el lanzamiento de una moneda es de 1 caso favorable dividido por 2 casos posibles; 1/2. Supongamos q se realiza y anota si sale cara o cruz en cada tirada, así como la frecuencia relativa en cada caso. Los resultados podrían ser los q se presentan. Si se representan gráficamente, según ↑ el nº de lanzamientos, la línea quebrada q une las frecuencias se ajusta + a la horizontal trazada en la ordenada 1/2 (0,5) o valor teórico de la probabilidad definida por Laplace. Por tanto, la frecuencia relativa tiende a estabilizarse cuando el nº es muy elevado. **Ley del azar o ley de regularidad estadística**: fenómeno de estabilización de las frecuencias.

No obstante, no siempre es fácil aplicar este concepto, muchas veces no es posible repetir un experimento aleatorio un gran nº de veces, y si lo es, no es práctico.

* _____

Andréi Nicoláyevich Kolmogórov - **teoría axiomática de la probabilidad**: aplicación de la teoría de conjuntos a los sucesos q componen el espacio muestral. 2 ventajas:
- Recoge las definiciones de probabilidad anteriores; cumplen la axiomática.
- Permite el desarrollo matemático de la teoría de la probabilidad.

Definición axiomática de probabilidad: Dado un espacio muestral *E*, se denomina probabilidad de un suceso A_i, definido en el espacio muestral E y designado por $P(A_i)$, a un nº real asignado al suceso A_i, q cumple las sig. propiedades:

- $0 \leq P(A_i) \leq 1$
- $P(E) = 1$
- Si $A_1, A_2 \ldots A_k$ son sucesos incompatibles dos a dos, entonces:

$$P(A_1 \cup A_2 \cup \ldots\ldots \cup A_k) = P(A_1) + P(A_2) + \ldots\ldots + P(A_k)$$

$$P(A) = \frac{n_A}{n} = \frac{0}{6} = 0$$

$$P(A) = \frac{n_A}{n} = \frac{n}{n} = P(E) = 1$$

Las 2 primeras indican q la probabilidad es cuantificable numéricamente con un nº entre 0 y 1.
- Se asigna un 0 a un suceso imposible A:
- Se otorga un 1 a un suceso seguro A, q se corresponde con el espacio muestral E.

La 3ª propiedad indica q la unión de sucesos incompatibles (aquellos cuya intersección es el conjunto vacío: $A_i \frown A_j = Ø$) es igual a la suma de sus probabilidades individuales.

De estas propiedades se deriva q P(A) + P(A) = 1 siendo ambos sucesos exhaustivos y excluyentes (si no ocurre A, necesariamente lo hará A). Esto implica q P(A) = 1 - P(A), o lo q es lo mismo, la probabilidad del suceso complementario de A, (A), es igual a 1 menos la probabilidad de ocurrencia de A. A partir de la definición axiomática se deducen una serie de teoremas, de los cuales se van a examinar el de la suma (concepto de unión de sucesos) y el del producto (concepto de intersección).

2. La **definición estadística** o **a posteriori**; no asume la equiprobabilidad. Se basa en la estabilidad de las frecuencias relativas cuando el nº de repeticiones de un suceso aleatorio es muy elevado y tiende a infinito.
   Lanzamos un dado al aire muchas veces, y anotamos las frecuencias relativas de un suceso. Estas tienden a estabilizarse en un valor constante, entre 0 y 1 = probabilidad del suceso.
   P(A) o probabilidad de un suceso *A*: límite al q tiende la frecuencia relativa de aparición de un suceso *A* cuando el nº de ensayos *n* o repeticiones tiende a infinito.

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

   A posteriori pq las probabilidades se atribuyen a cada suceso después de un gran nº de repeticiones del exp. aleatorio.

3. TEOREMA DE LA SUMA

Partiendo del axioma 3, referido a la unión de sucesos, establece q la probabilidad de q ocurra el suceso A o el B es igual a la probabilidad de q ocurran A + la de q ocurra B - la de q ocurran A y B (intersección de ambos sucesos).

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Si son mutuamente excluyentes (simultáneamente) o complementarios (ocurrencia de uno = no del otro), la regla se simplifica, resultando ser la suma de las probabilidades de cada suceso dado q $P(A \frown 8) = 0$

$$P(A \cup B) = P(A) + P(B)$$

4. PROBABILIDAD CONDICIONADA

Hasta ahora hemos estado tratando sucesos indepe., donde la probabilidad de uno no altera la del otro. Sin embargo, no siempre son tan simples; la aparición de u A puede depender de la aparición de B. En estos casos son dependientes, la probabilidad de A depende o está condicionada al suceso B: P (A|B)

Para dos sucesos cualesquiera A y B, la **probabilidad de A condicionada a B** (o de A supuesto B) es igual a la probabilidad de la intersección dividida por la probabilidad de B, siempre q P(B) ≠ 0

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

Siempre q P(A) ≠ 0

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

Si los sucesos son indep.:

$$P(A|B) = P(A) \quad y \quad P(B|A) = P(B)$$

5. TEOREMA DEL PRODUCTO

Situaciones en las q se quiere calcular la probabilidad de q aparezcan 2 sucesos simultáneamente; calcular la probabilidad de intersección entre 2 sucesos.
Se ha visto qe la probabilidad condicionada se define: $P(B|A) = \frac{P(A \cap B)}{P(A)}$

Despejando $P(A \frown B)$: $P(A \cap B) = P(A) \cdot P(B|A)$ q corresponde al teorema del producto.

Teorema del producto: probabilidad de ocurrencia de A y B es igual a la probabilidad de ocurrencia de A por la probabilidad de ocurrencia de B, dado q A ha ocurrido previamente (condición previa).

$$P(A \cap B) = P(A) \cdot P(B|A)$$

P(B|A): probabilidad de q ocurra B dado q ha ocurrido A.

Cuando son independientes P(B|A) = P(B): $P(A \cap B) = P(A) \cdot P(B)$

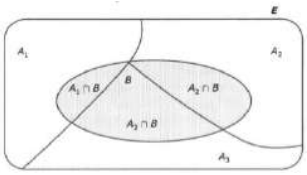
Cuando se extraen bolas o papeletas de una urna. Cuando se realiza +1 extracción, la probabilidad de q ocurra B dado q ha ocurrido A va a verse afectada por el hecho de q el elemento extraído vuelva a reponerse o no. **Extracción con reposición**; se mantiene siempre el mismo nº de bolas o papeletas; se devuelven a la urna. **Extracción sin reposición**: no se devuelven, las probabilidades de obtener una bola o papeleta concreta en esta 2ª extracción van a depender de lo obtenido en la 1ª.

6. TEOREMA DE LA PROBABILIDAD TOTAL

En un espacio muestral E, se dice q k sucesos A_1, A_2… A_k forman una **partición del espacio muestral** si se cumplen simultáneamente las siguientes condiciones:

- $A_i \cap A_j = \emptyset$ para cualquier par de sucesos A_i y A_j; son incompatibles y su intersección es el conjunto vacío.
- $A_1 \cup A_2 \cup \ldots \cup A_k = E$; la unión de todos los sucesos es igual al espacio muestral (son exhaustivos). En términos de probabilidad se cumple q: $P(A_1 \cup A_2 \cup \ldots \cup A_k) = P(A_1) + P(A_2) + \ldots + P(A_K) = 1$

Si sobre ese mismo espacio muestral se define un nuevo suceso *B*, el teorema de la probabilidad total permite calcular su probabilidad a partir de las $P(A_i)$ y de las $P(B|A_i)$.



Partición del espacio muestral en 3 sucesos A_1, A_2 y A_3 incompatibles entre sí, y otro suceso B en el mismo espacio muestral E. Para este caso concreto, la probabilidad de B se corresponde con la suma de las intersecciones de cada uno de los sucesos A_i con el suceso B:

$$P(B) = P(A_1 \cap B) + P(A_2 \cap B) + P(A_3 \cap B)$$

Aplicando el teorema del producto:

$$P(B) = P(A_1) \cdot P(B|A_1) + P(A_2) \cdot P(B|A_2) + P(A_3) \cdot P(B|A_3)$$

De manera general, el teorema de la probabilidad total, Proporciona el valor de la probabilidad de B en función de la probabilidad de los sucesos A; y de las probabilidades de B condicionadas a los A;

$$P(B) = \sum_{i=1}^{k} P(A_i) \cdot P(B|A_i)$$

.

- Otra forma de calcularlas es el **diagrama de árbol**. A la izq, las probabilidades de los sucesos A_1, A_2 y A_3 . Las probabilidades + a la derecha son las de los sucesos B y su complementario B condicionadas a los sucesos A_1, A_2 y A_3, respectivamente. Siempre la suma de las probabilidades del mismo debe ser 1. *$P(A_1)+P(A_2)+P(A_3)=1$; $P(B|A_1)+P(\bar{B}|A_1) = 1$*
  Las probabilidades condicionadas vienen dadas (derecha del diagrama) y las de intersección $P(A_1 \cap B)$, $P(A_2 \cap B)$ y $P(A_3 \cap B)$ se pueden calcular multiplicando las probabilidades de cada rama. Así, la probabilidad del suceso B se determina como:

$$P(B) = P(A_1) \cdot P(B|A_1) + P(A_2) \cdot P(B|A_2) + P(A_3) \cdot P(B|A_3)$$

7. TEOREMA DE BAYES

En sucesos dependientes, la probabilidad condicionada tiene en cuenta inf. de un suceso para conocer la probabilidad de otro. El teorema de Bayes va más allá y permite calcular cómo se modifican las probabilidades de det. sucesos cuando se conoce alguna inf. adicional. El teorema de la probabilidad total permitía obtener la probabilidad de un suceso B y este, las probabilidades condicionadas de los sucesos A_i dado el suceso B.

Teniendo en cuenta la formula de probabilidad condicionada, se aplica el teorema del producto al numerador:

$$P(A_i|B) = \frac{P(A_i) \cdot P(B|A_i)}{P(B)}$$

Y aplicando el teorema de la probabilidad total al denominador se obtiene el teorema de Bayes:

$$P(A_i|B) = \frac{P(A_i) \cdot P(B|A_i)}{\sum_{i=1}^{k} P(A_i) \cdot P(B|A_i)}$$

$P(A_i)$: probabilidades de cada suceso A_i
$P(B|A_i)$: probabilidad del suceso B condicionada a los sucesos A_i

$$\sum_{i=1}^{k} \mathbf{P(A_i) \text{ -} P(B|A_i) = P(B)}$$

El teorema proporciona las probabilidades de A_i condicionadas por B en función de las probabilidades de los sucesos A_i y de las probabilidades de B condicionadas por dichos sucesos.

8. APLICACIONES DE LA PROBABILIDAD CONDICIONADA EN PSICOLOGÍA DE LA SALUD

- Epidemiología --> bastante frecuencia. Una situación habitual podría ser una decisión como diagnosticar a un paciente q presenta unos síntomas (suceso X), y q puede tener una enfermedad (suceso E). *Determinar la probabilidad de tener un infarto (E) si antes se ha tenido un dolor agudo del brazo derecho (síntoma).*

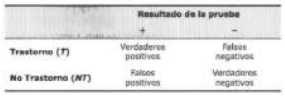
  Prevalencia: proporción de casos existentes de una enfermedad en un momento det. Indica la probabilidad de personas q tienen una carac. o enfermedad en relación a la población.
  Incidencia: proporción de casos nuevos de una enfermedad en una población durante un período det. Probabilidad de personas nuevas q pueden tener una carac. o enfermedad en un periodo concreto.

  Existe una relación entre ambos; si los casos nuevos (incidentes) no se resuelven, se hacen crónicos (prevalentes). Además, una ↓ en la incidencia = ↓ prevalencia y al revés.

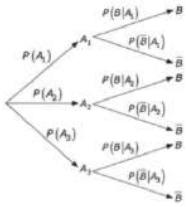
- Análisis de factores de riesgo o probabilidad de q ↑ un problema o enfermedad al estar expuesto a un riesgo—> parte de la base de q sujetos expuestos a un factor (X+) tienen más posibilidades de sufrir una enfermedad o tener un problema psicológico (E+) en comparación con el grupo no expuesto (E-) a dicho factor (X-).

- Valoración de la calidad de las pruebas diagnósticas. Tenemos una prueba para la evaluación diagnóstica de un trastorno (T), q nos va a permitir distinguir a las personas sanas o sin trastornos (NT) de las q lo tienen en función de un punto de corte establecido previamente. Se supone q la prueba dispone de 2 indicadores: (+) indica q tiene el trastorno y (-) q está sana y no tiene el trastorno.
  Los datos se presentan en una tabla de doble entrada:

  Una buena prueba diagnóstica presenta:
  - **Alta sensibilidad** P (+|T) o probabilidad de discriminar a los verdaderos positivos; la probabilidad de q los que tengan un trastorno den positivo (+).
  - **Alta especificidad** P (-|NT) o probabilidad de detectar a los verdaderos negativos; probabilidad de los q no tienen trastorno den negativo (-).



  Sin embargo, las pruebas nunca son exactas y siempre se trabaja con un margen de error en la evaluación. A consecuencia, encontramos 2 tipos de valores predictivos:
  - **Valor predictivo positivo** P(T|+) o probabilidad de q todos los q den positivo tengan el trastorno. Pero hay sujetos q dan (+) y no lo tienen (falsos positivos), por lo q el valor predictivo positivo disminuye.
  - **Valor predictivo negativo** P(NT|-) o probabilidad de q todos los (-) estén sanos y no tengan trastorno. Sin embargo, hay sujetos q dan (-) y tienen el trastorno (falsos negativos) = valor predictivo negativo reducido.



De esta forma, si se determina el grado de sensibilidad, especificidad y valores predictivos de una prueba se puede conocer su calidad. Una prueba sería muy sensible si al aplicarse a un conjunto de personas q tienen el trastorno dan positivo (+) en un porcentaje muy alto; y muy específica si un porcentaje muy elevado de personas NT dan negativo (-). Lo mismo se puede deducir en relación a los valores predictivos positivos y negativos, cuanto más próximos a 100 (o a 1 en términos de probabilidad) más valor predictivo tienen y mejor es su calidad.

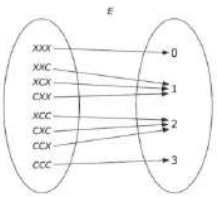
T7: VARIABLES ALEATORIAS Y MODELOS DISCRETOS DE PROBAILIDAD

En los primeros 5 temas se han estudiado las V estadísticas, el conjunto de valores resultantes de medir una carac. de interés sobre cada elemento individual de una población o muestra a través de un procedimiento de asignación numérica mediante la aplicación de det. reglas, dando lugar a los valores de la V estadística. Estos conjuntos de datos los hemos descrito mediante su correspondiente distribución de frecuencias.

Por otro lado, en el tema anterior se estudiaron los fundamentos de la teoría de probabilidades. Q un exp. se denomina aleatorio cuando su resultado no podemos predecirlo con certeza. Si se realiza una sola vez se obtendrá un único resultado del espacio muestral. Pero, a medida q aumenta el nº de ensayos irán apareciendo todos los resultados posibles, cada uno con su correspondiente probabilidad.

Con los conocimientos adquiridos ya es posible pasar del cálculo de probabilidades al estudio de las distribuciones de probabilidad. Éstas son las distribuciones teóricas correspondientes a la probabilidad de ocurrencia de un suceso si repitiésemos el exp. un número infinito (muy grande) de veces y q se refieren a la población.

1. CONCEPTO DE VARIABLE ALEATORIA



Esencial en la teoría de probabilidades. Surge ante la necesidad de cuantificar los resultados de los exp. aleatorios y poder realizar un estudio matemático de los mismos. ''Cualquier carac. medible q toma dif. valores con probabilidades det.
V aleatoria (v.a.), X, una función q asigna un nº real, y sólo uno, a cada resultado posible de un espacio muestral E de un exp. aleatorio. Se representan por letras mayúsculas del alfabeto latino, y se utilizan las letras minúsculas con subíndice para referirnos a los valores concretos q toman. Así X, Y... representan V aleatorias, X_1, X_2, ... y_1, y_2... representan los valores concretos q toman.

