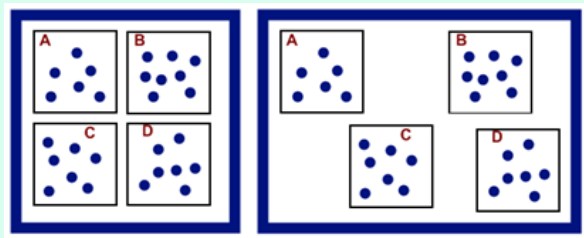
Temas de Estadística práctica elemental



Colección Hojamat.es

http://www.hojamat.es

PRESENTACIÓN

Esta publicación recoge el contenido de nuestro Curso elemental de Estadística práctica.

(http://www.hojamat.es/estadistica/iniestad.htm)

Hemos deseado ofrecer juntos los conceptos básicos del mismo, sin la distracción que supone todo el conjunto de enlaces, prácticas y ejercicios que contiene. Es una alternativa en modo texto, lo que no impide la realización de prácticas y ejemplos al final de cada tema, acudiendo a la web hojamat.es.

El nivel de esta publicación es de tipo medio, el equivalente en España de las asignaturas de Estadística en estudios de Ciencias Sociales. Por eso no contiene demostraciones matemáticas, que complicarían el aprendizaje.

Cada tema comienza con una cuestión-ejemplo. Algunas contienen enlaces a hojas de cálculo, que si no se desean activar, con las imágenes insertadas permitirán seguir la explicación. Se incluyen porque en pocos párrafos introducen los conceptos con facilidad.

CONTENIDO

MÉTODO ESTADÍSTICO

La Estadística es odiada cordialmente en parte por la pesadez de sus cálculos. Disponer de un instrumento como la Hoja de Cálculo permite poder insistir en los conceptos más que en los aspectos numéricos. El uso de estos modelos en las clases puede organizarse de forma que su confección sea simultánea con el uso y el aprendizaje de los temas. Por ejemplo, para estudiar los datos de tipo cuantitativo los alumnos pueden ir construyendo las tablas de frecuencias absolutas y relativas de forma simultánea a la explicación de los conceptos. Hemos experimentado con éxito esta modalidad en grupos con pocos alumnos y alto grado de diversidad.

Por otra parte, a veces sólo se desea aplicar técnicas estadísticas aunque la teoría haya quedado algo oscura para el alumnado, porque se considere útil usar dichas técnicas si se conoce su sentido dentro de los trabajos

de investigación. Por ejemplo, en cursos elementales se puede usar la desviación típica como medida de la dispersión aunque no se conozca su fórmula.

OBJETIVO

Confección de estudios de tipo estadístico

Por experiencia propia, me atrevo a afirmar que la única forma de aprender la Estadística Elemental es mediante la realización de estudios, en sus fases de:

- Recogida de datos
- Tabulación y graficación
- Análisis de datos
- Descripción e inferencias

Por esta razón, el saber estudiar estadísticamente una situación constituirá el principal objetivo de estas páginas.

MÉTODO ESTADÍSTICO

Problemas que resuelve la Estadística

La ciencia de la Estadística, aunque en sus inicios sólo pretendió recoger datos y presentarlos, en nuestros tiempos ha extendido mucho sus aplicaciones, de tal modo que muchos la consideran como la Tecnología del método científico. Hoy en día no se puede emprender un estudio serio en cualquier área del conocimiento sin poseer conocimientos estadísticos.

DESARROLLO DE UN ESTUDIO

Tipo de cuestiones que podemos tratar con métodos estadísticos

- Deben tener parte aleatoria y parte explicable.
- Es preferible que se basen en datos recogidos en nuestro entorno o procedentes de Anuarios o Internet
- No deben ser demasiado numerosos

CUESTIÓN . EJEMPLO

¿Cómo influyen las horas de estudio en el rendimiento de los alumnos?

Imaginemos que estamos interesados en estudiar cómo influyen las (escasas) horas de estudio en el rendimiento de los alumnos y alumnas. Para organizar nuestro estudio es aconsejable seguir los siguientes pasos:

Planteo del problema

Un estudio estadístico debe comenzar con una definición lo más concreta posible de los elementos del mismo:

Población y/o muestra: Se debe elegir con toda claridad sobre qué alumnado se efectuará el estudio: Edad, centro de enseñanza, grupos, etc.

Variables que se estudiarán: Hay que pensar qué entendemos por rendimiento: Un test, pruebas, exámenes en alguna asignatura...

Variables explicativas: ¿Cómo medimos las horas de estudio? ¿A quién preguntamos? ¿Cómo eliminar datos sospechosos? ¿Introducimos la variable Capacidad para el estudio?

Uso de un modelo

Este paso sólo lo incluiremos si deseamos sacar consecuencias de nuestro estudio. En caso contrario nos limitaremos a describir los datos recogidos.

El modelo casi siempre se puede concretar en una fórmula. Por ejemplo, en este caso, podíamos suponer que el rendimiento obedece a una fórmula aproximada parecida a esta:

Rendimiento = A*Capacidad + B*Número de horas + Parte aleatoria o individual + C

donde A,B y C son constantes que hay que determinar.

Recogida de datos

Una vez que hemos decidido el modelo, deberemos recoger datos de nuestra población. Según nuestras posibilidades se puede:

- Recoger los datos que tenemos a mano, tal como nos vienen
- Plantear un muestreo aleatorio
- Diseñar un experimento

Estimación de parámetros

A partir de los datos de un muestreo se pueden estimar los valores de los parámetros. En el ejemplo, deberíamos intentar encontrar valores adecuados para A,B y C.

Simplificación del modelo

A la vista de los datos, pueden existir parámetros repetidos o de valor tan pequeño que se puedan despreciar. En nuestro ejemplo, si C es casi cero, podríamos plantearnos eliminarlo y volver a estimar.

Crítica y diagnosis

Si nuestro modelo presenta un buen ajuste a los datos, deberemos proceder a redactar informes, describir los puntos débiles y volver a aplicar el modelo a otros colectivos para comprobar su eficacia.

PROBLEMAS QUE RESUELVE LA ESTADÍSTICA

DESCRIPCIÓN DE LOS DATOS

Los estudios estadísticos, para poseer una cierta fiabilidad, deben basarse en la recogida de muchos datos, cuantos más mejor. Por ello, la Estadística dispone de técnicas para

- Recoger datos, lo más representativos posible
- Resumir datos en tablas ordenadas de frecuencias
- Representar gráficamente las tablas obtenidas
- Efectuar medidas representativas del conjunto recogido: Centrales, de dispersión, asimetría y aplastamiento, correlaciones, etc.

Se suele llamar Estadística Descriptiva al estudio de estos aspectos. La mayoría de los trabajos escolares de tipo estadístico se limitan a estos aspectos.

MUESTREO

La recogida de datos no se debe efectuar sin un planteamiento previo. Las técnicas que nos ayudan a elegir muestras representativas de las poblaciones constituyen la Teoría del Muestreo. Suele ser una parte bastante aburrida y que sólo la estudian los especialistas en encuestas, sondeos o controles de calidad.

La operación fundamental del muestreo es *estimar* los parámetros de la población a partir de los datos de la muestra. El conjunto de técnicas usadas es parte de la Estadística Inferencial.

Es posible, con apoyo teórico mínimo, que los alumnos de Bachillerato (y quizás algunos de ESO) puedan plantearse estimaciones elementales.

CONTRASTE DE HIPÓTESIS

En la actualidad es la parte de la Estadística más usada en todo tipo de investigaciones. Ningún trabajo de nivel universitario o profesional se admite sin estar basado en un Diseño de Experimentos y en el uso de las

técnicas de Contraste de Hipótesis. Consiste en plantear una hipótesis (llamada nula) frente a otra alternativa, recoger datos representativos y comprobar si estos son consistentes con las hipótesis o no.

Por ejemplo, ningún fármaco entra en el mercado sin estudios estadísticos que apoyen la hipótesis de que cura una dolencia determinada.

Las técnicas usadas forman un cuerpo de teoría muy amplio llamado Estadística Inferencial y Diseño de experimentos.