En el tema anterior definimos el espacio muestral E (espacio de resultados) como el conjunto de todos los resultados posibles de un exp., entendiendo por este un proceso q se puede repetir indefinidamente en las mismas condiciones y cuyo resultado no se puede predecir. Pues bien, cuando a cada suceso elemental o resultado posible del espacio muestral le asignamos un valor numérico se obtiene una V aleatoria denominada X, q toman valores numéricos, y se pueden definir dif. V sobre los resultados de un mismo exp. *Sobre el de «lanzar una moneda al aire en 3 ocasiones» podemos definir una V aleatoria como nº de caras obtenidas, nº de cruces obtenidas, o una V q toma el valor 1 cuando el nº de caras es mayor q el de cruces y toma el valor 0 en el otro caso. El azar interviene en el resultado q obtenemos y no en la V o función.*

2. TIPOS DE VARIABLES ALEATORIAS

- **Discretas**; adopta valores enteros; fijados 2 valores consecutivos, no puede tomar ninguno intermedio.
  Una V aleatoria X es discreta **(v.a.d.)** cuando sólo puede tomar un conjunto finito de valores o un conjunto infinito y numerable de valores. *Nº de caras q salen al lanzar 2 veces una moneda o el conjunto de los nº enteros, e puede adoptar un conjunto infinito y numerable de valores (los nº negativos, 0 y los nº positivos*).
- **Continuas**; dados dos valores, siempre se puede encontrar un 3er valor incluido entre los 2 primeros.
  Una V aleatoria X es continua **(v.a.c.)**: puede tomar infinitos valores o un conjunto de valores no numerable. *El tiempo de reacción ante un E, la estatura o el cociente intelectual.*

Variable aleatorias discretas

Función de probabilidad de una V.A. discreta

La descripción del comportamiento matemático se realizará similar a la 1ª parte del libro con las V estadísticas. En este caso, su distribución venía dada por los valores q toma la V y su correspondiente frecuencia; vendrá dada por los valores q la V puede tomar (X_1, X_2, …, X_n) y su correspondiente probabilidad .

Se llama función de probabilidad de una V.A. discreta X, y se representa por f(x), a aquella función q asocia a cada valor de la V la probabilidad de q ésta adopte ese valor:

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

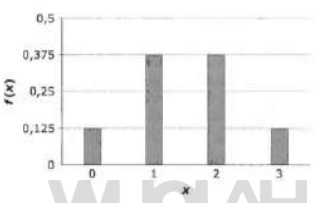
Exp. aleatorio consistente en lanzar una moneda al aire en 3 ocasiones. Si definimos una VA X como nº de caras, obtenemos la siguiente tabla:

Espacio muestral	X	P
⊕ ⊕ ⊕	$x_1 = 0$	1/8 = 0,125
⊕ ⊕ ☺		
⊕ ☺ ⊕	$x_2 = 1$	3/8 = 0,375
☺ ⊕ ⊕		
⊕ ☺ ☺		
☺ ⊕ ☺	$x_3 = 2$	3/8 = 0,375
☺ ☺ ⊕		
☺ ☺ ☺	$x_4 = 3$	1/8 = 0,125

La 1ª columna recoge el espacio muestral del experimento E = {XXX, XXC, XCX, CXX, XCC, CXC, CCX, CCC}, siendo cada fila un suceso. El nº de sucesos o elementos del espacio muestral = 8. En la 2ª columna se muestran los valores q puede tomar la variable X y en la 3ª sus correspondientes probabilidades. Éstas se pueden calcular fácilmente teniendo en cuenta la definición clásica de probabilidad como la de obtener 3 cruces o ninguna cara ($X_1 = 0$) será 1/8, ya q hay un resultado favorable de ocho posibles.

Por tanto, la función de probabilidad de X es:

x	0	1	2	3
f(x)	0,125	0,375	0,375	0,125



La función de probabilidad de una V.A.D puede representarse mediante un diagrama de barras donde en el eje de abscisas se recogen los valores q toma la V y en el de ordenadas sus probabilidades.

Las 2 propiedades fundamentales q debe cumplir la función de probabilidad:

- Para cualquier valor de x, f(x) siempre toma valores positivos o nulos: $\forall x \in X \quad f(x) \geq 0$
- La suma de todas las probabilidades correspondientes a cada valor de X es =1. $\sum f(x) = f(x_1) + f(x_2) + \ldots + f(x_n) = 1$

Puede observarse q no son más q una adaptación de la definición axiomática de la probabilidad, aplicada al V.A..

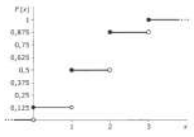
Función de distribución de una V.A. discreta

La **función de distribución** de probabilidad de una V.A. X se representa con la misma letra q su función de probabilidad, pero en mayúscula: F(x). indica la probabilidad de q la V tome un valor menor o igual q un valor concreto x. Asocia a cada valor de la V la probabilidad de q ésta adopte ese valor o cualquier otro inferior.

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

Si ordenamos de menor a mayor los valores x, esta función se obtiene acumulando o sumando los valores de la función de probabilidad:

$$F(x_k) = P(X \leq x_k) = f(x_1) + f(x_2) + \ldots + f(x_k)$$



En la función de probabilidad se le asigna a la probabilidad un valor concreto; esta es acumulativa, se le asigna la probabilidad a un valor concreto y todos los anteriores.
<— representación gráfica.
F (x) va «dando saltos» precisamente en los valores de la V (0, 1, 2 y 3).
El círculo blanco no incluye esos valores.
F(2) = 0,875 pero F(l,9999...) = F(1) = 0,5.

Se pueden deducir, sin demostraciones matemáticas, las propiedades fundamentales q debe cumplir:

- Todos los valores q toma la función son positivos o nulos.

$$\forall x \qquad F(x) \geq 0$$

- F(x) es nula o vale 0, para todo valor inferior al menor valor de la variable aleatoria, x_1:

$$F(x) = 0 \quad \text{si} \quad x < x_1$$

- F(x) es igual a uno para todo valor igual o superior al mayor valor de la V.A. Si llamamos x_n al mayor valor:

$$F(x) = 1 \qquad \text{si} \qquad x \geq x_n$$

- La función F(x) es no decreciente ya q es una acumulación o suma de probabilidades siempre positivas o nulas.
- La probabilidad P de q laV.A. X tome valores x mayores q x_1 y menores o iguales q x_2 ($x_1 < x \leq x_2$) es la diferencia entre los valores de la función de distribución correspondientes a su valor superior menos su valor inferior.

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

Media y varianza de una V.A. discreta

Aprendimos a describir una distribución de frecuencias de una V.Estadística a través de los índices de tendencia central y de dispersión. Lo mismo se puede hacer con una V.A.: calcular su media y su varianza.

Para una V.E.D. X se puede calcular su **media** obteniendo el sumatorio del producto de cada uno de los valores de la V por su frecuencia relativa o proporción. Pues para la media (letra griega «μ») de una V.A.D. X calcularemos el sumatorio de los productos de cada uno de los valores q toma la V por su correspondiente probabilidad:

$$\mu = E(X) = \sum x \cdot f(x)$$

Tb se denomina **esperanza matemática** o **valor esperado de X** y se representa por E(x). Este término tiene sus raíces en los juegos de azar y fue introducido con el fin de estimar las ganancias esperadas, si se repitiese el juego un elevado nº de veces. Referido a una V.A. representa el promedio teórico q tomaría si se repitiese el exp. infinitas veces. Por eso empleamos las letras griegas (μ), ya q se trata del parámetro correspondiente a la población de resultados del exp.

Para obtener la **varianza** de una V.A. X, δ_2 o V(X), debemos calcular el sumatorio del producto de cada uno de los valores q toma la V menos su media elevados al cuadrado multiplicados por su correspondiente valor de la función de probabilidad. Recordar la similitud con el índice estadístico correspondiente a la varianza de una V.E., T3.

$$\sigma^2 = V(X) = \sum (x - \mu)^2 \cdot f(x)$$

Una fórmula alternativa:

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

$E(X^2) = \sum x^2 f(x)$

$[E(X)]^2$ es la media de la variable elevada al cuadrado, μ^2.

Por tanto, la varianza puede definirse tb como la esperanza de los cuadrados de X, $E(X^2)$, menos el cuadrado de la esperanza de X, $[E(x)]^2$.

De manera análoga a las V.E, la **desviación típica δ** de una V.A.D. X es la raíz cuadrada de la varianza:

$$\sigma = \sqrt{\sigma^2} = \sqrt{\sum (x - \mu)^2 \cdot f(x)}$$

3. DISTRIBUCIONES DISCRETAS DE PROBABILIDAD

No siempre es necesario construir las funciones de probabilidad y de distribución, en función de las condiciones de partida del exp. y de las carac. de la VA, podemos ajustar estas distribuciones a alguna (modelo teórico de probabilidad) ya conocida, q simplifica mucho el trabajo si se ajuste según sus propiedades.

Existen diversas distribuciones teóricas para VD, bien conocidas, por utilizarse frecuentemente como modelo, o por su interés como instrumento estadístico:

- De **Bernoulli** y **binomial**;
- **Poisson** o ‘’sucesos raros’' q se utiliza bajo o las mismas condiciones de la binomial para V dicotómicas, pero con un elevado nº de ensayos y un valor de p muy pequeño.
- **Multinominal**; ensayos q ofrecen +2 resultados posibles. Supone una generalización de la binomial o un caso particular de aquella.

De muchas se han elaborado una serie de tablas q facilitan su aplicación a problemas concretos. Por lo general, en Psicología y Ciencias de la Salud se trabaja con VAD q sólo pueden tomar 2 valores (dicotómicas) y q habitualmente representaremos por 1 y 0. En estos casos, resultan muy útiles la distribución de Bernoulli, y, especialmente, su generalización a *n* ensayos, q es la distribución binomial.

La distribución de Bernoulli

Lanzar una moneda al aire 1 admite sólo 2 resultados; cara o cruz. El acierto o fallo a una pregunta con 2 alternativas respondida al azar, el lado izq. o derecho de un laberinto en forma de T elegido por una rata no entrenada... son algunos de los múltiples ej. A una de ellas se le denomina «**éxito o acierto**» (q habitualmente, se codifica con 1) y a la otra «fracaso o error» (q se codifica como 0), sin q estos términos tengan connotaciones ni positivas ni negativas. Es el fundamento y la base de otras distribuciones discretas, entre las q destaca la distribución binomial.

De este modo, la VAD q sigue el modelo de Bernoulli se define como una VA dicotómica X, con 2 posibles valores mutuamente exclusivos: 1 (éxito) con probabilidad *p* y 0 (fracaso) con probabilidad *q*; la suma de ambas =1.

f (1) = P(X = 1) = p
f (0) = P (X = 0) = q
p + q = 1, por lo q q = 1 -p

Una VA X q sigue el modelo de Bernoulli con parámetro *p*, se denota como **X → Ber (P)** y presenta las carac.:

- Función de probabilidad: $f(x) = P(X = x) = p^x q^{1-x}$
- Función de distribución: $F(x) = P(X \leq x) = \sum p^x q^{1-x}$
- Media: μ = *p*
- Varianza: $\delta^2 = p(1 - p) = pq$
- Desviación típica: $\delta = \sqrt{pq}$

x puede adoptar el valor 0 (fracaso) o 1 (éxito)
p = probabilidad de éxito en el único ensayo del exp.
q = probabilidad de fracaso (1-p) en el único ensayo del exp.

La distribución binomial

Generalización de la distribución de Bernoulli en la q el exp. se repite +1 vez. Se repite *n* veces, y de forma indep., un ensayo Bernoulli en el q la probabilidad de «éxito», p, se mantiene constante en cada uno de los *n* ensayos.

Una VA X sigue una distribución binomial (con parámetros n y p) si expresa el nº de éxitos en n realizaciones indep. de un exp. con probabilidad *p* de obtener «éxito» y, por tanto, (1-p) de obtener «fracaso». Suele representarse por ***B*(n, p)**.

***B:* binomial,**
***n:* nº de ensayos o veces q se repite un exp. Bernoulli.**
***p:* probabilidad de «éxito».**

La distribución de Bernoulli descrita sería un caso particular de la binomial con parámetro *n* =1 (un único ensayo), el parámetro *p* sería la probabilidad de «éxito», y se representaría como una binomial B(1, p).

Carac. de una distribución B(n, p):

- Función de probabilidad: $f(x) = P(X = x) = \binom{n}{x} p^x q^{n-x}$
- Función de distribución: $F(x) = P(X \leq x) = \sum \binom{n}{x} p^x q^{n-x}$
- Media: μ = np.
- Varianza: δ^2 = npq
- Desviación típica: $\delta = \sqrt{npq}$
- El nº combinatorio «n sobre x» $\binom{n}{x}$ es igual a $\binom{n}{x} = \frac{n!}{x!\,(n-x)!}$
  - el factorial de un nº *n* es: n ! = n •(n - 1) •(n - 2) • ... •(n - (n - 1))
  - el factorial 1! = 1 y el factorial 0!= 1

x = nº de aciertos
q = probabilidad de fracaso (1-p) en cada ensayo

Aunq es relativamente fácil deducir las carac. anteriores no lo vamos a hacer aquí formalmente, recurriremos a su aplicación. Si se lanza una moneda 3 veces y definimos la VA X como «nº de caras obtenidas», esta seguirá el modelo de distribución binomial con parámetros *n*=3 y *p*=0,5. Diremos q X sigue un modelo B(3, 0,5) pq en cada lanzamiento sólo son posibles 2 resultados: «éxito» (cara) y «fracaso» (cruz); los ensayos son indep. entre sí (el resultado en un ensayo no depende de lo q haya salido o no en los anteriores) y la probabilidad de «éxito» se mantiene constante.
La utilización de funciones de probabilidad y distribución requiere cálculos tediosos. Las tablas de la función de probabilidad y de la función de distribución binomial, Tablas I y II del Formulario, nos evitan el cálculo de las probabilidades a partir de la ecuación de esas funciones, facilitando su obtención con elevado *n*.

Tabla I, función de probabilidad binomial:

- 1ª columna, *n*, incluye los valores desde 1 hasta 20.
- 2ª columna, nº de «éxitos» (x) q esperamos obtener para ese nº de ensayos y q abarcan desde 0 hasta ese nº.
- 1ª fila, algunos valores de la probabilidad de «éxito» (p) desde 0,01 a 0,5.
- En el interior se encuentran las probabilidades correspondientes. La probabilidad buscada, para unos valores concretos de n y x, se encuentra en la intersección de su fila con la correspondiente columna de p. *La probabilidad de obtener 2 éxitos en 3 ensayos con una probabilidad de éxito de 0,3 - Tabla en la posición q se indica y vale 0,1890.*

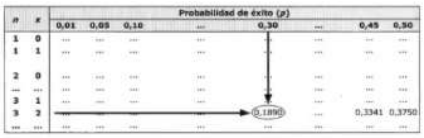


Tabla II:

- Idéntica a la anterior.
- Las probabilidades del interior son acumuladas.

Sin embargo, en ambas, sólo se contienen valores de p desde 0,1 hasta 0,5. ¿Q hacer cuando tengamos una p > 0,5? Hay q intercambiar las condiciones de «éxito» y «fracaso». Y, el nº de ensayos *n* sólo llega hasta 20. Esto no plantea ningún problema, para valores superiores podemos hacer una aproximación de la binomial a la distribución normal, como se verá en el próximo tema.

T8: MODELOS CONTINUOS DE PROBABILIDAD

Se han estudiado las VAD, unificando conceptos como la distribución de frecuencias y la probabilidad, q nos ha permitido definir su función de probabilidad, esperanza matemática y varianza teórica. De forma análoga se pueden definir estos mismos conceptos para las VAContinuas. Sin embargo, el problema es q estas no toman un nº finito de valores, toma infinitos. Tenemos q acudir a un modelo probabilístico q permite determinar, mediante el cálculo integral, la probabilidad de un intervalo de la V y no de un valor concreto como en las discretas. El proceso del cálculo de integrales no se va a tratar en este libro ni es necesario. Para estas distribuciones se han elaborado tablas q contienen los valores de las probabilidades correspondientes.

1. CARACTERÍSTICAS DE LAS VARIABLES ALEATORIAS CONTINUAS

En el tema anterior se definió una VAC como aquella q puede adoptar infinitos valores o un conjunto de valores no numerables. Dado q dentro de cada intervalo de valores existen a su vez infinitos valores posibles, la probabilidad de q tome un valor det. es nula. Es decir, obtener un det. valor de X = 0, por lo q las probabilidades se van a asignar a un det. intervalo. Para ello, se acude al concepto de función de densidad de probabilidad en torno a un valor, en lugar de función de probabilidad de un valor de las VAD.

Función de densidad función de distribución

Función de densidad de probabilidad de una VAC, f(x): función q cumple las 2 condiciones siguientes:

1. Los valores de f(x) son siempre = 0 o positivos, nunca negativos.
2. Una integral, q en VC es el análogo al sumatorio en VDiscretas. El área total (desde -∞ hasta +∞ en la variable X) bajo la curva es igual a uno. De ahí q se aplique para la determinación de las probabilidades correspondientes a las variables continuas.

a) $f(x) \geq 0$

b) $\int_{-\infty}^{+\infty} f(x)\,dx = 1$

Podemos calcular la probabilidad de q X se encuentre en un det. intervalo [a,b] mediante el cálculo integral:

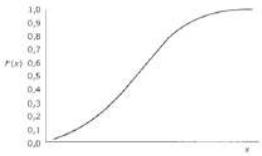
$$P(a \leq X \leq b) = \int_a^b f(x)\,dx$$

f(x): función de densidad de probabilidad de X
La integral está definida para el intervalo [a,b]

f(x) no se corresponde con un valor puntual de probabilidad sino de densidad q, aplicándole el cálculo de integrales, permite obtener la probabilidad para un intervalo. Dicho de otro modo, f(x) no es una probabilidad, pero la integral de f(x) para un det. intervalo [a, b] de X si nos proporciona un valor de probabilidad .

La función de distribución, F(x), se describe como para las VD; la probabilidad acumulada hasta un cierto valor de la V.
Función de distribución acumulada o función de distribución de probabilidad de una VAC, F(x), aquella función q asocia a cada valor de la variable X la probabilidad de obtener valores menores o iguales q un valor dado (= menor, ya q la probabilidad de ser igual al valor dado es 0).

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(x)\,dx$$



Para la función de distribución en el caso continuo se mantienen las mismas propiedades q en las VAD:

- F(x) ≥ 0 —> todos los valores q toma la función de distribución son nulos o positivos F(x).
- F(-∞) =0 y F(+∞) =1 por lo q 0 ≤ F(x) ≤ 1 —> F(x), al ser una probabilidad, está acotada entre 0 y 1.
- ∀a ≤ b, P(a ≤ X ≤ b) = F(b) - F(*a*) —> la probabilidad de q X se encuentre en el intervalo [a,b] es la diferencia entre la función de distribución para X= b, F(b) y la función de distribución para X= a, F(a).

Media y varianza de una variable aleatoria continua

Como las VD, las VAC tb presentan una media o valor esperado y una varianza, q pueden obtenerse mediante procedimientos análogos a las VD, pero adaptados para el caso continuo.

Sea X una VAC, la **media** o **valor esperado**, μ o E(X):

$$\mu = E(X) = \int_{-\infty}^{+\infty} x \cdot f(x)\,dx$$

La **varianza**:

$$\sigma^2 = V(X) = \int_{-\infty}^{+\infty} [x-\mu]^2 \cdot f(x)\,dx$$

VARIABLES ALEATORIAS DISCRETAS	VARIABLES ALEATORIAS CONTINUAS
Probabilidad para X = x $f(x) = P(X = x)$ $f(x)$ = función de probabilidad	***Probabilidad para el intervalo [a,b]*** $P(a \leq X \leq b) = \int_a^b f(x)\,dx$ $f(x)$ = función de densidad de probabilidad
Función de Distribución $F(x_k) = P(X \leq x_k) = f(x_1) + f(x_2) + \ldots + f(x_k)$	***Función de Distribución*** $F(x) = P(X \leq x_k) = \int_{-\infty}^{x_k} f(x)\,dx$
Media o Valor Esperado $\mu = E(X) = \sum x \cdot f(x)$	***Media o Valor Esperado*** $\mu = E(X) = \int_{-\infty}^{+\infty} x \cdot f(x)\,dx$
Varianza $\sigma^2 = V(X) = \sum (x-\mu)^2 \cdot f(x)$	***Varianza*** $\sigma^2 = V(X) = \int_{-\infty}^{+\infty} [x-\mu]^2 \cdot f(x)\,dx$

Las propiedades de la media y la varianza para VC son las mismas q las discretas. Por otra parte, para el cálculo de las q estudiaremos es necesario utilizar el cálculo integral, pq se han derivado fórmulas directas para la obtención de dichos parámetros.

2. DISTRIBUCIÓN NORMAL - CAMPANA DE GAUSS / CURVA NORMAL

Definida por De Moivre en un intento de encontrar las probabilidades acumuladas en una distribución binomial cuando n (nº de ensayos) es grande.

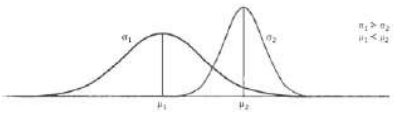
Función de densidad de probabilidad para una variable X q tiene distribución normal:

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2} \quad \text{para} \quad -\infty < X < \infty$$

μ: parámetro media o valor esperado de la distribución.
δ: parámetro desviación típica de la distribución.
π = 3,1416
***e* = 2,718 (base de los logaritmos neperianos).**

Si una variable X tiene una función de densidad q se ajusta a la fórmula anterior, lo expresaremos por: X —> *N*(μ, **δ**), indicando q tiene una distribución Normal (N) con parámetros μ y **δ**.

No se trata de una única distribución, corresponde a toda una familia caracterizada por sus parámetros media, μ, y desviación típica, **δ**. Como puede observarse su forma de «campana» es más apuntada cuanta menor desviación típica. Su figura nos indica q la puntuación de la mayoría de los individuos, en una V q sigue esta distribución, se encuentra en torno a la media y, a medida q nos alejamos va disminuyendo la frecuencia.



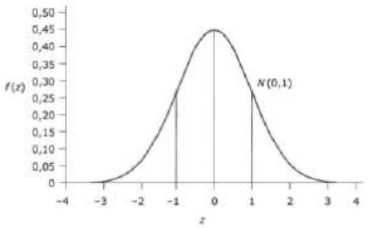
Según una de sus propiedades fundamentales, si a una variable X q se distribuye normalmente, con μ y **δ** le aplicamos una transformación lineal de la forma Y=bX+*a*, la nueva variable Y tb se distribuirá normalmente pero con media $\mu_Y=b\mu_X+a$ y desviación típica $\delta_Y=|b|\cdot\delta_X$

Por otra parte, si restamos la media y dividimos por la desviación típica obtenemos una nueva V q designamos por z. Esta se distribuirá normalmente con media =0 y desviación típica igual =1 **z → N(0,1)**. La demostración de $\mu_2=0$ y $\delta_2=1$ excede el presente curso.

$$z=\frac{X-\mu}{\sigma}$$

La **función de densidad de probabilidad de z**:

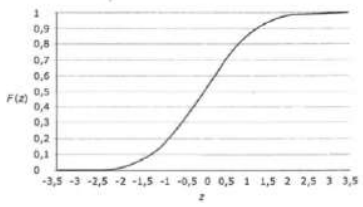
$$f(z)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{z^2}{2}} \quad \text{para} \quad -\infty < z < \infty$$



Esta distribución se denomina normal tipificada o normal estandarizada. Para la aplicación a problemas concretos recurriremos a las Tablas III y IV del Formulario. Si observamos la figura entre las propiedades fundamentales de una distribución normal podemos destacar:

- Es simétrica en torno a su media, μ, q coincide con su mediana y su moda.
- La curva normal tiene 2 puntos de inflexión, donde la curva pasa de ser cóncava a convexa. Estos están situados a una distancia de una desviación típica de la media.
- Es asintótica en el eje de abscisas; se extiende desde -∞ hasta +∞ sin llegar nunca a tocar el eje X.

Su función de distribución —>



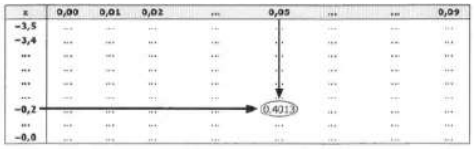
Utilización de las tablas

En las Tablas III y IV se recoge la función de distribución de la curva normal estándar. Se presentan todas las puntuaciones típicas desde -3,59 hasta +3,59 con intervalos de 0,01.

- 1ª columna, encabezada con la letra z, consta de un nº con un decimal, la puntuación típica.
- 1ª fila (derecha de la letra z), 2º decimal de la puntuación z.
- Todos los valores interiores representan probabilidades = llevan un cero delante de la coma.

Tabla III—> puntuaciones típicas negativas (por debajo de la media). Tabla IV —> las positivas (por encima de la media).

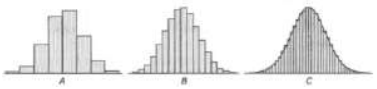
La puntuación típica z = -0,25 (Tabla III) deja por debajo de sí una probabilidad de 0,4013. Tabla III. La puntuación típica z=0,25 (Tabla IV) deja por debajo de sí una proporción de 0,5987. Al ser una distribución simétrica puede comprobarse que la proporción que queda por debajo de z =-0,25 es igual a la proporción q queda por encima de z =0,25 (1- 0,5987 =0,4013). Si la Tabla no recoge el valor exacto de z q se busca, se puede utilizar el + próximo. Algunos casos concretos son:



1. **Cálculo de la probabilidad para valores menores o iguales q una det. puntuación típica.**
   Se busca directamente en la Tabla.
2. **Cálculo de la probabilidad para valores mayores q una det. puntuación.**
   Se mira en la tabla la probabilidad q esa puntuación deja por debajo y se resta de 1.
3. **Cálculo de la probabilidad entre dos puntuaciones det.**
   Se restan las probabilidades q dejan por debajo de sí las 2 puntuaciones típicas.

Histograma y distribución normal

Datos de una muestra en una variable X (A). Si se hacen los intervalos + pequeños (B) y dibujamos el polígono de frecuencias (C) llegamos a una distribución similar a la normal. La puntuación de la mayoría de los casos se encuentra en torno a la media y, a medida q nos alejamos va disminuyendo. Esto va a permitir aplicar las propiedades de la curva normal a nuestros datos y utilizar las tablas de la misma forma q se ha visto.

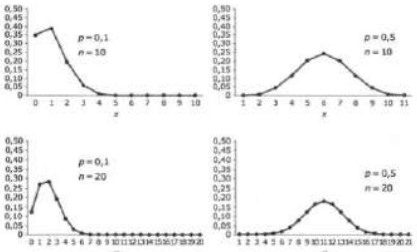


$$z_i = \frac{X_i - \overline{X}}{S_x}$$

Si se dispone de los datos originales de un grupo de sujetos en una det. variable X, y ésta se distribuye normalmente, para resolver det. cálculos utilizando Tablas III y IV, debemos transformar las puntuaciones directas en puntuaciones típicas mediante la fórmula de la izq.

Aproximación de la binomial a la normal

Cuando para la distribución binomial tenemos un n superior a 20, debemos aproximar la distribución binomial a la normal. Esta aproximación mejora a medida q *p* (probabilidad de éxito) se aproxima a 0,5 y *n* es grande:



Se sabe q una variable, X, q sigue una distribución binomial tiene una media μ =np y una desviación típica $\delta = \sqrt{npq}$. Por tanto, se puede transformar su función de probabilidad (q es discreta) a la normal:

La distribución normal es continua y, como para cualquier distribución continua, la probabilidad de q la variable X tome un valor concreto es cero: P (X= x) = O. Para aproximar la distribución binomial a la normal se establecerá un intervalo entre 0,5 unidades a la izq. y a la derecha de la puntuación:

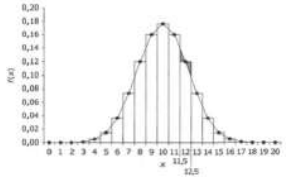
$$P(X = x) = P[(x - 0,5) \leq x \leq (x + 0,5)]$$

A continuación, transformamos las puntuaciones en típicas:

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

Debido a q μ = *np* y $\delta = \sqrt{npq}$, la **aproximación de la binomial a la normal** se define como:

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



En el caso anterior, sumar y restar el valor 0,5 se llama **corrección por continuidad**, permitiendo utilizar las puntuaciones discretas, X, como si fuesen continuas. Para ello, se interpreta cada puntuación, X, como si fuesen los puntos medios de sus intervalos. Se intenta asegurar q el intervalo incluya los valores discretos de la binomial.

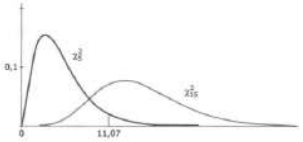
3. LA DISTRIBUCIÓN DE PEARSON

Ya se ha visto en el T4 el estadístico ji-cuadrado (x^2), para referirse a la correlación entre V cualitativas. A partir de ahora, x2 se va utilizar para hacer referencia a una distribución continua de probabilidad.

Sean X_1, X_2, ..., X_n un conjunto de *n* variables aleatorias indep. con una distribución N(0,1), entonces una nueva V aleatoria $X = x^2_1 + x^2_2 + \ldots + X^2_n$ sigue una distribución $X \rightarrow X^2_n$ (Ji- cuadrado con *n* grados de libertad)·

- Media o valor esperado μ =n
- Varianza $\delta^2 = 2n$

Los grados de libertad (n) indican q cada una de las *n* variables aleatorias puede tomar cualquier valor de sus posibles valores, sean cuales sean los tomados por las *n*-1 restantes.



Esta distribución se usa fundamentalmente en pruebas de bondad de ajuste (para contrastar si la distribución de una V se ajusta a una distribución det.l). Al igual q otras distribuciones, es una familia de curvas como las presentadas, q varían en función de los grados de libertad.

- Nunca adopta valores menores de 0.
- Es asimétrica positiva pero a medida q ↑ sus grados de libertad se va aproximando a la distribución normal.
- Para *n* > 30 la podemos aproximar a una distribución N (n, $\sqrt{2n}$).

La Tabla V del Formulario permite obtener las probabilidades acumuladas a algunos valores de toda la familia de distribuciones.

- 1ª fila —> probabilidades o proporciones.
- 1ª —> columna los grados de libertad correspondientes.
- En el interior —> los valores de la V. *Para una V q sigue una distribución x^2 con 5 grados de libertad, X ➔ X^2_5, el valor 11,07 deja por debajo una proporción de 0,95. Por tanto, P (X ≤ 11,07) = 0,95. Esta puntuación se corresponde con el percentil 95. Suele presentarse $_{0,95}X^2_5 = 11,07$.*


Tabla V del Formulario

Si lo q interesa es hallar P (X ≥ 11, 07):

$$P(X \geq 11,07) = 1 - P(X \leq 11,07) = 1 - 0,95 = 0,05$$

4. LA DISTRIBUCIÓN T DE STUDENT

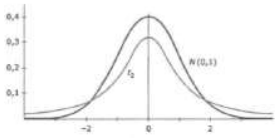
A la h de definir este tipo de distribución de probabilidad, al igual q se hizo anteriormente con x^2, se hará en función de otras distribuciones ya conocidas.

Sean X e Y dos V aleatorias independientes, donde X sigue una distribución N(0,1) e Y una distribución X^2_n. Entonces, la V aleatoria $T = \frac{X}{\sqrt{Y/n}}$ una distribución *t* con *n* grados de libertad y se expresa: T ➔ t_n Sus parámetros son :

- Media o valor esperado μ = 0
- Varianza δ^2 = n/n-2

Distribución t: el cociente entre una variable *N*(0,1) y la raíz cuadrada de una variable X^2_n dividida por sus grados de libertad. Su nombre se debe a su descubridor, el matemático Gosset, q publicó bajo el seudónimo de Student.