MEDIDA DE RELACIONES

En la Sociedad y la Naturaleza se pueden descubrir relaciones y paralelismos que, en algunos casos, permiten representar mediante una fórmula una relación entre dos o más variables. Llamaremos *Correlación* al estudio de estas relaciones. Aunque el caso más frecuente es el de la comparación de dos variables, cada día se emprenden más estudios en los que algunas variables se relacionan con varias otras, llamadas *explicativas*.

Estos paralelismos no se deben confundir con relaciones causa-efecto, que tienen tratamientos más profundos.

ESTABLECIMIENTO DE PREDICCIONES

Los datos recogidos en un estudio pueden presentar tendencias o ciclos que quizás nos permitan predecir qué va a ocurrir fuera del rango de datos obtenido. Se puede intentar predecir qué va a ocurrir en el futuro, por ejemplo, lo que constituiría una extrapolación, o bien en valores intermedios, y la llamaríamos interpolación. Esta técnica también puede servir para completar datos perdidos o erróneos.

En el establecimiento de las predicciones es fundamental conocer los márgenes de error. Las técnicas consiguientes están recogidas en la Teoría de la Regresión.

RECOGIDA, TABULACIÓN Y ORGANIZACIÓN DE DATOS

CUESTIÓN - EJEMPLO

¿Qué número de letras suelen tener las palabras en nuestro idioma?

Un trabajo muy ameno en las clases de Estadística es efectuar un recuento del número de letras que suelen tener las palabras en nuestro idioma. Se puede organizar un recuento de datos en varios niveles. Por ejemplo, algunos equipos de alumnos pueden elegir textos de prensa, otros de libros técnicos, de revistas de Informática y otros, e intentar descubrir diferencias entre la distribución de letras en las palabras de cada tema. También puede ser interesante comparar unos idiomas con otros.

Una variante de este trabajo puede ser el descubrir la vocal más frecuente en cada uno de los idiomas, o el reparto de vocales y consonantes en las palabras, o la abundancia de adjetivos o ciertas conjunciones.

A continuación puedes ver los datos que hemos obtenido con tres recogidas diferentes:

1. Prensa: Textos procedentes de varios ejemplares de prensa, con párrafos elegidos aleatoriamente
2. Técnicos: Párrafos extraídos de revistas de Informática
3. Sociales: Algunos textos procedentes de libros de Ética y Sociología

Núm. letras	Prensa	Técnicos	Sociales	Total
1	14	6	17	37
2	83	103	97	283
3	44	50	43	137
4	34	36	28	98
5	38	29	41	108
6	25	30	23	78
7	24	33	31	88
8	19	24	25	68

9	**12**	**18**	**25**	**55**
10	**18**	**5**	**12**	**35**
11	**5**	**10**	**6**	**21**
12 o más	**6**	**8**	**12**	**26**
Totales	**322**	**352**	**360**	**1034**

Se ha detenido el recuento cuando se ha superado el número de trescientos y se ha explorado el último párrafo completo. Este hecho explica el que los totales sean diferentes en las tres columnas.

Si lo deseas, abre el modelo de Hoja de Cálculo letras.ods.

(http://www.hojamat.es/estadistica/tema1/open/letras.ods)

Observarás que contiene la tabla que acabas de leer y que constituye un ejemplo claro de la naturalidad con la que una Hoja de Cálculo maneja las tablas de tipo estadístico.

Recuerda que lo único que necesitas saber de hojas de cálculo para seguir este curso son los conocimientos mínimos sobre la estructura de filas y columnas, la existencia de fórmulas en algunas celdas y la edición de las mismas. El resto lo irás aprendiendo sobre la marcha. No obstante, puedes leer el primer capítulo de las Guías presentadas más arriba para aprender lo que necesitas en esta primera sesión.

Lo que nos interesa en este momento es la teoría estadística que hay detrás de esta tabla de recuento. Analízala:

Variable estadística

La primera columna de la tabla constituye la *variable estadística* que estamos estudiando. Es una variable porque puede tomar más de un valor y suele representar la característica que deseemos estudiar. En este caso oscila entre 1 y 12 o más. Cuando leas el resumen teórico aprenderás más sobre variables. De momento, basta considerar que la variable que nos interesa es el número de letras de las palabras.

La variable contiene una *característica*, que en este ejemplo es el número de letras de cada palabra recogida. Es una característica *cuantitativa*, porque se

expresa mediante un número. Si no se pudiera representar por números la llamaríamos *cualitativa*.

Cuando se estudia una variable puede ser interesante concretar el tipo de medida que se usa para recoger los datos. Este concepto de *tipo de medida* es fundamental en Estadística, pero en este momento no tienes que profundizar demasiado en él. Acude al resumen teórico para leer las definiciones.

Las medidas que usamos en este caso son;

- **Cuantitativas**, puesto que se basan en números
- **Discretas**, ya que entre 2 letras y 3 letras no hay otras posibilidades.
- **De intervalo**, porque tiene una unidad (una letra) y tiene sentido restar dos medidas para compararlas.
- **No es de razón**, dado que no nos interesa el cero ni el cociente entre dos medidas. Si hubiéramos querido, también podíamos haberlas considerado como de razón, pero no aporta nada en este caso.

Así pues, la variable de estudio está medida **a nivel de intervalo, es cuantitativa y discreta**, porque sólo puede tomar los valores aislados 1,2,3,...

Frecuencias

Si la primera columna contiene *la variable* que estudiamos, las siguientes columnas representan *las frecuencias*, que recogen el número de veces que ha aparecido cada valor 1,2,3... Al haber tres fuentes de datos, hay también tres columnas de frecuencias, pero eso no es habitual. En la tabla falta la última columna de totales, que la rellenarás tú en las Prácticas. Es importante que entiendas lo que es la frecuencia y como se representa:

El número de veces que se repite un valor concreto en una recogida de datos se llama ***frecuencia absoluta*** o simplemente frecuencia. Se representa por la letra **n** o por la **f**, según los distintos textos. Aquí usaremos **n**. La suma de todas las frecuencias coincide con el número total de elementos estudiados, al que representaremos por N.

Así que en nuestra tabla las columnas segunda a cuarta representan frecuencias absolutas: la frecuencia de las

palabras de 8 letras en los textos técnicos es de 24, la de 11 letras en Sociales es de 6, etc.

Intentaremos descubrir las diferencias que pueden existir entre las tres columnas (Prensa, Técnicos y Sociales). Nos tenemos que plantear este estudio porque **es imposible comparar directamente las frecuencias**, a causa de los distintos totales que presentan las tres modalidades (322, 352 y 360). Esto nos obliga a acudir a frecuencias relativas o porcentajes, como verás en las Prácticas.

TIPOS DE MEDIDA

Características y modalidades

Llamamos *característica* a cualquier propiedad de objetos o personas que deseamos estudiar en Estadística. Las distintas formas de presentarse esta característica se llaman *modalidades*. Por ejemplo, 1,82 y 1,65 cm. son dos modalidades de la característica *altura*, y varón y mujer dos modalidades de la característica *sexo*.

Si una característica sólo tiene dos modalidades la llamaremos *dicotómica*.

Medida

Es la operación de asignar un número a cada una de las modalidades de una característica, convirtiendo algunas relaciones entre modalidades en sus correspondientes relaciones entre los números que representan su medida. Por ejemplo, los ciudadanos españoles se corresponden con su DNI, su peso con los kg que da la balanza, y el sexo se puede corresponder con los símbolos V y M, etc.

Escala de medida

Es un conjunto básico de modalidades y números (considerados como sus medidas) a partir del cual se construye un procedimiento para medir las restantes modalidades. Así, la escala centígrada de temperaturas se basa en asignar 0º a la temperatura de fusión del agua y 100º a la de ebullición. La medida de la dureza de los minerales se basa en un conjunto determinado de ellos, desde el talco hasta el diamante.

En Estadística consideramos cuatro tipos básicos de escalas, desde la *nominal*, que apenas permite análisis, hasta la de *razón*, que es la más completa.

Se diferencian unas de otras esencialmente en las operaciones que permiten.

Escala nominal

Una escala se llama *nominal* si la única relación que tiene en cuenta es la de *igualdad* (y su contraria la desigualdad). Suele estar formada por nombres, códigos o números considerados como etiquetas (como el DNI). Así, son nominales los apellidos, la Comunidad Autónoma, el distrito postal, etc.