Distribución t con dos grados de libertad, junto a la distribución normal estándar —>



Características:

- Es simétrica, con μ = 0. Su forma es muy parecida a la *N*(0,1), aunq menos apuntada.
- Puede tomar cualquier valor entre -∞ y +∞.
- A medida q ↑ los grados de libertad, la distribución se aproxima más a una distribución normal.
- La curva es asintótica al eje de abscisas.

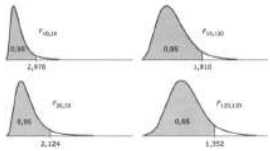
Fundamentalmente se utiliza en estadística inferencial. En la Tabla VI del Formulario se presentan los valores positivos.

- 1ª columna: grados de libertad.
- 1ª fila: distintas probabilidades o proporciones de valores menores o iguales q un valor positivo dado.
- Como es una distribución simétrica podemos hallar las probabilidades asociadas a valores negativos a partir de los valores positivos de la Tabla VI.

5. LA DISTRIBUCIÓN DE F DE FISHER-SNEDECOR

Si X_1 y X_2 son V aleatorias indep., con distribución x^2 con n_1 y n_2 grados de libertad respectivamente; una nueva variable F —> $F = \frac{X_1/n_1}{X_2/n_2}$ sigue una distribución F con n_1 y n_2 grados de libertad (F_{n1n2}). Siendo n_1 los grados de libertad del numerador y n_2 los del denominador.

- Su media o valor esperado viene definido por $\mu = \frac{n_2}{n_2 - 2}$ para $n_2 > 2$
- Su varianza por $\sigma^2 = \frac{2n_2^2(n_1 + n_2 - 2)}{n_1(n_2 - 4)(n_2 - 2)^2}$ para $n_2 > 4$



Fundamentalmente en el contraste de hipótesis (Análisis de Varianza...). Figura = representación según distintos grados delibertad.

Características:

- Es asimétrica positiva, nunca toma valores menores q 0.
- Propiedad recíproca; si X es una V con distribución F con n_1 y n_2 grados de libertad, entonces la variable Y = 1/X es tb una distribución F con n_2 y n_1 grados de libertad:
  p y 1 - *p*: probabilidades acumuladas asociadas al valor de la V. Esta propiedad es útil para obtener algunos percentiles o probabilidades q no aparecen en la tabla.
- La Tabla VII recoge solamente la probabilidad de q X sea menor o igual q 0,900; 0,950; 0,975; 0,990 y 0,995, los valores utilizados habitualmente.

$$pF_{n_1,n_2} = \frac{1}{1-pF_{n_2,n_1}}$$

T9: MUESTREO Y DISTRIBUCIÓN MUESTRAL DE UN ESTADÍSTICO

En temas anteriores se ha presentado la materia correspondiente a la Estadística Descriptiva. En los 2 siguientes temas se presentarán las bases sobre las q se asienta la **Inferencia Estadística** (disciplina q abarca las técnicas y métodos q permiten deducir las propiedades desconocidas de la población a partir de los datos obtenidos en la muestra). La inf. q obtenemos de las muestras permite estudiar el comportamiento de las VA (discretas o continuas) y de los índices estadísticos q las representan (q son tb VA). Sobre esta base, apoyándonos en la teoría de muestreo, podremos estimar los valores de los parámetros a partir de los valores de los correspondientes índices estadísticos.

ESTADÍSTICA DESCRIPTIVA
Recogida, organización y análisis de los datos.
PROBABILIDAD
Permite y legitima el salto de las carac. (conocidas) de la muestra hasta las desconocidas de la población.
ESTADÍSTICA INFERENCIAL
Obtener inf. acerca de la población a partir de la aportada por la muestra mediante la combinación de los modelos de probabilidad y de los estadísticos. Comprende los campos estimación de parámetros y contraste de hipótesis. La piedra angular de ambos es el concepto de distribución muestral de un estadístico, q establece la relación entre las carac. de la población y el comportamiento de los estadísticos de las muestras q las representan.

1. MUESTREO

Al plantearse la recogida de datos suele ocurrir q por razones de coste económico, tiempo o imposibilidad real, se hace muy difícil estudiar a todo el grupo sobre el q se quiere trabajar. Por lo general es imposible trabajar con el conjunto total población. Por tanto, tendremos q bus- car una muestra (subconjunto). Lo ideal es elegir las muestras de tal forma y modo q representen y sean fiel reflejo de las carac. relevantes a la investigación en la población de trabajo.

Es posible trabajar con «unos pocos» datos extraídos de un conjunto más amplio y q las conclusiones sean válidas para el conjunto total. Pero, ¿cuáles son los procedimientos de selección q lo garantizan? ¿es posible cuantificar el riesgo de equivocación/ error al realizar afirmaciones sobre las poblaciones?

Conceptos básicos en el muestreo

- **Población**: colección, finita o infinita, de elementos q comparten ciertas carac. *Todos los SH componen la población de hombres y mujeres; todas las personas que de forma habitual se expresan en francés componen la población de francófonos*. El concepto es indep. de la cantidad de elementos, puede estar compuesta por 1 solo; *población de satélites de la Tierra, la Luna*. 1 elemento pueda pertenecer a +1 población si cumple los criterios necesarios. Un elemento pertenece a una población solo durante el tiempo q se cumplen las propiedades. Atendiendo al nº de elementos, las poblaciones pueden ser finitas o infinitas. En general, son muy grandes y esto hace q sea prácticamente inviable trabajar con ellas y lo habitual es trabajar con muestras. Los índices q representan los valores q resumen las carac, de las poblaciones = **parámetros**. Son constantes, se calculan con todos los elementos de la población.
- **Censo**: en det. ocasiones resulta posible estudiar a todos y cada uno de los elementos q componen la población, realizándose el estudio de todos los elementos q componen la población. *Censo poblacional.*
- **Muestra**: cuando no es posible o conveniente trabajar con la población. Es representativa cuando reúne las mismas carac. q la población. La representatividadegarantiza q los resultados puedan ser generalizados. Los índices q representan los valores q resumen las carac. de las muestras = **estadísticos** y son VA cuyos valores varían en función de los elementos q compongan la muestra.
- **Muestreo**: conjunto de procedimientos y técnicas q permiten extraer muestras de una población garantizando la representatividad o de procedimientos y reglas para poblaciones. La generalización, desde la parte al todo, supone siempre un cierto error q se puede cuantificar y controlar a través de la Estadística mediante dos conceptos: error máximo (E_{max}) y nivel de confianza ($1-\alpha$). Pasos a seguir:
  A. Definir los casos (participantes u otros seres vivos, objetos, fenómenos o comunidades) sobre los cuales se habrán de recolectar los datos.
  B. Delimitar la población mediante una carac. q defina, de forma exhaustiva y excluyente, a los individuos.
  C. Elegir el método de selección de la muestra. (T9)
  D. Calcular el tamaño de la muestra. (T10)
  E. Aplicar el procedimiento de selección.
  F. Obtener la muestra.

Tipos de muestreo

- **Muestreo probabilístico**: se conoce la probabilidad q tiene cada elemento de la población de ser elegido para formar parte de la muestra y se conoce el **marco muestral** (listado de elementos q componen la población). Su ventaja es garantizar la representatividad = permitir hacer inferencias.
- **Muestreo no probabilístico**: no se conoce la probabilidad q tiene cada elemento de ser elegido. No garantiza la representatividad = conclusiones circunscritas a la situación sin posibilidad de generalizar.

Dentro de cada una de estas categorías hay dif. métodos según la forma en q se extraen los elementos:

Tipos de muestreo	Métodos de muestreo
Probabilistico	Aleatorio Simple Aleatorio Sistemático Aleatorio Estratificado Aleatorio por Conglomerados
No probabilistico	Por Cuotas Intencional (opinático) Incidental (casual) Bola de Nieve

Métodos de muestreo probabilístico

MUESTREO ALEATORIO SIMPLE
Consiste en tomar de una población de tamaño *N* una muestra de tamaño *n*, utilizando algún procedimiento q garantice q todos los elementos de la población tienen la misma probabilidad de ser elegidos.

- Se asigna un nº a cada elemento de la población.
- A través de algún medio mecánico o informático *(bolas dentro de una bolsa, tablas de nº aleatorios, nº aleatorios generados con una calculadora u ordenador…)* se eligen tantos sujetos como sea necesario para completar el tamaño de muestra.

Todas las muestras posibles son **equiprobables**, tienen la misma probabilidad de ser elegidas. Pero, la probabilidad de cada una de ellas y la probabilidad de pertenencia de los elementos será distinta en función de la forma en q se genere la muestra. Para formar la muestra se puede actuar de 2 modos:

- **Con reposición**: Tras elegir un elemento, éste se reincorpora a la población para q pueda ser elegido en la siguiente extracción, así la población siempre tiene el tamaño *N*. Cada elemento de la población (formada por N elementos) tiene la misma probabilidad (1/N) en cualquiera de las extracciones de pertenecer a la muestra (formada por n elementos). La probabilidad de obtener una muestra concreta de *n* elementos es: (1/N)x(1/N)x(1/N)x ... x(1/N) =$(1/N^n)$. Es decir, q de las N^n muestras posibles de tamaño *n*, todas tienen la misma probabilidad de ser elegidas.
- **Sin reposición**. Una vez seleccionado un elemento de la población no se reintegra, la población va perdiendo tamaño. En la 1ª extracción, el tamaño es N, en la 2ª es N-1, en la 3ª es N-2… El tamaño de la población cambia con cada extracción. Aunq todos los elementos tienen la misma probabilidad de ser elegidos, esa va cambiando y el resultado de una extracción no es indep. del resultado obtenido en las demás. En este caso, la función de probabilidad conjunta de las V es diferente del producto de sus funciones de probabilidad individuales. Existen $\binom{N}{n}$ muestras posibles de tamaño *n*, y la probabilidad de una muestra concreta es $\frac{1}{\binom{N}{n}}$.

La gran mayoría de procedimientos de la Estadística Inferencial exigen el principio de indep. en la obtención de muestras, lo cual no se cumple en el muestreo aleatorio sin reposición (muy habitual en investigación). Esto se resuelve considerando q, cuando el tamaño de la población (N) es grande con respecto al de la muestra (n), las probabilidades calculadas con ambos muestreos (con y sin reposición) son prácticamente iguales. Finalmente, este tipo de muestreo, aunq resulta atractivo por su sencillez, tiene poca o nula utilidad práctica cuando la población es muy grande.

MUESTREO ALEATORIO SISTEMÁTICO
Cuando los elementos de la población están ordenados o pueden ordenarse. Es necesario asignar un nº a todos los elementos de la población, pero, en lugar de extraer *n* números aleatorios sólo se extrae uno (*i*). El nº i del q se parte es elegido al azar, y los elementos q compondrán la muestra ocupan los lugares i, i+k, i+2k, i+3k ... i + (n - 1)k, siendo k el resultado de dividir el tamaño de la población entre el tamaño de la muestra (k = N / n). Se toman los individuos de k en k partiendo del sujeto en la posición i. El nº i q empleamos como punto de partida será al azar entre 1 y k. En este tipo de muestreo no todos los elementos tienen la misma probabilidad de ser extraídos, y las extracciones no son indep. El riesgo está en aquellos casos en q se dan periodicidades en la población, ya q al seleccionar una periodicidad constante, los elementos seleccionados para la muestra pueden no ser representativos.

MUESTREO ALEATORIO ESTRATIFICADO
Cuando la población no es homogénea, existen grupos o estratos heterogéneos entre sí con gran homogeneidad dentro del estrato (se puede estratificar *según la profesión, municipio de residencia, sexo, estado civil…*). Se pretende asegurar q todos los estratos de interés estén representados adecuadamente en la muestra. Cada uno funciona de forma indep., pudiendo aplicarse dentro de ellos el muestreo aleatorio simple o sistemático. El procedimiento de composición de la muestra en los dif. estratos se denomina **afijación**, y puede ser de dif. tipos:

- **Afijación simple**: a cada estrato le corresponde igual nº de elementos muestrales.
- **Afijación proporcional**: la distribución se hace de acuerdo con el peso (tamaño) de la población en cada estrato.

La muestra total se forma por la suma de las muestras de cada estrato. Cada submuestra es indep. Permite aplicar técnicas de selección dif. dentro de cada estrato y obtener estimaciones separadas en cada una.

MUESTREO ALEATORIO POR CONGLOMERADOS

Los 3 métodos de muestreo presentados hasta ahora están diseñados para seleccionar directamente los elementos de la población; las unidades muestrales (los sujetos) son los elementos de la población. En el muestreo por conglomerados, la unidad muestral es un grupo de elementos de la población q conforman una unidad más amplia. *Áreas sanitarias, dpt. universitarios, una caja de det. producto…*

El procedimiento consiste en seleccionar aleatoriamente un cierto nº de conglomerados (el necesario para alcanzar el tamaño muestral) y trabajar con todos los elementos pertenecientes a los conglomerados elegidos. Si el nº de es muy amplio seleccionamos algunos al azar; **muestreo por conglomerados bietápico**. En general se habla de muestreo por etapas o polietápico cuando hay +2 etapas. Para aplicarlo en cada etapa se van seleccionando conglomerados de menor tamaño hasta q en la última etapa se trabaja con los *n* elementos q componen esos conglomerados. *En un estudio sobre la población universitaria española se seleccionan Universidades; dentro de ellas Facultades, dentro de ellas carreras específicas y dentro los cursos, el último conglomerado. La muestra estaría formada por todos los individuos de los cursos.*

Métodos de muestreo no probabilístico

Cuando no es posible realizar un muestreo probabilístico pq desconocemos la probabilidad de inclusión de cada elemento en la muestra y/o tiene un excesivo costo económico o de tiempo. Se acude a métodos no probabilísticos, aun siendo conscientes de q no sirven para realizar generalizaciones pq no se tiene certeza de q la muestra extraída es representativa, ya q no todos los sujetos de la población tienen la misma probabilidad de ser elegidos. En general, se selecciona siguiendo det. criterios, procurando q la muestra resultante sea lo más parecida posible a la población.

MUESTREO POR CUOTAS / ACCIDENTAL

Base; un buen conocimiento de los estratos de la población y/o de los individuos + adecuados para la investigación. Semejanzas con el muestreo aleatorio estratificado, pero no tiene su carácter de aleatoriedad.

Se fijan unas «**cuotas**» (nº de individuos q reúnen unas det. condiciones). Una vez determinadas se eligen los primeros q se encuentren q cumplan esas carac. Se utiliza mucho en las encuestas de opinión. *La Consejería de Sanidad de una CA desea estudiar la incidencia de las drogas en la adolescencia. A través de los informes de la Consejería de Educación, se conoce cuáles son los centros educativos + afectados, se fija un nº de sujetos a entrevistar proporcional a cada uno de los centros fijados y, finalmente, se deja en manos de los encuestadores a q sujetos concretos se entrevista.*

MUESTREO OPINÁTICO O INTENCIONAL

Se caracteriza por un esfuerzo deliberado de obtener muestras «representativas» mediante la inclusión en la muestra de grupos supuestamente típicos. Muy frecuente su utilización en sondeos preelectorales de zonas q en anteriores votaciones han marcado tendencias de voto. *Caso del Estado de Ohio (EEUU) en relación con las elecciones a presidente en los EEUU, es un estado cambiante (swing state), no tiene una tendencia del voto. Desde comienzos del sXX sólo en 2 ocasiones falló la predicción; no coincidió el presidente elegido según los resultados obtenidos en el estado con el q el q realmente salió elegido. ¿Por q coinciden tantas veces? La respuesta viene det. por la gran diversidad q presenta su población, hay representantes de todos los tipos de votantes: de áreas rurales, de grandes ciudades, cristianos conservadores, afroamericanos y muchos trabajadores. Por tanto, es un Estado muy representativo, la gran mayoría de los dif. tipos de votantes están presentes en él.*

MUESTREO CASUAL O INCIDENTAL

Proceso en el q el investigador selecciona directa e intencionadamente los individuos de la población. El caso más frecuente es utilizar como muestra los individuos a los q se tiene fácil acceso *(profesores de universidad a sus alumnos)*.

MUESTREO DE BOLA DE NIEVE

Las unidades muestrales van incorporándose paulatinamente a partir de la referencias de los sujetos q ya han participado. Se localizan algunos individuos, los cuales conducen a otros, y estos a otros. Muy frecuente cuando se hacen estudios con poblaciones marginales, difíciles de identificar y localizar.

2. DISTRIBUCIÓN MUESTRAL DE UN ESTADÍSTICO

A una población se le mide una carac., *la altura*. Con estos datos se podrá hacer una distribución de frecuencias (los resultados se referirán a toda la población). Se podrá calcular su media y varianza. Si se hace con una muestra se obtiene la distribución muestral; se construye su distribución de frecuencias y calcula la media y varianza, q son en este caso los estadísticos, $\bar{X}$ y S^2.

Por otra parte, en una población cualquiera es posible extraer +1 muestra dif. del mismo tamaño. Por tanto, el valor concreto de un estadístico dependerá de los valores concretos q tomen cada uno de los elementos de la muestra. El estadístico obtenido ya no será una constante sino una V, su valor concreto dependerá de la muestra en la q se haya calculado. La distribución de probabilidad de todos los posibles valores del estadístico en las dif. muestras = **distribución muestral del estadístico**. Dicho de otro modo, dada una población de la q se van a extraer varias muestras, y para cada una se calcula un estadístico *(por ej la X)* de una V aleatoria X cualquiera. Puesto q la media (X) toma dif, valores, dependiendo de cada muestra, el conjunto de las distintas medias forman a su vez una VA q tendrá su propia distribución de probabilidad con sus carac.: forma, media y varianza, los parámetros q la definen q se representarán por letras griegas con un subíndice, q nos indica a q estadístico nos estamos refiriendo.

Por tanto, la **distribución muestral de un estadístico** es la distribución de probabilidad teórica de los valores de un estadístico cuando estos se calculan sobre las k muestras (siendo k muy grande, teóricamente infinito) de tamaño *n*, extraídas de la población y obtenidas mediante muestreo aleatorio simple.

Cada vez q queramos estimar un parámetro, ¿debemos extraer tantas muestras como sea posible, calcular la media o la proporción en todas esas muestras y luego obtener la media de todas las medias o la media de todas las proporciones calculadas?¿Y lo mismo con los estadísticos varianza, mediana o coeficiente de correlación de Pearson, por ej.? Para responder se retomará el concepto de esperanza matemática de una VA, la mejor opción para estimar parámetros desconocidos utilizando los modelos de probabilidad.

Si de cualquier población con media μ y desviación típica **δ**, se toman todas las posibles muestras aleatorias con reposición, cada una de tamaño *n*, la distribución muestral del estadístico media tiene como parámetros:

- $\mu_{\bar{x}} = \mu$ —> la media de las medias es igual a la media poblacional.
- Desviación típica o error típico de la media $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

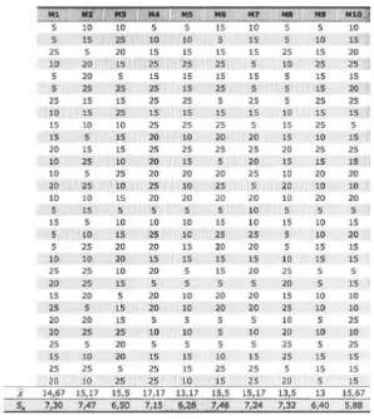
En el caso de muestras finitas y muestreo sin reemplazamiento, la desviación típica deberá multiplicarse por el factor de corrección definido por $\sqrt{\frac{N-n}{N-1}}$. En este caso:

- $\mu_{\bar{x}} = \mu$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

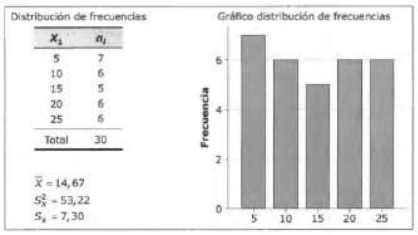
Análisis con 10 muestras aleatorias de tamaño n = 30

10 columnas = valores observados en la V para cada muestra, desde la Muestra 1 (M1) hasta la 10 (M10).
En las 2 últimas filas figuran la media y la desviación típica en cada muestra.



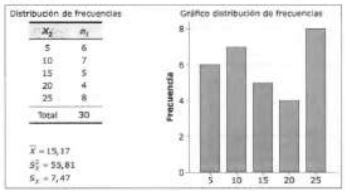
A continuación, se presentan los datos, la representación gráfica de la distribución de la variable X y los estadísticos muestrales de la MI.

Muestra 1

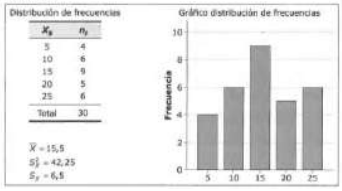


Tras repetir el exp. 9 veces +, se obtuvieron las 9 columnas siguientes de la tabla (en los gráficos se muestran las representaciones de cinco de estas muestras con los valores de sus estadísticos).

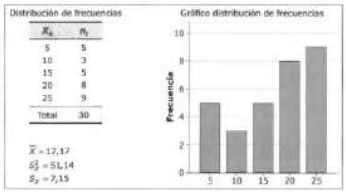
Muestra 2



Muestra 3



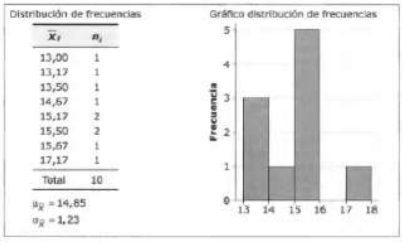
Muestra 4



...

Distribución muestral de las 10 medias

Distribución de frecuencias, representación gráfica y cálculos de estadísticos correspondientes a la media y desviación típica de las 10 medias.

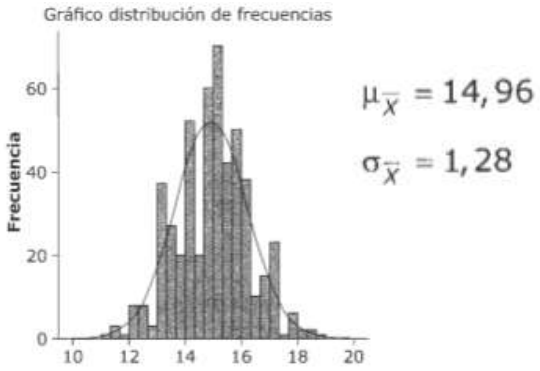


Como se puede observar, la media de todas las medias para 10 muestras ($\mu_{\bar{x}} = 14,85$) se aproxima bastante a la media poblacional (μ =15). Sin embargo, la desviación típica de las medias no tiene, en principio, nada q ver con la desviación típica poblacional 1,23 ≠ 7,07 como era de esperar. No obstante, se puede comprobar q el error típico de la media de este estudio empírico es a x = 1, 23, siendo el teórico: $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{7,07}{\sqrt{30}} = 1,29$

La diferencia entre ambos se debe al nº de muestras tan reducido para construir esta distribución muestral.

Análisis con 500 muestras aleatorias de tamaño n = 30

11,00	11,50	11,90	12,00	12,00	12,10	12,20	12,20	12,20	12,30
12,30	12,50	12,50	12,50	12,70	12,90	12,90	13,00	12,50	13,17
13,17	13,17	13,17	11,50	11,50	13,17	13,20	13,20	13,20	13,20
13,20	13,20	13,20	13,20	12,50	13,20	13,20	13,20	13,20	13,20
13,20	13,20	13,20	13,20	13,20	13,20	13,20	13,20	13,20	13,20
13,20	13,20	13,23	13,23	13,23	13,30	13,30	13,30	13,44	13,50
13,50	13,50	13,50	13,50	13,50	13,50	13,50	13,50	13,50	13,50
13,50	13,50	13,50	13,50	13,50	13,50	13,50	13,50	13,50	13,50
13,55	13,60	13,60	13,60	13,67	13,67	13,70	12,50	13,70	13,70
13,70	12,50	13,70	13,70	13,70	13,70	13,70	13,70	13,80	13,80
13,90	13,90	13,90	13,90	13,90	13,90	14,00	14,00	14,00	14,00
14,00	14,00	14,00	14,00	14,00	14,00	14,00	14,10	14,10	14,10
14,11	14,17	14,17	14,17	14,17	14,20	14,20	14,20	14,20	14,20
14,20	14,20	14,20	14,20	14,20	14,20	14,20	14,20	14,20	14,20
14,20	14,20	14,20	14,20	14,20	14,20	14,20	14,21	14,30	14,30
14,30	14,30	14,30	14,30	14,30	14,30	14,40	14,40	14,44	14,50
14,50	14,50	14,50	14,50	14,55	14,60	14,60	14,60	14,60	14,60
14,60	14,66	14,66	14,66	14,67	14,67	14,67	14,70	14,70	14,70
14,70	14,70	14,70	14,70	14,70	14,70	14,70	14,70	14,70	14,70
14,70	14,70	14,70	14,70	14,70	14,70	14,70	14,70	14,70	14,70
14,70	14,80	14,80	14,80	14,80	14,80	14,80	14,80	14,80	14,80
14,80	14,80	14,80	14,80	14,80	14,80	14,80	14,80	14,80	14,90
14,90	14,90	14,90	14,90	14,90	14,90	14,90	14,90	14,90	14,90
14,90	14,90	14,90	14,90	15,00	15,00	15,00	15,00	12,50	15,00
15,00	15,00	15,00	15,10	15,10	15,10	15,10	15,10	15,10	15,10
15,10	15,10	15,10	15,10	15,10	15,10	15,10	15,10	15,10	15,12
15,17	15,17	15,17	15,17	15,17	15,17	15,20	15,20	15,20	15,20
15,20	15,20	15,20	15,20	15,20	15,20	15,20	15,20	15,20	15,20



… (faltan)

Tabla - medias de 500 muestras simuladas sobre la población antes definida (ahora sin los datos de cada muestra), cuyo tamaño muestral es de n= 30.

La media de todas las medias ($\mu_{\bar{x}}$ = 14, 96) se aproxima aún más a la media poblacional (μ = 15) y el error típico de la media ($\delta_{\bar{x}}$= 1,28) empírico tb se aproxima bastante al error típico teórico $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{7,07}{\sqrt{30}} = 1,29$

A medida q aumenta el nº de muestras, los valores de los parámetros de la distribución muestral de la media ($\mu_{\bar{x}}$ y $\delta_{\bar{x}}$) se aproximan + a la media poblacional (μ) y al error típico de la media antes definido $\left(\frac{\sigma}{\sqrt{n}}\right)$.

3. DISTRIBUCIÓN MUESTRAL DEL ESTADÍSTICO MEDIA

La **distribución muestral del estadístico media** se puede caracterizar mediante su valor esperado o media ($\mu_{\bar{x}}$), su error típico ($\delta_{\bar{x}}$) y la forma de la distribución. Para obtener esta inf., hay q tener en cuenta la distribución de la V de partida X en la población y si la varianza poblacional es conocida. 3 casos:

1. La variable X se distribuye según la curva normal con varianza conocida;
2. La variable X se distribuye de forma normal con varianza desconocida;
3. La variable X no sigue la distribución normal o no conocemos su distribución.

Distribución normal de la variable X con varianza poblacional conocida

Sea una variable aleatoria X con distribución normal y parámetros poblacionales μ y δ^2. Si se extrae un nº muy amplio de muestras de tamaño n, la distribución de sus medias tiende a una distribución normal definida como N (μ, $\delta/\sqrt{n}$) a medida que n tiende a infinito. Esto significa q:

$$\mu_{\bar{X}} = \mu$$

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

y q la distribución tb se ajusta a la normal. Para verificar q $\mu_{\bar{x}}$=μ y $\delta_{\bar{x}}$=$\frac{\sigma}{\sqrt{n}}$ se tendrán q recordar las propiedades de la media y varianza aplicándolas a la esperanza matemática (valor esperado) y varianza teórica.

La **esperanza matemática** de la VA media muestral será:

$$E(\bar{X}) = E\left(\frac{1}{n}\sum X_i\right) = \frac{1}{n}E(X_1 + X_2 + \ldots + X_n)\frac{1}{n}[E(X_1) + E(X_2) + \ldots E(X_n)] =$$

$$= \frac{1}{n}(\mu_1 + \mu_2 + \ldots + \mu_n) = \frac{1}{n}\sum \mu_i = \frac{1}{n} n\mu = \mu$$

$$\boxed{E(\bar{X}) = \mu}$$

La **varianza teórica** será:

$$\sigma^2_{\bar{X}} = \frac{1}{n^2}\sigma^2_{X_1 + X_2 + \ldots + X_n} = \frac{1}{n^2}\sigma^2_{X_1} + \frac{1}{n^2}\sigma^2_{X_2} + \ldots + \frac{1}{n^2}\sigma^2_{X_n} =$$

$$= \frac{1}{n^2}\sum \sigma^2_n = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$$

$$\sigma^2_{\bar{X}} = \frac{\sigma^2}{n}$$

Desviación típica o error típico:

$$\boxed{\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}}$$

Por tanto, la distribución muestral de la media X es: X ➔ N(μ,δ/$\sqrt{n}$) q se lee: Normal con media μ y error típico $\frac{\sigma}{\sqrt{n}}$
Consecuentemente, ya q se trata de una distribución normal, podremos tipificar la variable X; calcular su puntuación Z_x y la distribución sigue siendo normal.

$$Z_{\overline{X}} = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$ q, como se sabe, se distribuye N(0,1).

Supuestos:
- La variable X; se distribuye según la normal.
- Se conoce δ^2
- Se extraen muestras de tamaño n mediante muestreo aleatorio simple.

Entonces la distribución muestral de las medias forma una VA q se distribuye N(μ,δ/$\sqrt{n}$) ¿Q ventaja aporta saberlo? Q permite aplicar todo lo q se conoce de la distribución normal = hacer todos los cálculos relativos a las probabilidades correspondientes a los valores de la media, aplicando el concepto de tipificación.

A medida q ↑ el tamaño de la muestra, el error típico de la media disminuye. En la distribución muestral de la media el error típico es inversamente proporcional al tamaño de la muestra (n), por lo tanto, a medida q ↑*n* la distribución muestral de las medias se hace más homogénea (presenta menor variabilidad).

Distribución normal de la variable X con varianza poblacional desconocida
En el caso anterior se parte del hecho de q se conoce la varianza poblacional, lo cual no es muy frecuente. En estos casos, se deberá estimar el valor del error típico de la media mediante la cuasidesviación típica muestral dada por:

$$\sigma_{\overline{X}} = \frac{S_{n-1}}{\sqrt{n}}$$

En estas condiciones la distribución muestral de la media ya no se ajusta a la normal, sigue la distribución t de Student. Al no conocer la varianza poblacional, la transformación viene dada ya por la distribución:

$$T = \frac{\overline{X} - \mu}{S_{n-1}/\sqrt{n}}$$, q sigue una distribución t de Student con n - 1 gl.

Supuestos:
- La variable X_i se distribuye N(μ, δ)
- Se desconoce el valor de δ^2
- Extraemos muestras mediante muestreo aleatorio simple de tamaño *n*.

Entonces, la distribución muestral de las medias forma una VA q se distribuye $t\left(\mu, S_{n-1}/\sqrt{n}\right)$ con n-1 grados de libertad.

La variable X no se distribuye normalmente
Por lo general, las situaciones anteriores no son habituales. Lo que ocurre generalmente es q las V no se ajustan al modelo de la normal o se desconoce su varianza poblacional. En estas situaciones, la Estadística aporta el ***Teorema del Límite Central***, q permite calcular las probabilidades asociadas a los valores de las medias sin necesidad de conocer la forma de la distribución de las V, siempre q las muestras tengan tamaño suficiente (n ~ 30).

Sea $X_1, X_2,..., X_n$ un conjunto de VA, indep. e idénticamente distribuidas con media μ_i y varianza $\delta^2_i \neq 0$ Si n es suficientemente grande (n ≥ 30) la distribución muestral de la media de las X_i se aproxima a la distribución normal a medida q n aumenta, independientemente de las distribuciones q presenten $X_1, X_2,..., X_n$.