Si una escala nominal se construye con números (como los distritos postales), sólo se admitirán entre ellos las relaciones de igualdad y desigualdad, pero no operaciones como suma o producto. No podemos sumar dos Comunidades Autónomas. Tampoco tendría utilidad multiplicar dos números de teléfono.

Escala ordinal

La escala *ordinal* añade a la nominal la posibilidad de ordenar los datos, es decir, considera las relaciones de *mayor* y *menor*, aunque no se plantea una distancia entre unas medidas y otras. La escala de Insuficiente,

Suficiente, Bien, Notable y Sobresaliente es ordinal. No se considera si entre Bien y Notable existe la misma diferencia que entre Notable y Sobresaliente. Son ordinales muchas de las medidas en Psicología o Ciencias de la Educación.

Escala de intervalos

Se introduce una medida tipo (o patrón) llamada *unidad* y se tiene en cuenta cuantas unidades están comprendidas entre dos medidas distintas. Tienen sentido, además de la igualdad y el orden, las *diferencias* entre dos medidas. Podemos sumar y restar medidas, pero no tienen sentido sus cocientes. Son de intervalo la gran mayoría de las escala de las ciencias experimentales: temperatura, peso, velocidad, intensidad de la corriente eléctrica, etc.

Escala de razón

En esta escala se le da también un sentido a las *razones* entre dos medidas, es decir, las veces que una medida contiene a la otra. Fue la medida por excelencia de la Geometría griega y se ha trasladado a todas las Ciencias Sociales y de la Naturaleza. Se distingue también por la existencia de un *cero verdadero*, no

convencional. Así, la escala centígrada de temperatura es sólo de intervalo y la Kelvin es de razón.

Podemos dividir medidas, pero sólo para su comparación o razón.

Resumen

· En escala nominal sólo distinguiremos la igualdad o desigualdad entre dos modalidades.

· La escala ordinal añade la posibilidad de establecer un orden.

· Si se usa una unidad y tienen sentido las diferencias, se trata de una escala de intervalo.

· Por último, si de pueden comparar dos medidas mediante un cociente o razón, la escala es de razón.

CONSTANTES Y VARIABLES

Llamaremos *constante* a una característica que sólo admite una modalidad, por ejemplo la constante de gravitación universal. Por el contrario, *variable* es

aquella que admite varias modalidades, a las que también llamaremos *datos* o *valores*.

Tipos de variables

Una variable se llama ***cualitativa*** si sólo admite una medida nominal. Son cualitativas la localidad de nacimiento de cada persona, el color de su piel o su domicilio.

Llamaremos ***casi cuantitativa*** a aquella que admite como máximo una medida ordinal, como podría ser la motivación en el estudio, el grado de extraversión o la valoración de una prueba de gimnasia rítmica.

Por último, llamaremos ***cuantitativa*** a aquella variable que admita medidas de intervalo o de razón. Si entre cada dos valores pueden existir infinitos otros, la llamaremos ***continua***, como el peso, la estatura, etc. y si solo admite un número finito de valores entre cada dos, recibirá el nombre de ***discreta*** (edades medidas en años, número de hermanos, etc.). A causa de la falta de precisión en las medidas, muchas variables continuas se pueden tratar como discretas, y lo contrario sucede cuando existen tantos valores distintos

que puede ser útil tratar como continua una variable discreta.

Los datos de cualquier tipo de variable pueden ser simples, de un solo valor, y se llaman en este caso ***unidimensionales***. Llamaremos ***bidimensionales*** a los datos compuestos de dos valores, como un resultado de Baloncesto (89 a 76). Existen además datos ***tridimensionales*** y, en general, multidimensionales.

RECOGIDA DE LOS DATOS

Los datos se recogen de conjuntos reales, por lo que debemos hacer algunas distinciones:

Población y muestra

Llamaremos ***población*** a un conjunto bien definido por ciertas características que deseamos estudiar: La población de una Comunidad Autónoma, los aprobados de 2º de Bachillerato en mi Centro, los profesores de E.S.O. en la Delegación Norte, etc.

Como las poblaciones pueden contener muchos elementos distintos, elegiremos una ***muestra***, o subconjunto de ellas, que sea más fácil de estudiar que la población. Existe toda una ciencia para conseguir muestras representativas de la población.

Un estudio estadístico de la población recibe el nombre de ***censo*** y el de la elección y estudio de una muestra, ***muestreo***.

Un número que caracterice o describa una población recibe el nombre de ***parámetro***. La estatura media de los alumnos y alumnas de 16 años es un parámetro de esa población, o la Renta per cápita de la población española. Si ese mismo número lo calculamos en una muestra, recibe el nombre de ***estadístico***. Si mido la estatura media de los alumnos de mi clase de Bachillerato estaré calculando un estadístico. Cuando se calcula un parámetro a partir de un estadístico, estaremos realizando una ***estimación***. Un caso representativo es el de los sondeos antes de unas elecciones.

Fuentes de datos

Si nos restringimos al trabajo con alumnos, destacaríamos los siguientes:

Encuesta y/o entrevista

Se usan cuestionarios de preguntas generalmente cerradas (caso de la encuesta) y abiertas dentro de un diálogo, lo que constituiría la entrevista.

Observación

Puede tener distinta forma según el modo de observar. Podríamos distinguir entre

Experimento: Se organizan las condiciones exactas de la recogida, controlando bien todas las variables menos las que nos interesan. Por ejemplo, deslizar un bola por un plano inclinado y medir el tiempo.

Observación no planificada: Los datos nos vienen dados sin que podamos controlar nada, como sería el número de personas que pasan por la calle en cada minuto.

Consulta

Los datos ya existen publicados y se procede a extraerlos de su fuente: bibliotecas, documentos oficiales, los contenidos en Internet y otros.

Simulación

En este caso el interés sería más bien teórico, ya que los datos procederían de una simulación aleatoria, generada mediante un programa de ordenador o similar. Así se puede estudiar, por ejemplo, la distribución binomial.

Recuento

Los censos y muestreos necesitan siempre un ***recuento de datos***.

Si los recuentos de datos se efectúan manualmente, se suelen representar mediante trazos verticales:

María |||||

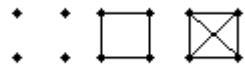
José Mari |||||||

aunque son más útiles configuraciones de cinco en cinco o de diez en diez.

Así, esta configuración de recogida representa el número 13 mediante cuadrados con diagonal que representan el número 5

También es útil representar el número 5 con cuatro barras y la quinta tachando a las demás.

Si se desea contar el número diez, se puede seguir esta secuencia de puntos y rectas:



(cuatro puntos más cuatro rectas más dos diagonales: 10)

ORGANIZACIÓN DE LOS DATOS

Unas vez efectuado el recuento de datos dispondremos de unos números llamados ***frecuencias*** asignados a los distintos valores de las variables. A la operación de construir ese conjunto se le suele llamar coloquialmente ***confeccionar una estadística***.

Frecuencias

El número de veces que se repite un valor concreto en una recogida de datos se llama ***frecuencia absoluta*** o simplemente frecuencia. Se representa por la letra **n** o por la **f**, según los distintos textos. Aquí usaremos **n**. La suma de todas las frecuencias coincide con el número total de elementos estudiados, al que representaremos por N.

Representaremos esto así

$$\sum n = N$$

Para poder comparar distintos conjuntos de datos es más útil el uso de las ***frecuencias relativas o proporciones*****,** que son los cocientes de dividir cada frecuencia absoluta entre el total de valores N. Se representan por **f** (así lo haremos nosotros) o por **h**.

$$f = \frac{n}{N}$$

La suma de todas las frecuencias relativas es igual a uno:

$$\sum f = 1$$

La frecuencia relativa es una razón (o cociente). Por tanto se puede convertir en ***porcentaje*** multiplicándola por 100, y así representa el tanto por ciento del total que representa cada dato. Lo representamos por ***p***.

Por tanto se cumplirá que $p = f \times 100$

y que $\sum p = 100$

Frecuencias acumuladas

Cuando la variable está medida al menos a nivel ordinal permite la acumulación de frecuencias.

La frecuencia acumulada de un valor es el número de datos del conjunto que son *menores o iguales* a él. Por tanto, se calculará sumando todas las frecuencias de datos menores o iguales al mismo. Podemos acumular

las frecuencias absolutas y también las relativas y los porcentajes.