Este teorema establece q sí n es suficientemente grande, las V q se combinan son indep., tienen distribuciones idénticas y valor esperado y varianza finitas, entonces la distribución muestral del estadístico tiende a N (μ,δ/$\sqrt{n}$). En caso de no conocer δ, se utiliza como estimador la cuasidesvación típica con X ➔ N (μ , $S_{n-1}/\sqrt{n}$)

4. DISTRIBUCIÓN MUESTRAL DEL ESTADÍSTICO PROPORCIÓN

En Psicología y en Ciencias de la Salud son muy habituales estudios en los q están involucradas una o varias proporciones (o porcentajes) medidas en alguna V de interés *(proporción de universitarios varones frente a mujeres adictos a la cocaína)*. Para poder hacer los pertinentes estudios deberemos conocer la distribución muestral de este estadístico, distinguiendo entre muestras pequeñas y suficientemente grandes.

Sea una población en la q se mide una V q solo puede tomar 2 valores: éxito (1) o fracaso (0) (V q como sabemos se ajusta al modelo de Bernoulli). Definimos π como la proporción de aciertos en la población. Si extraemos todas las posibles muestras de tamaño *n* y medimos en cada una de ellas la variable aleatoria X= nº de éxitos en las n extracciones, y sea P = proporción de éxitos en las *n* extracciones, constante en todas las muestras, es decir, $P_1 = P_2 = ... = P_n = \pi$. Entonces, podemos definir las distribuciones muestrales de X y P según las muestras sean pequeñas o grandes.

Distribución muestral del estadístico P para muestras pequeñas

n < 30, la distribución muestral del estadístico X (nº de éxitos en *n* ensayos) es $X \to B\left(n\pi, \sqrt{n\pi(1-\pi)}\right)$

donde: $\mu_X = E(X) = n\pi$ y $\sigma_X = \sqrt{n\pi(1-\pi)}$. Dado q P es una mera transformación lineal de X(P=X/n), se demuestra q la distribución muestral del estadístico P (proporción de aciertos en n ensayos) es $P \to B\left(\pi, \sqrt{\frac{\pi(1-\pi)}{n}}\right)$ donde: $\mu_P = E(P) = \pi$ y $\sigma_P = \sqrt{\frac{\pi(1-\pi)}{n}}$

En efecto, siendo P = X/n, entonces: $E(P) = \frac{1}{n} n\pi = \pi$

** se ha cambiado la notación y, en lugar de llamar P a la probabilidad de éxito en la población, la denominamos n (letra griega correspondiente a P) .*

La varianza: $\sigma_p^2 = \frac{\pi(1-\pi)}{n}$ cuya raíz o **error típico de medida** es:

$$\sigma_P = \sigma\left(\frac{1}{n}X\right) = \frac{1}{n}\sigma_X = \frac{1}{n}\sqrt{n\pi(1-\pi)} = \sqrt{\frac{\pi(1-\pi)}{n}}$$

En la tabla de la binomial se pueden determinar las probabilidades para diferentes tamaños muestrales y valores de π.
Supuestos:

- La VA X es una variable Bernoulli (solo 2 valores éxito o fracaso).
- Se conoce π proporción en la población.
- Las *n* observaciones son independientes.

Entonces: La distribución muestral de la variable P es Binomial, definida por $B\left(\pi, \sqrt{\frac{\pi(1-\pi)}{n}}\right)$

Distribución muestral del estadístico P para muestras suficientemente grandes

Por regla general, las muestras suelen ser grandes. Por el Teorema del Límite Central se sabe q, a medida q *n* crece, la distribución de las proporciones se aproxima a la distribución normal con parámetros:

$$\mu_p = E(P) = \pi$$

$$\sigma_p^2 = \frac{\pi(1-\pi)}{n}$$

Aplicando la tipificación tendremos:

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

Supuestos:

- La VA X es una variable Bernoulli (solo dos valores éxito o fracaso).
- Se conocen π proporción en la población (o el valor P en la muestra)
- Las *n* observaciones son independientes.
- La muestra es suficientemente grande, es decir, n ≥ 30. Si n < 30 deberá cumplirse: n x π ≥ 5 y n x (1- π) ≥ 5

Entonces, los valores de las proporciones se aproximan a una normal definida por: $P \to N\left(\pi, \sqrt{\frac{\pi(1-\pi)}{n}}\right)$

5. DISTRIBUCIÓN MUESTRAL DEL ESTADÍSTICO VARIANZA

El razonamiento para obtener la distribución muestral de la varianza es el mismo q para la media y la proporción.
Si X es una variable que se distribuye en la población N(μ,δ), se extraen todas las muestras posibles de tamaño *n* con media X, varianza S^2_x y cuasivarianza S^2_{n-1}.

Entonces, las variables aleatorias $\frac{nS_X^2}{\sigma^2}$ y $\frac{(n-1)S_{n-1}^2}{\sigma^2}$ siguen una distribución x^2 con n-1 grados de libertad. Es decir, ambas V se distribuyen según X^2_{n-1}(n -1,), por lo q: μ =n- 1 y desviación típica, $\delta = \sqrt{2(n-1)}$

Conociendo la distribución se deduce q las distribuciones muestrales de la varianza y cuasivarianza son:

La **distribución muestral del estadístico varianza (S^2_x)**, sigue una distribución X^2_{n-1} con parámetros: $\mu_{S_x^2} = E\left(S^2\right) = \frac{n-1}{n}\sigma^2$

Y error típico: $\sigma_{S_x^2} = \sigma^2 \frac{\sqrt{2(n-1)}}{n}$

La **distribución muestral del estadístico cuasivarianza (S^2_{n-1})** - distribución X^2_{n-1} con parámetros: $\mu_{S_{n-1}^2} = E\left(S_{n-1}^2\right) = \sigma^2$

Y error típico: $\sigma_{S_{n-1}^2} = \sigma^2 \sqrt{\frac{2}{n-1}}$

En efecto, en una distribución x^2 con n-1 grados de libertad, la esperanza es $E(x^2_{n-1}) = n-1$ y la varianza igual a $V(X^2_{n-1}) = 2(n-1)$. Así podemos demostrar las fórmulas anteriores, recordando las propiedades de la media y la varianza.
En cuanto a la varianza muestral:

$$\frac{nS_X^2}{\sigma^2} = \chi^2_{n-1} \rightarrow S_X^2 = \frac{\sigma^2\chi^2_{n-1}}{n}$$

Calculamos la esperanza y la varianza:

$$E\left(S_X^2\right) = E\left(\frac{\sigma^2\chi^2_{n-1}}{n}\right) = \frac{\sigma^2}{n}E\left(\chi^2_{n-1}\right) = \frac{\sigma^2}{n}(n-1)$$

$$V\left(S_X^2\right) = V\left(\frac{\sigma^2\chi^2_{n-1}}{n}\right) = \frac{\sigma^4}{n^2}V\left(\chi^2_{n-1}\right) = \frac{\sigma^4}{n^2}2(n-1) \rightarrow \sigma_{S^2} = \sqrt{\frac{\sigma^4}{n^2}2(n-1)} =$$

$$= \sigma^2\frac{\sqrt{2(n-1)}}{n}$$

En cuanto a la cuasivarianza muestral:

$$\frac{(n-1)S^2_{n-1}}{\sigma^2} = \chi^2_{n-1} \rightarrow S^2_{n-1} = \frac{\sigma^2\chi^2_{n-1}}{n-1}$$

Calculamos la esperanza y la varianza:

$$E\left(S^2_{n-1}\right) = E\left(\frac{\sigma^2\chi^2_{n-1}}{n-1}\right) = \frac{\sigma^2}{n-1}E\left(\chi^2_{n-1}\right) = \frac{\sigma^2}{n-1}(n-1) = \sigma^2$$

$$V\left(S^2_{n-1}\right) = V\left(\frac{\sigma^2\chi^2_{n-1}}{n-1}\right) = \frac{\sigma^4}{(n-1)^2}V\left(\chi^2_{n-1}\right) = \frac{\sigma^4}{(n-1)^2}2(n-1) = \frac{2\sigma^4}{n-1} \rightarrow \sigma_{S^2_{n-1}} =$$

$$= \sqrt{\frac{2\sigma^4}{n-1}} = \sigma^2\sqrt{\frac{2}{n-1}}$$

Lo mostrado determina los valores de los parámetros de las distribuciones muestrales de la varianza y cuasivarianza. Para calcular las probabilidades asociadas a las varianzas, mediante la tabla de ji-cuadrado, se utilizan directamente las expresiones:

$$\chi^2 = \frac{nS_X^2}{\sigma^2} \quad \text{y} \quad \chi^2 = \frac{(n-1)S^2_{n-1}}{\sigma^2}$$

q se distribuyen $X^2_{n-1}(n-1, \sqrt{2(n-1)})$, siempre y cuando la variable X siga la distribución normal $N(\mu,\boldsymbol{\delta})$ en la población.

Aproximación a la normal de la distribución muestral de la varianza.
Cuando la muestra es suficientemente grande, n= 100, la distribución de las VA se aproxima a la normal con media μ= n-1 y desviación típica, $\boldsymbol{\delta} = \sqrt{2(n-1)}$

En este caso podemos tipificar las V, obteniendo:

$$\sigma = \sqrt{2(n-1)}.$$

Varianza:

$$Z = \frac{\left(\frac{nS_X^2}{\sigma^2}\right) - (n-1)}{\sqrt{2(n-1)}}$$

Cuasivarianza:

$$Z = \frac{\left(\frac{(n-1)S^2_{n-1}}{\sigma^2}\right) - (n-1)}{\sqrt{2(n-1)}}$$

Ambas con distribución aproximada N(0,1), q facilita la búsqueda de las probabilidades asociadas en la tabla de la Normal.

T10: ESTIMACIÓN DE PARÁMETROS Y CÁLCULO DEL TAMAÑO MUESTRAL

En el T9 se han estudiado el muestreo y el concepto de distribución muestraL de un estadístico; base de la Inferencia Estadística, q permite derivar a partir de los resultados de la muestra, los resultados q con una cierta probabilidad se pueden determinar para la población. Para llevar a cabo su proceso de análisis, la Inferencia combina los modelos de probabilidad con los estadísticos. La aplicación supone q el investigador se plantea preguntas con contenido teórico. Una vez formuladas las preguntas y evaluadas en relación con los conocimientos previos (a través del estudio y análisis crítico de la bibliografía), se traducen a términos estadísticos. Con estas formulaciones estadísticas se comprueba si la situación planteada se parece a algún modelo de los q nos ofrece la Estadística y la Probabilidad. Si es así, obtendremos una R estadística a la pregunta, q debe llenarse con el contenido teórico q suscitó la pregunta inicial.

Esta forma de trabajo exige garantizar q la realidad q se investiga y los modelos matemáticos (probabilístico y/o estadístico) q se aplican son similares. Es decir, q la realidad representada en el modelo teórico se ajusta al modelo matemático elegido para dar la R estadística. Esto significa q la comparación entre la pregunta estadística (expresión matemática del modelo teórico) y los modelos de probabilidad y/o estadísticos debe hacerse cumpliendo las exigencias matemáticas o supuestos referentes a :

- La métrica de las V o nivel de medida: cuantitativas, cuasi-cuantitativas, cualitativas y sus correspondientes escalas de medida, donde se pueden establecer det. relaciones matemáticas (igualdad en la escala nominal; orden de las posiciones en la escala ordinal; magnitud del intervalo, es decir, igualdad o desigualdad de diferencias en la escala de íntervalo; e igualdad o desigualdad de razones en la escala de razón). Estas operaciones matemáticas permiten cada una realizar unos det. cálculos.
- La forma de la distribución de la V: Bernoulli, Binomial, Nor- mal, Student…

Para entender mejor lo descrito anteriormente, se va a desarrollar el siguiente ejemplo.

La **Inferencia Estadística** tiene 2 ramas:

- Estadística **paramétrica**: la distribución de las V en la población es conocida (normal, binomial…), la muestra se selecciona por muestreo aleatorio simple y los datos están medidos al menos en escala de intervalos. Para cumplir con sus objetivos emplea 2 procedimientos -se basan en el conocimiento teórico de la distribución muestral del estadístico correspondiente al parámetro o parámetros q se quieren estimar-:
  - Estimación de parámetros: asignar un valor numérico o determinar un intervalo de valores numéricos al parámetro q deseamos conocer. Permite hacer conjeturas del tipo: *si en una muestra seleccionada al azar de una población en la q se ha medido una variable X_i, siendo la media, $\bar{X}$, ¿cuál será el valor más próximo o el intervalo de valores entre los q se encuentre el valor de media de la población μ con un cierto grado de confianza?*
  - Contraste de hipótesis: tiene como objetivo comprobar si un det. supuesto, referido a un parámetro o parámetros poblacionales, es compatible con la evidencia empírica q nos proporciona la muestra. Responde a preguntas del tipo: *¿es el valor de la media poblacional de una variable X; un valor determinado según el grado de confianza q consideramos suficiente?*

    * Ambos se basan en los mismos modelos probabilísticos y estadísticos. Sin embargo, en la estimación de parámetros se parte de los datos muestrales para responder a una pregunta sobre la población; en el contraste se hace una afirmación sobre la población q luego se contrasta.
- Estadística **no paramétrica**: la distribución de las V no se ajusta a ninguna distribución conocida o los datos están medidos en una escala inferior a la escala de intervalo.

1. ESTIMACIÓN DE PARÁMETROS

Básicamente, inferir el valor desconocido de un parámetro. 4 tipos de estimaciones:

- **Estimación puntual**. Procedimiento mediante el cual asignamos un único valor al parámetro desconocido, a partir del resultado obtenido en una muestra. *Tras la aplicación de un programa de intervención dental para niños, encontramos q el 60% (P =0,60) se lavan los dientes 3 veces al día. Una estimación puntual nos llevaría a suponer q la proporción π en la población de niños q se lavarían los dientes 3 veces al día si participasen en el programa sería: $\boldsymbol{\pi} = 0{,}60$.*
- **Estimación por intervalos**. Daremos un rango de posibles valores, dentro del cual estimamos se encuentra el verdadero valor del parámetro con un det. grado de confianza. *Podríamos afirmar q después de participar en el programa de intervención, la proporción se encuentra entre 0,50 y 0,70; $0{,}50 < \pi < 0{,}70$ con un cierto margen de confianza.*
- **Estimación Bayesiana**. En lugar de considerar a los parámetros como constantes, se presentan como VA con una cierta distribución a priori. Las observaciones o datos aportan inf. q transforman las probabilidades a priori en probabilidades a posteriori.
- **Estimación Bootstrap**. Se basan en el remuestreo y técnicas de simulación = requieren el uso de ordenadores. Extraer de una misma muestra varias (muchas) muestras y estudiar el conjunto de las obtenidas. Se puede asimilar a un muestro aleatorio simple con reposición q se realizase en una población de un tamaño pequeño. De cada muestra extraída se calcula el estadístico de interés y se estudia su distribución.

Antes de estudiarlas se plantean una serie de cuestiones sobre si un estadístico es un buen estimador de un parámetro o no, como: ¿sirve cualquier estadístico para estimar un parámetro? ¿cualquier estadístico es un buen estimador? *¿La media es un buen estimador de la media poblacional?, ¿La proporción muestral es un buen estimador de la proporción poblacional?, ¿La varianza muestral es un buen estimador de la varianza poblacional?*

Para q un estadístico pueda considerarse un buen estimador de un parámetro deberá; Carecer de sesgo, tener eficiencia, consistencia y suficiencia.

Propiedades de los estimadores

Carencia de sesgo

Sea θ el parámetro a estimar y $\hat{\theta}$ el valor del estimador (valor obtenido en la muestra), diremos q $\hat{\theta}$ es un **estimador insesgado** o carente de sesgo si E($\hat{\theta}$) = θ para cualquier valor de θ. Es decir, se cumple q la media de la distribución muestral (esperanza matemática de la distribución) coincide con el parámetro estimado.

$$E(\hat{\theta}) = \theta$$

Comprobación de si los principales estimadores (media, desviación típica, varianza y cuasivarianza) la cumplen:

- ¿Es la media un estimador insesgado de la media poblacional?