Las frecuencias acumuladas serán crecientes y la última absoluta coincidirá con N, la última relativa con 1 y el porcentaje último con 100.

Distribución de frecuencias

El conjunto formado por los valores de la variable y sus frecuencias (hasta seis columnas) constituye la ***distribución de frecuencias*** de la población o muestra, y se representa en las ***tablas de frecuencias***, primer paso obligado de un estudio estadístico.

Tabla de frecuencias

Dato X	n_i	f_i	p_i
2	3	0,06	6%
3	7	0,14	14%
4	12	0,24	24%
5	18	0,36	36%
6	7	0,14	14%
7	3	0,06	6%
Total	50	1	100%

AGRUPACIÓN DE DATOS

Si la variable que se estudia es continua, o discreta con muchos valores distintos, se organizarán sus datos en forma de intervalos. Para ello se fija un valor mínimo y otro máximo, de forma que todos los datos estén comprendidos entre ellos (a veces esto no se garantiza y quedan intervalos abiertos). La diferencia entre ambos se denomina ***rango*** de los datos y posteriormente se divide en un número de *intervalos* mediante valores intermedios.

Esto se construye para después situar cada dato en su intervalo correspondiente y hacer el recuento, con lo que cada intervalo poseerá una *frecuencia*.

Para representar cada intervalo disponemos de

Extremo inferior: Es el valor mínimo que puede tener un valor incluido en ese intervalo.

Extremo superior: Es el valor máximo posible. Se considera no alcanzable. Así si un intervalo comprende desde 5 hasta 10, incluiremos en el mismo los valores comprendidos entre estos dos, incluyendo el 5 y sin incluir el 10.

Marca de clase: Promedio entre los dos extremos (o punto medio del intervalo), que se elige como representante de todos los valores comprendidos. Esto constituye una pérdida de exactitud, compensada por el mejor manejo de los datos agrupados.

Amplitud: Es la diferencia entre los dos valores extremos.

Tabla de datos agrupados

Extremo inf.	Extremo sup.	Marca de clase	frecuencia
120	130	125	4
130	140	135	23
140	150	145	31
150	160	155	7

En la figura podemos ver una tabla con cuatro intervalos de amplitud 10, en los que se representan los dos extremos y las marcas de clase, junto a sus frecuencias.

El número aconsejado de intervalos suele estar entre 5 y 15, aunque se pueden elegir con libertad según las características del estudio. Una regla empírica nos

aconseja elegir como número de intervalos **la raíz cuadrada entera del número de observaciones**. También se suelen manejar intervalos todos iguales, aunque a veces se altera la amplitud de algunos para destacar algo de interés en la distribución.

En algunas cuestiones se puede suponer que los datos incluidos en un intervalo se distribuyen en el mismo de manera uniforme, pero otras veces es mejor suponer que todos coinciden con la marca de clase. Ambas hipótesis son falsas, lo que supone que la agrupación en intervalos supone siempre una pérdida de información.

UN CASO PRÁCTICO

Recogida de los resultados de una encuesta con respuestas de texto libres.

Un grupo de alumnos y alumnas participa en una excursión a la montaña, en la que se organiza una ruta para practicar la orientación y la interpretación de señales. Después de una comida en común se organizan juegos y dinámicas. Antes de regresar se les

entrega una encuesta con preguntas de respuesta libre. Una de las preguntas es "¿Qué te ha gustado más de la excursión?

Recogidas las respuestas, presentan este resultado, que representamos sin estructurar, como si una persona las hubiera copiado sin más.

El paisaje - Las actividades - Los juegos - El viaje - No me ha gustado nada - La montaña - Los acompañantes - La comida - Los árboles - El trayecto en autocar - El viaje al campo - La ruta de orientación - Mi equipo - Los monitores - Los pinos - El paseo entre árboles - El buscar la ruta - Los guías eran muy simpáticos - El autocar - La subida entre árboles - Los bocatas - Los árboles - La ruta - Seguir las señales -Los monitores - Mis amigos - El puente romano - La escalada - Los pájaros -La charla de la mañana - El viaje en autocar - El río - Los compañeros de equipo - El haber llegado los primeros - Seguir los círculos de los árboles - La montaña - El buen tiempo - El valle - Las dinámicas - Los montes - Mis amigos.

¿Cómo tratar estadísticamente estos datos?

El problema de estas encuestas con respuesta libre es su dispersión, y que varias respuestas pueden significar

el mismo sentir, pero expresado de diversa forma. Esto obliga a filtrar esas respuestas en diversas categorías. En concreto, las características de estas encuestas son:

- Gran dispersión en las respuestas, lo que hace que abunden los resultados con frecuencia 1 o 2.
- Ambigüedad en las respuestas. Por ejemplo "Me ha encantado", "Ha estado muy bien" , "Muy buena impresión",... pueden significar lo mismo, pero quizás en algún caso concreto se pretenda introducir un matiz.
- Necesidad de un filtro. Para conseguir un significado estadístico deberemos agrupar las respuestas en categorías, lo que introduce una componente subjetiva, que puede falsear el estudio.

En este ejemplo se perciben tres categorías básicas, según el objeto de cada comentario: **Naturaleza**, **Personas** y **Actividades**. Además, hay respuestas aisladas, como "*El haber llegado los primeros*", que se resisten a esa clasificación y se pueden agrupar en el apartado de Otros.

En cada categoría se pueden introducir subcategorías. Así, la Naturaleza puede comprender:

- Montes NM
- Árboles NA
- Paisaje en general NG
- Otros NO

A cada categoría le hemos asignado un código.

Los comentarios sobre personas se pueden dividir en

- Amigos o compañeros PA
- Monitores o guías PM

Y los de actividades en

- Viaje AV
- Ruta AR
- Juegos o dinámicas AJ
- General AG
- Otros AO

El apartado Otros lo dejaremos sin dividir en subcategorías. OTR

El paisaje	NG	El viaje al campo	AV	Los árboles	NA	Los compañeros de equipo	PA
Las actividades	AG	La ruta de orientación	AR	La ruta	AR	El haber llegado los	OTR
Los juegos	AJ	Mi equipo	PA	Seguir las señales	AR	primeros	AR
El viaje	AV	Los monitores	PM	Los monitores	PM	Seguir los círculos de los	NM
No me ha gustado nada	OTR	Los pinos	NA	Mis amigos	PA	árboles	OTR
La montaña	NM	El paseo entre árboles	AR	El puente romano	NO	La montaña	NG
Los acompañantes	PM	El buscar la ruta	AR	La escalada	AR	El buen tiempo	AJ
La comida	AG	Los guías eran muy		Los pájaros	NO	El valle	NM
Los árboles	NA	simpáticos	PM	La charla de la mañana	AJ	Las dinámicas	PA
El trayecto en autocar	AV	El autocar	AV	El viaje en autocar	AV	Los montes	
		La subida entre árboles	AR	El río	NO	Mis amigos.	
		Los bocatas	OTR				

Para más tarde efectuar un recuento con Calc:

Recuento

Naturaleza		Personas		Actividades		Otros	
NG	2	PA	4	AR	8		
NM	3	PM	4	AJ	3		
NA	3			AG	2		
NO	3			AV	5		
Total	11	Total	8	Total	18	Total	4
						Total recuento	41

Naturaleza	11
Personas	8
Actividades	18
Otros	4

De esta forma hemos conseguido resumir un conjunto tan variado de respuestas, pero a costa de sacrificar la espontaneidad.

Una vez hecha la tabulación, podremos comentar los datos:

- Lo que más se destaca es el conjunto de actividades (18), especialmente la ruta de orientación (8)
- Las respuestas se ajustan bastante a la idea de los entrevistadores, pues sólo 4 respuestas han sido inclasificables.

- Destaca la importancia que se le ha dado al viaje en autocar (5), que no pertenecía propiamente a la actividad.
- Los aspectos naturales (árboles, montañas,...) presentan variedad de respuestas (2+3+3+3)

A partir de estos datos también podemos emprender un estudio gráfico, por ejemplo, de los aspectos:

CUESTIÓN – EJEMPLO

¿Cómo ha influido mi cambio de método?