El estadístico media sigue una distribución normal o una t de Student (según conozcamos o no la varianza poblacional).

Conocida σ: $\bar{X} \to N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ No conocida σ: $\bar{X} \to t\left(\mu, \frac{S_{n-1}}{\sqrt{n}}\right)$

En cualquiera de las 2 distribuciones E(x̄) = μ. Por lo tanto, es un estimador insesgado de la media poblacional.

- ¿Es la proporción es un estimador insesgado de la proporción poblacional? Es decir, ¿E(P) = π?

La distribución muestral de la proporción se define como:

$$P \to B\left(\pi, \sqrt{\frac{\pi(1-\pi)}{n}}\right)$$

Como se observa, E(P) = π por lo q P es un estimador insesgado de *n*.

- ¿Son la varianza y la cuasivarianza estimadores insesgados de la varianza poblacional?

Según se desprende de su distribución muestral: $E\left(S_X^2\right) = \frac{n-1}{n}\sigma^2$

Es obvio q la esperanza matemática de la varianza de la muestra, S^2_x no es exactamente su correspondiente valor poblacional δ^2, por lo q decimos que se trata de un estimador sesgado, siendo precisamente su sesgo el factor n-1/n.

Por el contrario, la cuasivarianza sí es un estimador insesgado, $E(S^2_{n-1}) = \delta^2$. Si queremos realizar una estimación puntual de δ^2, es preferible utilizar la cuasivarianza.

Eficiencia

Dados dos estimadores $\hat{\theta}_1$ y $\hat{\theta}_2$ del mismo θ (parámetro), diremos q. $\hat{\theta}_1$ es más eficiente si su varianza (la de su distribución muestral) es menor. Es decir: $\sigma^2_{\hat{\theta}_1} < \sigma^2_{\hat{\theta}_2}$

Por tanto, entre 2 estimadores insesgados será preferible seleccionar el q presente una menor varianza (menor error típico de la distribución muestral del estadístico). El error típico refleja el mayor o menor alejamiento de los posibles valores del estadístico a su esperanza matemática (media de la distribución muestral). Un estimador es tanto mejor cuanto su distribución muestral esté más concentrada = varianza más pequeña.

Por ej, en relación con la varianza y la cuasivarianza, los errores típicos de ambos son: $\sigma_{S_i^2} = \sigma^2\frac{\sqrt{2(n-1)}}{n}$ y $\sigma_{S^2_{n-1}} = \sigma^2\sqrt{\frac{2}{n-1}}$
El de la cuasivarianza es mayor, por lo q diremos q la varianza es un estimador + eficiente q la cuasivarianza.

La eficiencia de un estimador siempre es relativa, ninguno puede ser perfectamente eficiente, dado q el error típico acompaña a cualquier distribución muestral. Se define la **eficiencia relativa, ER**, de un estimador $\hat{\theta}_1$ con respecto a otro $\hat{\theta}_2$, como la razón:

$$ER = \frac{\sigma_{\hat{\theta}_1}}{\sigma_{\hat{\theta}_2}}$$

- Cociente =1, ambos estimadores son igualmente eficientes.
- ER > 1, el estimador del denominador es más eficiente.
- ER < 1, el estimador del numerador es más eficiente.

Consistencia

$\lim_{n\to\infty} P\left(|\hat{\theta} - \theta| < \delta\right) = 1$

Indica q a medida q el tamaño muestral se hace grande (q tiende a infinito), el valor del estadístico se aproxima al valor del parámetro. Por tanto, un estimador es consistente cuando la probabilidad de q su valor se acerque al del parámetro es mayor a medida q aumenta el tamaño de la muestra. En otras palabras, si *n* tiende a infinito, la probabilidad de q $|\hat{\theta} - \theta|$ sea menor q cualquier valor **δ**, por pequeño q sea éste, tiende a 1.

Suficiencia

Se refiere a la capacidad del estimador de utilizar toda la inf. existente en la muestra en relación al parámetro; q el estimador emplee todos los valores de los datos (inf.) para estimar el parámetro.
Sabemos q la media de la muestra, ~~X~~, es un buen estimador de la media poblacional, μ. Tb podríamos utilizar otros estimadores como: la mediana, el promedio de los valores extremos de la distribución, la media de los cuartiles primero y tercero… Sin embargo, basta con observar las fórmulas para darse cuenta de q la media es un estimador suficiente.

Así, si utilizamos el promedio de los extremos de la distribución sería: $\frac{X_{i_{esp}} + X_{i_{inf}}}{2}$ sólo empleamos en el cálculo el valor de la puntuación más alta y + baja. Mientras q en la media $\overline{X} = \frac{\sum X_i}{n}$ es evidente q empleamos todos los valores de X_i. Así pues, el estimador suficiente de μ es ~~X~~.

Del mismo modo, la varianza y la varianza insesgada, así como la proporción son estimadores suficientes de δ^2 y π, respectivamente.

Como resumen, las **propiedades de los estimadores media**, **proporción** y **cuasivarianza** son:

1. La media muestral se considera buen estimador de la media poblacional (~~X~~ = μ). Cumple las propiedades de ser insesgado, consistente y suficiente.
2. La proporción muestral (*P*) se considera buen estimador de la proporción poblacional (P=*π*), ya q cumple las propiedades de ser insesgado, consistente y suficiente.
3. La varianza muestral (S^2_x) cumple las propiedades de ser consistente y suficiente pero no es insesgado. Por esta razón, la cuasivarianza se considera un buen estimador de la varianza poblacional $\left(\tilde{\sigma}^2 = S^2_{n-1}\right)$. Cumple las propiedades de ser insesgada, consistente y suficiente, aunq su eficiencia (q compara dos estimadores) es menor en relación con la varianza muestral.

Métodos de obtención de estimadores

Una vez ya se sabe q propiedades han de cumplir los estadísticos para ser empleados como estimadores, se estudia cómo se estima el parámetro = cómo se determina su valor o los valores entre los q se encuentra.

Existen varios métodos q garantizan las propiedades antes enunciadas y se aplican en circunstancias específicas; 2 los + empleados para los q habitualmente utilizamos en Psicología;

- **Método de mínimos cuadrados**: trata de obtener aquel estimador q minimice las distancias (al cuadrado) entre el valor estimado del parámetro y los resultados muestrales observados. No siempre es el mejor pero resulta muy útil para estimar los parámetros de la regresión, por ej.

$$\sum\left(X_i - \hat{\theta}\right)^2 \text{ sea mínimo, donde } i = 1, 2, \ldots, n$$

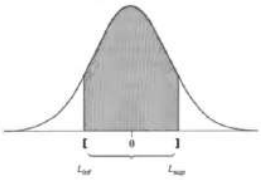
- **Método de máxima verosimilitud**: obtiene como estimador de un parámetro aquel valor del estadístico q hace lo más verosímil posible la muestra obtenida; se trata de elegir, de entre todos los posibles valores del parámetro, aquel q maximice la probabilidad de obtener el resultado particular observado en la muestra.

Estimación puntual

Independientemente del método escogido y partiendo de q el estimador seleccionado cumple las propiedades, la estimación puntual consiste en dar un valor numérico único al parámetro desconocido; utilizar el valor del estadístico para estimar el parámetro.
No deja de tener inconvenientes. Teniendo en cuenta el elevado nº de muestras q podemos extraer de la población y q de cada una de ellas podemos realizar una estimación, el nº de estimaciones podría resultar excesivo.
Además, aun cuando la muestra sea representativa, no se puede establecer ni la fiabilidad de la estimación ni el error q se comete. En estos casos, lo único q podemos afirmar es q el error cometido en la estimación se hará menor a medida q aumente el tamaño de la muestra. En definitiva, la estimación puntual no siempre es la más aconsejable y útil.

Estimación por intervalos

Obtener una medida del error (diferencia entre el estimador y el parámetro) q se comete al realizar la estimación con una det. probabilidad; es atribuir al parámetro un rango de valores posibles dentro del cual estará incluido el parámetro con una det. probabilidad.
Mediante la estimación por intervalos, en lugar de un solo valor (como en la estimación puntual), obtenemos un rango de posibles valores del parámetro, **intervalo de confianza** y sus límites, **límites del intervalo de confianza**.



Los corchetes indican los límites del intervalo y la llave los posibles valores estimados del parámetro θ. A la zona sombreada se le denomina **nivel de confianza (n.c.)** = probabilidad asociada al intervalo q contiene todos los posibles valores q puede tomar el parámetro θ. Se le llama de confianza y no de probabilidad ya q una vez extraída la muestra, contendrá al verdadero valor del parámetro o no. Lo q sabemos es q si repitiésemos el proceso con muchas muestras podríamos afirmar q el (1-α)% de los intervalos así construidos contendría al verdadero valor del parámetro. α es el nivel de significación y hace referencia a la cuantía del margen de error q se asume a priori.

Principal ventaja; se puede valorar la seguridad con la q se realizan las estimaciones mediante el nivel de confianza, el cual se expresa en términos de probabilidad.

Algunas carac.:

- **Relación entre amplitud del intervalo y nivel de confianza**

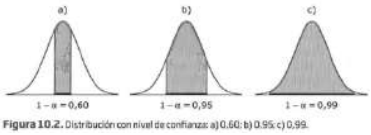


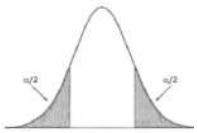
Figura 10.2. Distribución con nivel de confianza: a) 0,60; b) 0,95; c) 0,99.

<— 3 distribuciones en las q se han establecido 3 intervalos de confianza (zona sombreada) q van aumentado a medida q aumenta 1-α. Cuanto mayor es el intervalo de valores mayor es la probabilidad de q se encuentre dentro de él el verdadero valor de θ y la estimación es menos precisa = la precisión de la estimación se relaciona de forma inversa con el nivel de confianza, a mayor confianza, mayor intervalo (el rango de posibles valores del parámetro) = estimación - precisa. A mayor confianza menor precisión.

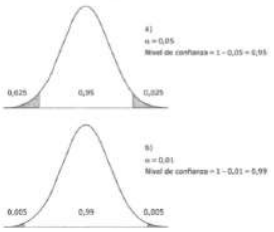
- **Fijación del nivel de confianza**

El investigador es quien decide y fija el valor del nivel de confianza en función de la valoración personal q hace sobre diversos aspectos: diseño de su trabajo, definición y obtención de la muestra, recogida de inf… Por convenio, en general se adoptan los niveles del confianza de 1-α = 0,95 ó 1-α = 0,99.

- **Nivel de riesgo o significación α**



El opuesto al nivel de confianza = **nivel de riesgo**, **margen de error** o **nivel de significación** y se representa por α. Indica la probabilidad de q el valor del parámetro no se encuentre dentro de los límites. α se reparte entre los 2 extremos de la curva q delimitan el intervalo de confianza (1-α). Es decir, q el margen de error se divide en 2 partes iguales siendo el área correspondiente a cada una α/2 (zonas sombreadas). Entre el nivel de confianza (1-α) y el de riesgo o significación a existe una relación inversa.



2. CÁLCULO DEL INTERVALO DE CONFIANZA

Para calcular el intervalo de confianza de un parámetro cualquiera es necesario conocer la distribución muestral del estadístico correspondiente y los parámetros q la definen. Es decir, la esperanza matemática y el error típico.

En general, para construir un intervalo de confianza para un parámetro se suma y se resta al estimador, θ el error máximo de estimación, E_{max}. Para comprender bien como se define el intervalo de confianza se retoma el concepto de puntuación Z (T3) ya q, junto al error de típico (T9) son los elementos necesarios para entender cómo se construyen los intervalos de confianza y su significado .

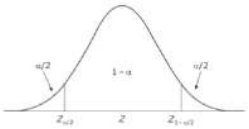
Intervalo de confianza para el parámetro μ con δ^2 conocida

Si la VA X sigue una distribución normal en la población, y se conoce la varianza poblacional, la distribución muestral de la media $\overline{X}$ es $\overline{X} \rightarrow N\left(\mu; \frac{\sigma}{\sqrt{n}}\right)$, siendo $E(\overline{X}) = \mu$ y $\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$.

Ya q la distribución muestral de la media es normal, se tipifica sin más que aplicar la transformación a Z:

$$Z = \frac{\overline{X} - E(\overline{X})}{\sigma_{\overline{X}}} = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$ donde Z es N(0,1).

Acudiendo a las tablas de la distribución normal se puede calcular la probabilidad de q la variable Z se encuentre entre 2 valores concretos. Si a se corresponde al margen de error (o nivel de significación q se ha fijado) tendremos la representación —>



$$P\left(Z_{\alpha/2} \le Z \le Z_{1-\alpha/2}\right) = 1 - \alpha$$

Sustituyendo Z por su valor:

$$P\left(Z_{\alpha/2} \le \frac{\overline{X} - \mu}{\sigma / \sqrt{n}} \le Z_{1-\alpha/2}\right) = 1 - \alpha$$

A partir de esta expresión, el objetivo es determinar los intervalos para μ: $P\left(Z_{\alpha/2}\,\sigma/\sqrt{n} \le \overline{X} - \mu \le Z_{1-\alpha/2}\,\sigma/\sqrt{n}\right) = 1 - \alpha$

Restando la media X en todos los términos: $P\left(-\overline{X} + Z_{\alpha/2}\,\sigma/\sqrt{n} \le -\mu \le -\overline{X} + Z_{1-\alpha/2}\,\sigma/\sqrt{n}\right) = 1 - \alpha$

multiplicando por -1: $P\left(\overline{X} - Z_{\alpha/2}\,\sigma/\sqrt{n} \le \mu \le \overline{X} - Z_{1-\alpha/2}\,\sigma/\sqrt{n}\right) = 1 - \alpha$

teniendo en cuenta q $Z_{\alpha/2} = -Z_{1-\alpha/2}$:

$$P\left(\overline{X} - |Z_{\alpha/2}|\,\sigma/\sqrt{n} \le \mu \le \overline{X} + |Z_{\alpha/2}|\,\sigma/\sqrt{n}\right) = 1 - \alpha$$

donde los límites del intervalo son: $L_{inf} = \overline{X} - |Z_{\alpha/2}| \sigma/\sqrt{n}$ $L_{inf} = P - |Z_{\alpha/2}| \sqrt{\frac{P(1-P)}{n}}$

La **semi-amplitud** (mitad de la amplitud) del intervalo de confianza se denomina **Error máximo de estimación**, siendo su valor $E_{max} = |Z_{\alpha/2}| \sigma/\sqrt{n}$. Indica q el investigador asume con un nivel de confianza del (1-α)% q la diferencia máxima entre el valor estimado a partir de la muestra y el valor real del parámetro es igual a:

$$E_{max} = |Z_{\alpha/2}| \frac{\sigma}{\sqrt{n}}$$

La obtención de los intervalos correspondientes a la media, conocida la varianza poblacional y siendo normal la distribución de la V es aplicable tb al caso en el q la V no siga la distribución normal siempre y cuando el tamaño de la muestra sea grande ($n > 30$). Como se vio en el T9, el Teorema del Límite Central establece q la distribución muestral de la media se aproxima a la distribución normal a medida q el tamaño de la muestra va aumentando (en la práctica con n > 30 el ajuste es bastante bueno) sin q necesariamente la VA X tenga una distribución normal. De este modo, cuando el tamaño muestral sea grande y queramos estimar la media poblacional, podemos utilizar el intervalo antes definido.

Intervalo de confianza para el parámetroμ con δ^2 desconocida

Si la VA X tiene distribución normal en la población, pero la varianza es desconocida, sabemos del T9 q la distribución muestral de la media X sigue la distribución t de Student definida por:

$$\overline{X} \to t\left(\mu; \frac{S_{n-1}}{\sqrt{n}}\right) \text{ con } n-1 \text{ grados de libertad.}$$

Esto implica q. $T = \frac{\overline{X} - \mu}{S_{n-1}/\sqrt{n}}$, tb sigue una distribución t con n-1 g.l.