Una profesora de idiomas decide cambiar el método de enseñanza del vocabulario. Pasa a sus alumnos una prueba que produce una calificación del 0 al 5. Establece en sus clases el nuevo método durante dos meses y vuelve a pasar una prueba de dificultad proporcionada al nuevo aprendizaje. Después de corregirla desea saber si ha subido el nivel de clase y si el nuevo método ha acercado los niveles de los alumnos o si, por contra, los ha dispersado.

Los resultados de ambas pruebas, ordenados de 0 a 5 (no por alumnos) han sido los siguientes

Antes del cambio de método	0 0 0 1 1 1 2 2 2 2 2 2 3 3 3 3 3 4 4 4 4 4 5 5 5
Después del cambio	0 0 1 1 1 1 2 2 2 2 3 3 3 3 3 4 4 4 4 4 4 5 5 5 5

Para estudiar diferencias que no se aprecian por simple inspección de las tablas (en el ejemplo parece que el nivel ha subido algo, pero no se tiene la seguridad) debemos calcular los *estadísticos* de la muestra. Llamamos estadísticos a todas las medidas realizadas sobre el conjunto (muestra) que se estudia, y que pueden resumir algunos aspectos interesantes del mismo:

Medidas centrales: *Mediana, media, moda*

Son medidas que buscan *el centro de los datos*. En nuestro ejemplo, para comprobar que el nuevo método es mejor, bastará ver que el centro (la media o la mediana) ha aumentado su valor.

Medidas de dispersión: *Desviación típica, rango,...*

Estas medidas evalúan la variabilidad de los datos. Si son grandes, es que los datos están más dispersos, y si son pequeñas, más homogéneos. En este ejemplo sería un peligro el hecho de que el nuevo método dispersara a los alumnos.

Medidas de asimetría y aplastamiento

No las usaremos en este ejemplo. Consulta la teoría si lo deseas.

En esta sesión sólo estudiaremos las medidas de tipo paramétrico, llamadas así porque se usan en estimaciones en las que se tienen en cuenta los *parámetros* de la población, que son las medidas del mismo tipo que los *estadísticos*, pero correspondientes a la población y no a la muestra. Es una distinción técnica. No te preocupes por ella en este momento.

MEDIDAS DE TENDENCIA CENTRAL

Se llaman medidas de tendencia central, de posición o *promedios* de una distribución de datos a aquellos valores que indican el centro de la distribución, y pretenden representar todos los datos en un solo punto.

MEDIA

Llamaremos *media aritmética* o simplemente *media* al valor resultante de sumar todos los datos y después dividir el resultado entre el número de ellos. Es, pues, el

resultado de un reparto igualitario de los valores. También se puede interpretar como el *centro de gravedad* de los datos.

Su fórmula más simple es

$$\bar{x} = \frac{\sum x}{N}$$

(suprimimos los subíndices en los sumatorios cuando no exista peligro de confusión)

Si los datos están agrupados en una tabla de frecuencias, la fórmula anterior se convertiría en

$$\bar{x} = \frac{\sum x \cdot n}{N} = \sum x \cdot f$$

que es la más usada en la práctica. Obsérvese que el uso de frecuencias relativas simplifica bastante su cálculo.

La suma de desviaciones de todos los datos respecto de la media es cero

$$\sum (x - \bar{x}) = 0$$

y la suma de los cuadrados de las desviaciones es la **mínima** respecto a la que resultaría al elegir otro valor cualquiera en la resta

$$\sum (x - \bar{x})^2 \leq \sum (x - k)^2$$

para cualquier valor de **k**.

La anterior suma se puede expresar también como

$$\sum (x - \bar{x})^2 = \sum x^2 - N \cdot \bar{x}^2$$

La media es propia sólo de datos cuantitativos. Si estos están agrupados, se establece la hipótesis de que están todos situados en la **marca de la clase** o punto medio.

La media es sensible a todos los datos. Si uno de ellos cambia, la media también se ve alterada. Por esta razón, influyen mucho en ella los valores extremos, que suelen ser poco fiables, por lo que a veces se desechan.

Su importancia radica en que es base de muchas técnicas estadísticas, pero no nos vale cuando existen intervalos abiertos o la medida solo tiene carácter ordinal.

OTRAS MEDIAS

También podemos usar

Media geométrica

Es la raíz enésima del producto de los datos. Se usa cuando el producto es más representativo que la suma, como ocurre cuando se promedian cocientes o razones.

$$mg = \sqrt[N]{\prod x} = e^{medLog(x)}$$

Media armónica

Es la media diseñada para promediar cantidades inversamente proporcionales y equivale al inverso de la media de los inversos de x

$$xa = \frac{1}{\sum \left(1/x\right) / N}$$

Media cuadrática

Es muy usada en la teoría de errores y en estudios sobre ajustes de datos. Es la raíz cuadrada de la media de los cuadrados de los datos.

$$mc = \sqrt{\frac{\sum x^2}{N}}$$

Media ponderada

Es muy interesante en la enseñanza, para promediar calificaciones con distintos pesos, además de múltiples utilidades en problemas de mezclas, centros de gravedad, cestas de inversiones, etc.

Se calcula asignando a cada dato un *peso*, cuya significación depende de cada problema, y suponer que cada dato se repite tantas veces como indica su peso (aunque este no sea entero. Esa es la diferencia entre peso y frecuencia). Se procede pues a multiplicar cada dato por su peso, para dividir después entre la suma de los pesos.

$$mp = \frac{\sum x \cdot p}{\sum p}$$

MEDIANA

Llamaremos *mediana* de un conjunto de datos de tipo ordinal (o de intervalo o razón) al dato que ocupa el punto medio de la distribución ordenada de datos. Es decir, es el punto que divide a la distribución en dos partes iguales: el total de frecuencias de los datos inferiores a la mediana es igual al de las frecuencias de los datos mayores. Si los datos son continuos o muy numerosos, se puede afirmar con cierta aproximación que ese total de frecuencias es el 50%.

En el cálculo de la mediana sólo se utiliza el **orden** de los datos y no su magnitud. Por eso es la medida adecuada para escalas de tipo ordinal, como las que se usan en Psicología y Ciencias de la Educación.

La mediana no es una medida bien establecida, pues contiene elementos convencionales. Tanto es así que

en algunos casos su definición cambia de unos manuales a otros.

Podemos establecer algunas convenciones para su cálculo:

Datos aislados

Si el número de datos es **impar**, se toma como mediana el dato central, que deja (n-1)/2 datos menores que él y otros tantos mayores: La mediana de 2 2 2 3 3 3 4 5 6 7 7 8 9 es **4**.

Si el número de datos es **par**, se toma como mediana el promedio de los dos datos centrales: la mediana de 2 3 4 5 6 8 9 10 es **5,5**.

Algunos autores dan reglas más precisas para el caso en el que los valores centrales están repetidos. En caso de duda es preferible agrupar los datos por frecuencias y usar la teoría correspondiente.

Datos agrupados

Si los datos están agrupados por intervalos, se usa la siguiente fórmula (que es fácil de justificar mediante

proporciones si se supone que todos los datos están distribuidos de manera uniforme en el intervalo)

$$Me = L_i + \frac{(N/2 - n_{ant})}{n_{interv}} \cdot Ampl.$$

En la fórmula los símbolos se interpretan así:

N/2 es la mitad del número de datos, con decimales si los hubiere.

Ampl. es la amplitud del intervalo mediano, el que contiene la frecuencia acumulada 0,5.

n_{ant} es la frecuencia acumulada anterior al intervalo mediano.

$n_{interv.}$ es la frecuencia absoluta de dicho intervalo.

L_i representa el límite inferior verdadero del intervalo mediano. Este concepto es muy importante, pues en caso de datos agrupados por frecuencias, pero no por intervalos, los límites verdaderos de un valor, por ejemplo 23, serían 22,5 y 23,5.

Propiedades de la mediana

La mediana es menos sensible a datos extremos que la media, por lo que se pueden considerar estos con menos peligro de sesgo en los resultados. También es muy estable para pequeñas variaciones en los datos.

Es muy útil si los datos son ordinales, o excesivamente asimétricos, o si en los intervalos existe alguno abierto, del tipo *60 o más*.

Cumple que la suma de los valores absolutos de las desviaciones respecto a ella es mínima

$$\sum |x - me|$$

es mínimo

Moda

Es la más pobre de las tres medidas y la más convencional. Llamaremos **Moda** al valor de la distribución de datos que presente una frecuencia mayor. Así, en el conjunto 2 2 3 3 3 4 4 5 6 6 la moda es 3, valor más repetido.

La definición se vuelve ambigua si existen varios intervalos con la misma frecuencia. En estos casos se distingue:

Valores separados: se consideran **moda** todos los valores de frecuencia máxima, y la distribución recibe el nombre de **bimodal, trimodal,** etc. Por ejemplo en el conjunto 3 3 4 4 4 5 6 7 7 7 8 8 9 las dos modas son 4 y 7 y la distribución será bimodal.

Valores consecutivos: Se usa, por convención la siguiente fórmula:

$$Mo = L_i + \frac{d_1}{d_1 + d_2} \cdot Ampl.$$

en la que L_i representa el límite inferior verdadero del intervalo, d_1 la diferencia de frecuencia con el intervalo anterior, la d_2 la diferencia con el siguiente y *Ampl.* la amplitud del intervalo.

La **moda** es una medida sencilla pero poco representativa del conjunto de datos. Es la única posible en datos cualitativos, pero se puede usar en intervalos abiertos. Su mayor inconveniente es la falta

de unicidad en su definición y su sesgo en distribuciones muy asimétricas. Por el contrario, es bastante estable frente a variaciones de los datos.

MEDIDAS DE VARIABILIDAD

Las medidas de variabilidad nos informan sobre el grado de concentración o dispersión que presentan los datos respecto a su promedio. Llamaremos ***homogénea***, concentrada o poco dispersa a aquella distribución en la que todos los datos están cercanos al centro, como 4 4 5 5 5 5 6 6 6 6 7, y ***heterogénea*** o dispersa a la distribución con datos más separados del centro, como 1 3 5 8 10 16 20.

Existen muchas formas de medir la variabilidad. Destacaremos las más importantes.

RANGO

También llamado ***Recorrido*** o ***Amplitud total***, es la diferencia entre el máximo valor del conjunto de datos y el mínimo de ellos. A mayor rango, mayor dispersión.

El rango del conjunto 4 6 4 7 8 6 5 3 4 7 7 9 6 5 es 6, la diferencia entre el máximo 9 y el mínimo 3.

A veces se usa el ***Rango verdadero*** que consiste en considerar cada dato rodeado de una unidad, por efecto de los redondeos, con lo que en el ejemplo anterior el mínimo sería 2,5 y el máximo 9,5. Con ello el rango se convertiría en 7.

No es una medida buena, pues ignora todo lo que ocurre dentro de ese rango.

DESVIACIÓN MEDIA

Es una medida de la dispersión consistente en la media aritmética de las desviaciones individuales respecto a la media, tomadas en valor absoluto. También se usan desviaciones respecto a la mediana.

Su fórmula, pues, será:

$$DM = \frac{\sum n_i \cdot \left| x_i - \bar{x} \right|}{N}$$

No es muy útil, pues carece de propiedades importantes. Además, tiene el inconveniente de que no es derivable.

VARIANZA

Si en la fórmula anterior sustituimos los valores absolutos por cuadrados (es otra forma de convertirlos en positivos), obtendremos la ***Varianza s^2***. Su fórmula será:

$$s^2 = \frac{\sum n_i \cdot (x_i - \overline{x})^2}{N} = \frac{\sum x_i^2 \cdot n_i}{N} - \overline{x}^2$$

Es fácil demostrar la equivalencia de las dos fórmulas.

Si los datos están aislados basta suprimir las frecuencias n_i de las fórmulas.

Es una medida muy sensible de la variabilidad y base de muchas técnicas estadísticas. Junto con la media forma el conjunto más importante de medidas.

Es propia de las medidas de intervalo o razón. Su inconveniente es que no usa la misma unidad que los datos, sino su cuadrado.

No se deben comparar varianzas en conjuntos de unidades muy distintas, como estatura e inteligencia.

En teoría del muestreo se sustituye por la ***cuasivarianza***, de idéntica fórmula, pero con cociente N-1 en lugar de N. En este caso no sería válida la segunda fórmula. En otro momento trataremos de ella.

DESVIACIÓN TÍPICA

Es la raíz cuadrada de la anterior. Su objeto es conseguir medir la variabilidad en las mismas unidades que los datos. Así, un conjunto medido en metros, tendrá la varianza medida en metros cuadrados, pero la desviación típica en metros. Tiene como fórmula

$$s = \sqrt{\frac{(x_i - \bar{x})^2 \cdot n_i}{N}} = \sqrt{\frac{\sum x_i^2 \cdot n_i}{N} - \bar{x}^2}$$

Como en la varianza, para datos aislados basta con suprimir las frecuencias n_i.

La desviación típica ***s*** es base de muchas técnicas, al igual que la media y la varianza. Su gran ventaja es

estar medida en las mismas unidades que los datos y la media, lo que permite establecer razones y proporciones entre ellas.

La desviación típica cumple la llamada ***desigualdad de Tchebychev***:

$$Pr\left(\left|x_i - \bar{x}\right| \leq ks\right) \geq 1 - \frac{1}{k^2}$$

según la cual, los datos que se alejan de la media una distancia igual o menor que **s** multiplicado por un coeficiente **k** suponen más de la proporción $1\text{-}1/k^2$. Así, el 75% de los datos al menos, se encuentra a menos de dos desviaciones típicas y el 89% a menos de tres.

En el estudio de las distribuciones teóricas podremos precisar más estas acotaciones.

COEFICIENTE DE VARIACIÓN

Cuando se comparan conjuntos de medias muy distintas (como podrían ser los diámetros de los planetas y la altura de mis alumnos) no sirve de nada comparar las distintas variabilidades. Entre las dos desviaciones típicas existiría una diferencia enorme en

magnitud. Por ello, se suele corregir la desviación típica dividiéndola entre su media. De esta forma obtenemos una medida *relativa* de la variablidad, que permite las comparaciones.

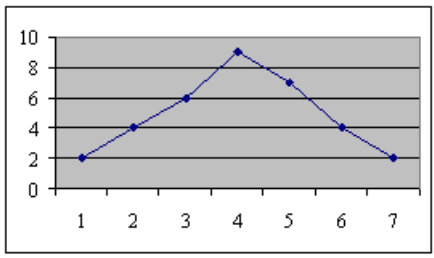
$$CV = \frac{\bar{x}}{s}$$

MEDIDAS DE ASIMETRÍA

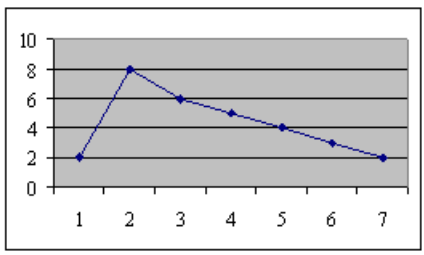
La ***asimetría*** o ***sesgo*** de una distribución es la característica por la que los datos pierden su simetría respecto a la media. Expresado de otra forma, es el mayor o menor grado de desviación que existe entre la media (reparto equitativo) y la mediana (punto medio de la distribución).

Si en un conjunto coinciden media y mediana, se presenta una ***simetría*** y cuanto más se separen, mayor será la asimetría de la distribución.

Será simétrica (aproximadamente) la distribución



y asimétrica esta otra



La distribución anterior, en la que existen muchas medidas bajas y pocas altas, diremos que presenta una ***asimetría positiva***. Si ocurriera lo contrario, diríamos que era ***asimétrica negativa***.