El intervalo de confianza para T sería: $P\left(t_{n-1;\alpha/2} \le T \le t_{n-1;1-\alpha/2}\right) = 1 - \alpha$

y sustituyendo en *T*: $P\left(t_{n-1;\alpha/2} \le \frac{\overline{X} - \mu}{\frac{S_{n-1}}{\sqrt{n}}} \le t_{n-1;1-\alpha/2}\right) = 1 - \alpha$

Siguiendo el mismo razonamiento q en el caso anterior (conocida) se tiene:

$$P\left(\overline{X} - |t_{n-1;\alpha/2}| \frac{S_{n-1}}{\sqrt{n}} \le \mu \le \overline{X} + |t_{n-1;\alpha/2}| \frac{S_{n-1}}{\sqrt{n}}\right) = 1 - \alpha$$

Los límites del intervalo de confianza son:

$L_{inf} = \overline{X} - |t_{n-1;\alpha/2}| \frac{S_{n-1}}{\sqrt{n}}$ $L_{sup} = \overline{X} + |t_{n-1;\alpha/2}| \frac{S_{n-1}}{\sqrt{n}}$

En caso de q la variable X tenga una distribución desconocida y el tamaño muestral sea n2 30, según el Teorema del Límite Central, la distribución muestral de la media es normal $\overline{X} \to Z\left(\mu; \frac{S_{n-1}}{\sqrt{n}}\right)$

Por tanto, los límites de los intervalos en esta aproximación a la normal:

$L_{inf} = \overline{X} - |Z_{\alpha/2}| \frac{S_{n-1}}{\sqrt{n}}$ $L_{sup} = \overline{X} + |Z_{\alpha/2}| \frac{S_{n-1}}{\sqrt{n}}$

Intervalo de confianza para el parámetro *π* (aproximación a la normal)

La distribución muestral del estadístico P cuando se cumple q la muestra es grande ($n \ge 30$; o $nP \ge 5$ y $n(1-P) \ge 5$) es $N(0,1)$ aprox., por lo q $Z = \frac{P - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}}$. A partir de la tipificación de *P* podemos construir el **intervalo de confianza** de la siguiente manera:

$$P\left(Z_{\alpha/2} \le \frac{P - \pi}{\sqrt{\frac{P(1-P)}{n}}} \le Z_{1-\alpha/2}\right) = 1 - \alpha$$

Aplicando el mismo razonamiento q en los demás estadísticos y, haciendo las correspondientes transformaciones, se tiene q la probabilidad de obtener un intervalo de confianza q contenga el parámetro es:

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

Siendo sus correspondientes límites:

$$L_{inf} = P - |Z_{\alpha/2}|\sqrt{\frac{P(1-P)}{n}} \qquad L_{sup} = P + |Z_{\alpha/2}|\sqrt{\frac{P(1-P)}{n}}$$

Intervalo de confianza para el parámetro δ^2

Al presentar la estimación puntual en el apartado (T9) se vio q la varianza S^2_x es un estimador sesgado de δ^2, siendo la cuasivarianza el estimador insesgado. Por ello vamos a utilizar la cuasivarianza muestral S^2_{n-1} como estimador de δ^2.

Además, sabemos q $\frac{(n-1)S^2_{n-1}}{\sigma^2}$ se distrubuye X^2_{n-1}

Podemos construir el intervalo de confianza con un nivel de confianza de 1-α: $P\left(\chi^2_{n-1,\alpha/2} \le \frac{(n-1)S^2_{n-1}}{\sigma^2} \le \chi^2_{n-1,1-\alpha/2}\right)$

dividiendo por (n-1) S^2_{n-1} los términos de la desigualdad: $P\left(\frac{\chi^2_{n-1,\alpha/2}}{(n-1)S^2_{n-1}} \le \frac{1}{\sigma^2} \le \frac{\chi^2_{n-1,1-\alpha/2}}{(n-1)S^2_{n-1}}\right) = 1-\alpha$

Por tanto, el intervalo de confianza para la varianza queda definido por la expresión:

$$P\left(\frac{(n-1)S^2_{n-1}}{\chi^2_{n-1,\alpha/2}} \ge \sigma^2 \ge \frac{(n-1)S^2_{n-1}}{\chi^2_{n-1,1-\alpha/2}}\right) = 1-\alpha$$

Sus límites inferior y superior son:

$$L_{inf} = \frac{(n-1)S^2_{n-1}}{\chi^2_{n-1,1-\alpha/2}} \qquad L_{sup} = \frac{(n-1)S^2_{n-1}}{\chi^2_{n-1,\alpha/2}}$$

El límite inferior se refiere a 1-(α/2) ya q se trata de una desigualdad dada por el límite superior del parámetro *π (apartado anterior)* q nos indica q $\sigma^2 \ge \frac{(n-1)S^2_{n-1}}{\chi^2_{n-1,1-\alpha/2}}$ o, lo q es lo mismo $\frac{(n-1)S^2_{n-1}}{\chi^2_{n-1,1-\alpha/2}} \le \sigma^2$, es decir, el límite inferior.

Cuando las muestras son grandes ($n > 100$), la distribución muestral de la varianza insesgada se puede aproximar a la normal. $N\left(\sigma^2; S^2_{n-1}\sqrt{\frac{2}{n}}\right)$. Por lo tanto, se puede construir el intervalo de confianza para la varianza definido como:

$$P\left(S^2_{n-1} - |Z_{\alpha/2}|S^2_{n-1}\sqrt{\frac{2}{n}} \le \sigma^2 \le S^2_{n-1} + |Z_{\alpha/2}|S^2_{n-1}\sqrt{\frac{2}{n}}\right) = 1-\alpha$$

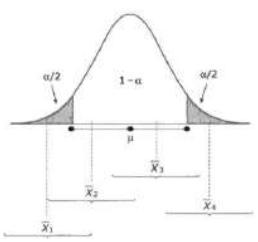
Los límites del intervalo son:

$$L_{inf} = S^2_{n-1} - |Z_{\alpha/2}|S^2_{n-1}\sqrt{\frac{2}{n}} \qquad L_{sup} = S^2_{n-1} + |Z_{\alpha/2}|S^2_{n-1}\sqrt{\frac{2}{n}}$$

3. SIGNIFICADO DEL NIVEL DE CONFIANZA

Hasta ahora; probabilidad de q el parámetro desconocido se encuentre entre los límites del intervalo. Sin embargo, no es del todo correcto, el concepto de probabilidad solo es aplicable a V y los valores de los límites del intervalo una vez calculados son valores constantes (no son variables).

Para interpretar correctamente el nivel de confianza asociado al intervalo se debe pensar: si se extraen un nº elevado de muestras (todas del mismo tamaño) y calculamos la media en cada una, obtendremos tantos intervalos de confianza como medias hayamos calculado. Pues bien, el 95% de todos los intervalos calculados tienen dentro al parámetro y el 5% no; de cada 100 intervalos, cabe esperar q 95 capten el valor del parámetro (intervalos correctos) y 5 no lo capten (incorrectos). Por tanto, una proporción de 1-α , de todos los intervalos de confianza contendrá al parámetro poblacional y una proporción α no los contendrá.



Los intervalos ~~X_1~~ y ~~X_4~~ no cubren el valor del parámetro; X_2 y X_3 sí. Las medias X_1 y X_4 están dentro de la zona sombreada y el intervalo de confianza NO contiene a la media poblacional. Sin embargo, X_2 y X_3 están en la zona no sombreada y contienen al parámetro. Por tanto, cualquier valor de la media en las zonas sombreadas da lugar a intervalos q NO contienen al parámetro, siendo la probabilidad de q esto ocurra de 0,05 (0,025 + 0,025). Por el contrario, el valor de la media en la zona no sombreada contendrá al parámetro y la probabilidad es de 0,95.
Se habla de probabilidad cuando se hace alusión a la V media, por eso al referirnos al intervalo hablaremos de confianza y no de probabilidad.

4. GENERALIZACIÓN DE LA CONSTRUCCIÓN DE INTERVALOS

Se puede generalizar el procedimiento de construcción de intervalos representados por:

$$IC = [\text{Estimador} \pm \text{Error máximo de estimación}] = [\hat{\theta} \pm E_{max}]$$

Para **estimadores con distribución muestral conocida**, los pasos para construir el intervalo son:

1. Determinar el parámetro q queremos estimar y el estadístico (estimador) q, cumpliendo con las propiedades q debe tener un buen estimador, lo estima.
2. Conocer la distribución muestral del estadístico (estimador) y los parámetros q la definen (media y error típico). La distribución muestral nos da las probabilidades asociadas a cada uno de los valores.
3. Fijar el nivel de significación α (el investigador por valoración personal de la seguridad de sus datos y del empleado por otros investigadores en situaciones similares) o el nivel de confianza 1-α. Suele ser α= 0,05 o 0,01.
4. Determinar el error máximo de estimación (E_{max}) definido por el producto del error típico de la distribución muestral del estadístico q estima al parámetro por el valor del estadístico (Z, T, F…) correspondiente al nivel de significación prefijado.

5. FACTORES QUE AFECTAN AL INTERVALO DE CONFIANZA

- Nivel de confianza: la mayor o menor amplitud (tb precisión) depende fundamentalmente del nivel de confianza con el q se decide trabajar. Con 1- α= 0,95 tendremos intervalos menos amplios (o más precisos) q =0,99.
- Error típico: no es más q una medida de la variabilidad de la distribución muestral del estadístico = depende del tamaño muestral *n* y de la homogeneidad de la muestra, afectando ambos factores al intervalo de confianza.
- Tamaño muestral: Inversamente proporcional al error típico; a mayor *n,* menor error típico, menor amplitud del intervalo y mayor precisión.
- Homogeneidad de la muestra; Sí lo son = la varianza es pequeña = tb la desviación típica (poblacional o muestral), = error típico será pequeño = amplitud del intervalo menor = precisión mayor.

6. CÁLCULO DEL TAMAÑO MUESTRAL

Todo estudio en el q se trabaja con muestras, además de garantizar la representatividad es imprescindible determinar el tamaño q ha de tener; suficiente para garantizar la precisión deseada y/o detectar de forma correcta diferencias entre los grupos en el caso de q existiesen, valorar la intensidad de la relación…

La Estadística Paramétrica (dentro de la Inferencia Estadística) tiene 2 procedimientos de trabajo: la estimación de parámetros y el contraste de hipótesis. En ambos se parte de los datos muestrales. Sin embargo, en la 1ª se responde a una pregunta sobre la población; la 2ª hace una afirmación sobre la población q luego se comprueba. Por tanto, hay 2 situaciones q se deben considerar a la hora de det. el tamaño muestral.

En la **estimación de parámetros**, los factores q influyen en la determinación del tamaño muestral son:

- El parámetro q se va a estimar.
- El error máximo (E_{max}) q el investigador está dispuesto a admitir.
- El nivel de confianza (1-α) con el q se trabaja.
- La precisión q se desea para el estudio.
- La variabilidad q presenta la población en relación a la V en estudio.

En el **contraste de hipótesis** los factores son:

- El error tipo I (α) y error tipo II (β) y la potencia estadística.
- Magnitud de la diferencia (o tamaño del efecto).
- Direccionalidad de la hipótesis.
- Variabilidad de la población respecto a la V en estudio.

En ambos casos, el tamaño muestral debe ser un nº entero y, cuando su cálculo de lugar a un nº decimal debe redondearse siempre al inmediato superior.

Tamaño muestral para el parámetro media

La determinación del tamaño muestral gira en torno a los conceptos de error típico y error máximo de estimación. Al igual q en la determinación del intervalo de confianza, hay 2 posibles situaciones: conocer o desconocer la varianza poblacional.

Conocida la varianza poblacional

El error máximo de estimación viene dado por $E_{max} = \frac{|Z_{\alpha/2}|\sigma}{\sqrt{n}}$. Si elevamos al cuadrado y despejamos *n* tendremos:

$$n = \left(\frac{|Z_{\alpha/2}|\sigma}{E_{max}}\right)^2 = \frac{\sigma^2 Z^2_{\alpha/2}}{E^2_{max}}$$

En el caso de poblaciones finitas y muestreo sin reposición debe multiplicarse por el factor de corrección: $\sqrt{\frac{N-n}{N-1}}$ donde N es el tamaño de la población. Entonces:

$$E_{max} = \frac{|Z_{\alpha/2}|\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

Y, por tanto:

$$n = \frac{\sigma^2 Z^2_{\alpha/2} N}{E^2_{max}(N-1) + \sigma^2 Z^2_{\alpha/2}}$$

Desconocida la varianza poblacional

Lo habitual en la investigación es q no se conozcan la media y varianza poblacional. En estos casos, la varianza tb debe ser estimada al mismo tiempo q la media, mediante su estimador insesgado, la cuasivarianza.

La distribución muestral del estadístico media (el q queremos estimar) se ajusta a una distribución t de Student con n-1 grados de libertad. En la varianza, el error máximo de estimación, E_{max} vendrá dado por: siendo $t_{n-1;\alpha/2}$ el valor de la distribución t de Student con n-1 grados de libertad .

$$E_{max} = |t_{n-1;\alpha/2}|\frac{S_{n-1}}{\sqrt{n}}$$

$E^2_{max} = \frac{t^2_{n-1;\alpha/2} S^2_{n-1}}{n}$, y despejando *n*:

$$n = \frac{t^2_{n-1;\alpha/2} S^2_{n-1}}{E^2_{max}}$$

En el caso de poblaciones finitas o muestreo sin reposición habrá q multiplicar por el factor corrector siendo, en consecuencia, el valor n:

$$n = \frac{t^2_{n-1;\alpha/2} S^2_{n-1} N}{E^2_{max}(N-1) + t^2_{n-1;\alpha/2} S^2_{n-1}}$$

Como se ha dicho, $t_{1-\alpha/2;n-1}$ sería el valor de t de Student en la Tabla VI (Formulario) con n-1 grados de libertad para la probabilidad (1- α) especificada.

si no se conoce todavía *n*, ¿cómo buscar en las tablas t de Student con n-1 grados de libertad? 2 soluciones posibles:

A. Trabajar mediante aproximaciones sucesivas por un procedimiento iterativo;
B. Aproximar la distribución t de Student mediante la curva normal.

Otra cuestión es q, si aún no hemos obtenido la muestra (de hecho estamos determinando su tamaño), ¿Cómo podemos saber cuál es la varianza insesgada en nuestra muestra? Existen dif. posibilidades pero el procedimiento + cómodo y efectivo es obtener un valor aprox. para la desviación típica insesgada (S_{n-1}) . A partir de estudios previos, o de la realización de uno piloto, partiendo de este valor de S_{n-1} se calcula el tamaño muestral (*n*) y se procede a la selección de la muestra, medida de las V…

Tamaño muestral para el parámetro proporción

Como ya se sabe la distribución muestral de P es:

$$P \to N\left(\pi; \sqrt{\frac{\pi(1-\pi)}{n}}\right)$$

Cuando n > 30 el error máximo es:

$$E_{max} = Z_{\alpha/2}\sqrt{\frac{P(1-P)}{n}}$$

Por tanto:

$$n = \frac{Z^2_{\alpha/2} P(1-P)}{E^2_{max}}$$

Para la determinación del error típico de *P* surge ahora un problema añadido a los comentados en la estimación de la media, puesto que *p* y *q* = (1-p) dependen directamente de *P*, y es precisamente este parámetro el q hay q estimar. Lo q se suele hacer es suponer q la varianza de la distribución muestral es máxima (p=q=1-p=0,5), con lo q la muestra será casi con toda seguridad superior a lo estrictamente necesario pero, por contra, no habrá q hacer suposiciones arriesgadas sobre el valor de *p*.

Si la población es finita, o el muestreo es sin reposición, habrá q corregir el tamaño muestral multiplicado la expresión por el factor de corrección: $\sqrt{\frac{N-n}{N-1}}$

$$n = \frac{Z^2_{\alpha/2} P(1-P)}{E^2_{max}}\sqrt{\frac{N-n}{N-1}} = \frac{Z^2_{\alpha/2} P(1-P) N}{E^2_{max}(N-1) + Z^2_{\alpha/2} P(1-P)}$$