ÍNDICES DE ASIMETRÍA

En las distribuciones asimétricas positivas la media es mayor que la moda (punto más alto de la gráfica), y lo contrario ocurre en las negativas. Por eso Pearson

sugirió medir la asimetría mediante su diferencia dividida entre la desviación típica.

$$A_{p1} = \frac{\overline{x} - Mo}{s}$$

Pero como existe una ley empírica por la cual la diferencia entre la media y la moda es el triple de la existente entre media y mediana, propuso también

$$A_{p2} = \frac{3(\overline{x} - Me)}{s}$$

Más precisa es la medida de Fisher, porque usa ***momentos de tercer orden***, es decir, los cubos de las desviaciones respecto a la media. El índice de Fisher no tiene unidades, es una razón o comparación, y su fórmula es

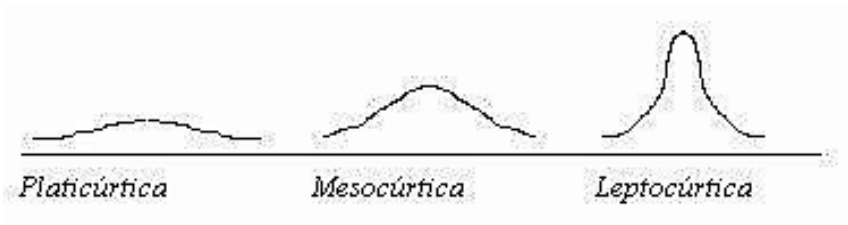
$$g_1 = \frac{\sum (x - \overline{x})^3 \cdot n_i}{s^3}$$

Será positivo o negativo cuando la asimetría también lo sea.

Existe otro índice, el de Bowley, que estudiaremos en otro momento.

MEDIDAS DE APLASTAMIENTO O CURTOSIS

Independientemente de su asimetría, una distribución puede presentar los datos con un reparto más uniforme, en el que las frecuencias sean muy parecidas. El gráfico aparecerá como aplastado y diremos que la distribución es ***platicúrtica*** o de ***poca curtosis***. En el otro extremo, si las frecuencias cercanas al centro son mayores (con diferencia) que las alejadas, diremos que es ***leptocúrtica*** o con ***gran curtosis***. Al caso intermedio lo denominaremos como distribución ***mesocúrtica***.



Para la medida del aplastamiento se usa otro índice de Fisher que usa el ***momento de cuarto orden***.

$$g_2 = \frac{\sum (x_i - \bar{x})^4 \cdot n_i}{s^4} - 3$$

En las leptocúrticas este índice es positivo, y negativo en las platicúrticas. Como veremos más adelante, en la distribución normal es nulo.

UN CASO PRÁCTICO

Un grupo de profesores y profesoras de un colegio han tenido unas jornadas de profundización en las técnicas de atención a la diversidad. Los organizadores desean conocer las opiniones de los asistentes respecto a varios aspectos de las jornadas, con respuestas de tipo ordinal: Muy mal, Mal, Regular, Bien y Muy Bien.

Las respuestas se han recogido en la siguiente tabla

Datos originales

CALIFICACIÓN	Muy mal	Mal	Regular	Bien	Muy bien
Contenidos			1	9	5
Metodología			3	12	6
Provecho		1	2	13	6
Cómo me he sentido			4	9	10
Participación personal			3	7	4
Participación en grupos		2	1	10	8
Duración del encuentro			3	12	6
Distribución del tiempo		1	3	11	7
Lugar				8	13
Organización				10	11

¿Cómo tratar estadísticamente estos datos?

El problema de esta encuesta es su carácter ordinal, lo que no permite el tratamiento con medidas de tipo paramétrico. Esto se puede solucionar de dos formas:

(A) Se mantiene su carácter ordinal y sólo se usa la Mediana y los Percentiles, con sus medidas asociadas. De esta forma se respecta la estructura ordinal de los datos, pero no se puede profundizar mucho en el análisis. Sería la postura más científica y propia de ambientes universitarios.

(B) Se puede convertir la escala Muy mal, Mal, Regular, Bien, Muy bien en una escala numérica 1, 2, 3, 4 y 5. El inconveniente es que si se usa la media o la desviación típica estaremos asumiendo que la variable es de intervalo, y que tienen sentido las diferencias entre tramos de la escala, hecho que no os consta. Nadie

puede afirmar que para un entrevistado la distancia entre Mal y Regular representa la misma diferencia en apreciación que la existente entre Regular y Bien.

Aquí asumiremos el riesgo de la opción (B), porque es más valioso que nuestros alumnos intenten analizar la tabla que respetar la pureza de la medida.

Ampliación de la tabla

Como en todo estudio estadístico, deberemos proceder a completar la tabla con totales o porcentajes.

CALIFICACIÓN	1	2	3	4	5	TOTAL	MEDIA	DIF. MEDIA
Contenidos			1	9	5	15	4,266666667	0,03
Metodología			3	12	6	21	4,142857143	-0,10
Provecho		1	2	13	6	22	4,090909091	-0,15
Cómo me he sentido			4	9	10	23	4,260869565	0,02
Participación personal			3	7	4	14	4,071428571	-0,17
Participación en grupos		2	1	10	8	21	4,142857143	-0,10
Duración del encuentro			3	12	6	21	4,142857143	-0,10
Distribución del tiempo		1	3	11	7	22	4,090909091	-0,15
Lugar				8	13	21	4,619047619	0,38
Organización				10	11	21	4,523809524	0,29
TOTAL FRECUENCIAS	0	4	20	101	76	201	4,23880597	

Hemos efectuado la traducción numérica a una escala del 1 al 5. Esto nos permitirá realizar cálculos, aunque hemos supuesto, sin ningún fundamento, que los intervalos entre las puntuaciones ordinales son equivalentes.

El primer cálculo que se ha efectuado es el de sumar las frecuencias de cada pregunta, con la función SUMA

CALIFICACIÓN	1	2	3	4	5	TOTAL
Contenidos			1	9	5	15
Metodología			3	12	6	21
Provecho		1	2	13	6	22
Cómo me he sentido			4	9	10	23
Participación personal			3	7	4	14
Participación en grupos		2	1	10	8	21
Duración del encuentro			3	12	6	21
Distribución del tiempo		1	3	11	7	22
Lugar				8	13	21
Organización				10	11	21
TOTAL FRECUENCIAS	0	4	20	101	76	201

Esta suma sólo nos indica que hay preguntas que no han sido respondidas por algunos asistentes.

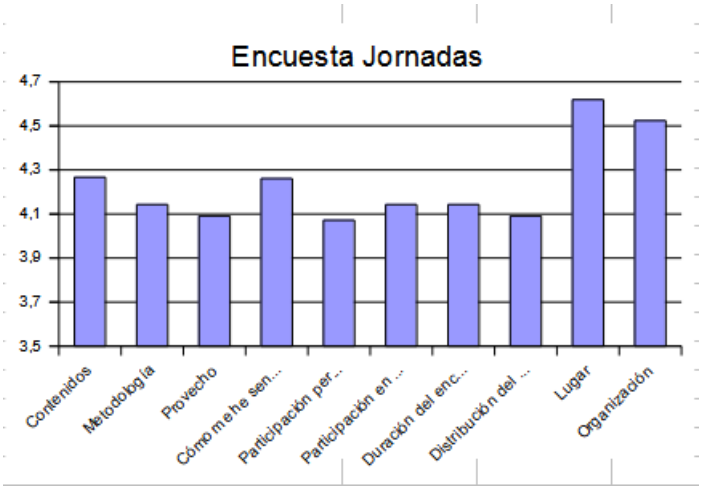
Promedios

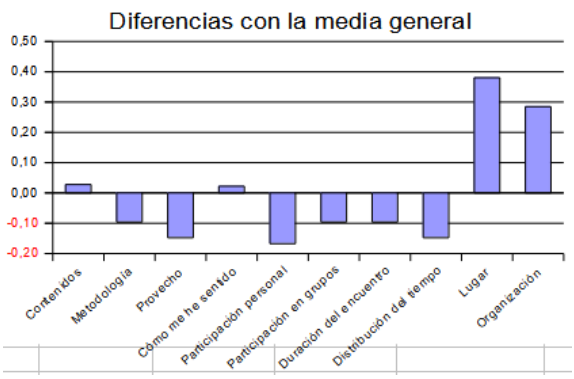
Para calcular la media a partir de las frecuencias deberemos acudir a la media ponderada. Si estudias la columna de medias, en todas ellas se multiplica el valor 1 por la primera frecuencia, el valor 2 por la segunda, y así hasta el 5, y después se suman todos los productos y se divide al final entre la suma de frecuencias. Estúdialo bien.

MEDIA	DIF. MEDIA
4,27	0,03
4,14	-0,10
4,09	-0,15
4,26	0,02
4,07	-0,17
4,14	-0,10
4,14	-0,10
4,09	-0,15
4,62	0,38
4,52	0,29
4,24	

Las medias nos ofrecen la forma de analizar la valoración de cada pregunta. Se ha añadido el cálculo de la media general y las diferencias de cada media con ella. De esta forma se nos aparece claramente qué preguntas se han valorado mejor (color verde) y cuáles peor (en rojo)

Podemos añadir un gráfico de barras para las medias y otro de diferencias con la media general. Con ellos se llega a la conclusión sorprendente de que lo mejor valorado de las jornadas fue el lugar de celebración, algo frustrante para los organizadores.





Estudio por niveles

El estudio de las respuestas según los niveles de la escala suele informarnos bastante bien del ambiente general del encuentro. El siguiente gráfico nos informa

de que el perfil de puntuaciones corresponde a personas con actitud positiva (máximo en el 4).

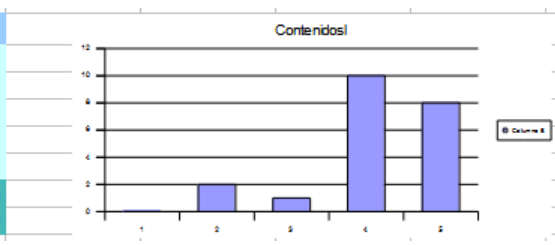
Estudio por nivel de respuesta

Frecuencia de cada respuesta	
1	0
2	4
3	20
4	101
5	76
Total	201

Estudio por preguntas

Puede ser interesante estudiar cada pregunta por separado, para ver si su perfil es semejante al general, o bien estudiar su media y desviación típica. Por ejemplo, el siguiente gráfico estudia la respuesta a “Participación en grupos”:

Participación en grupos	
1	0
2	2
3	1
4	10
5	8
Media	4,142857143
Desv. típ.	0,888321815



MEDIDAS TÍPICAS. ÍNDICES

CUESTIÓN - EJEMPLO

¿QUÉ NIVEL VERDADERO TIENE MI MARTITA?

Carmina y Luis son dos padres angustiados por las notas de sus hijos. En la segunda evaluación, su hija Martita trae una calificación de Bien en tres asignaturas: Informática, Lengua española y Matemáticas. Sin embargo, ella confiesa que la parte de Gramática Española no se le da muy bien, y que en Matemáticas cree que va entre las mejores. Los padres se plantean: ¿En qué zona de la clase se encuentra nuestra hija? ¿Entre los diez mejores? ¿A nivel intermedio? ¿Es de las peores?

Los padres investigan el origen de las tres calificaciones de su Martita, visitando a los profesores, y descubren lo siguiente:

El profesor de Informática califica sumando puntos según los trabajos realizados. La máxima nota ha sido

de 37, y a Martita, por obtener 25, le ha asignado un Bien.

El Bien de Lengua lo ha obtenido por un promedio de 3 en un rango entre 0 y 5.

Por último, en Matemáticas, obtuvo un 6 sobre 10, que también se interpretó como bien.

Los padres comparan la nota de su hija entre el máximo y obtienen esta proporción:

Informática	25/37	67,6%
Lengua española	3/5	60%
Matemáticas	6/10	60%

La cosa parece justa, pero ¿por qué su hija insiste en que le cuesta la Lengua más que las Matemáticas?

Vuelven a hablar con los profesores, insisten y obtienen estas tres preciadas tablas:

Informática - 15 equipos - Notas aisladas							
17	30	22	15	35	28	30	
37	20	25	20	15	28	32	28

Lengua Española - 2ª Evaluación	
Distribución de notas	
0	2
1	1
2	3
3	5
4	12
5	7

Matemáticas		Frec.
Calificaciones y equivalencias		
INS	4	12
SUF	5	10
BIEN	6	5
NOT	7,5	2
SOB	9	1

Así comprenden mejor las cosas, el de Mates es un hueso y ha suspendido a casi todos, luego el Bien de su Martita es muy valioso, sin embargo, en Lengua es de las peores. ¿Cómo podríamos expresar esto estadísticamente? Ese es el objetivo de los índices de posición.

CLASES DE PUNTUACIONES

Las medidas directas que se efectúan sobre una muestra no siempre informan claramente de algunos hechos o propiedades que permanecen ocultos, y que un cambio de escala o el uso de una medida derivada puede destacarlos. Este es el objeto de esta sesión teórica: el uso de medidas derivadas e índices que faciliten el conocimiento de hechos no percibidos en la medida inicial.

MEDIDAS O PUNTUACIONES TÍPICAS

Medida directa

Llamaremos *medida directa* en cualquier estudio o experimento, a aquella que se ha obtenido directamente sobre los objetos, individuos o entidades con los instrumentos usuales de medida.

Así, son medidas directas: la estatura en cm., la edad en años, la producción de una fábrica en toneladas, etc.

Sobre esta medida directa, mediante operaciones matemáticas o de ordenación, se pueden establecer otras medidas *derivadas* que informen del mismo fenómeno destacando otros aspectos. Por ejemplo: la producción de la fábrica en pesetas constantes de 1990, la estatura de una persona en comparación con la del año pasado, la edad comparada con el resto de su colectivo, etc.

La medida directa tiene el defecto de no informarnos sobre la posición o nivel que ese dato tiene dentro de su grupo.

Medida diferencial

Dada una medida directa X, llamaremos ***medida diferencial x*** a su diferencia con la media del grupo:

$$x = X - \bar{X}$$

Si recuerdas la propiedad de la media

$$\sum X = N\bar{X}$$

comprenderás que la suma de las medidas diferenciales será igual a cero y que, además, unas serán positivas y otras negativas

$$\sum x = 0$$

La consecuencia es que **la media de las medidas diferenciales siempre es cero,** y se puede demostrar que la desviación típica *s* de las medidas diferenciales es la misma que la de las medidas directas

La medida diferencial nos informa sobre lo cerca o lejos que se encuentra un dato respecto a la media. Es, por tanto, representativa de la situación del individuo dentro de su grupo, pero no nos permite evaluar si esa distancia es importante o no.

En realidad, es una simple traslación.

Medida típica Z

Si se divide una medida diferencial entre la desviación típica del grupo, se obtiene la *medida o puntuación típica Z*:

$$Z = \frac{X - \bar{X}}{\sigma}$$

Esta medida es muy importante, pues permite comparar dos colectivos distintos, debido a la siguiente propiedad:

La media de las puntuaciones Z siempre es cero y su desviación típica siempre es 1

De esta forma, mediante Z, las medidas obtenidas por cualquier sujeto en variables diferentes, siempre tendrán media 0 y desviación 1, con lo que Z mide **el verdadero nivel** dentro de cada grupo, al haber eliminado los parámetros de centro y dispersión. Es como si dos conjuntos los redujéramos a la misma escala para poderlos comparar.

Las medidas Z comprendidas entre -2 y 2 suponen como mínimo el 75% de los datos. Así, puntuaciones

superiores a 2 o inferiores a -2 son extraordinarias, en el sentido de que lo probable es lo contrario.

De igual forma, entre -3 y 3 están contenidos al menos el 89% de los datos. Puntuaciones más alejadas que 3 y -3 se consideran improbables.

OTRAS MEDIDAS

Hemos visto que la puntuación Z suele estar entre -3 y 3 y, por tanto, puede ser positiva o negativa y, en general, con decimales. Para simplificar su lectura, especialmente en Psicología y Ciencias de la Educación, se han introducido otras medias convencionales. Las más importantes son:

Escala T

La puntuación T se obtiene multiplicando Z por 10 y después sumando 50:

$$T = 10Z + 50$$

Lo normal es que una puntuación T oscile entre 20 y 80 puntos. Las medidas más extremas son improbables.

ÍNDICES DE POSICIÓN

CUANTILES

Si recordamos que la mediana es un punto convencional en una distribución de frecuencias caracterizado por la propiedad de tener a la mitad de los datos inferior a él, ¿no podríamos considerar también un punto que tuviera sólo un 25% inferior? ¿o el punto que sólo tiene el 10% de los datos superior a él? Estas consideraciones han llevado a la idea de **cuantil**:

Diremos que un número es el **cuantil de orden p** en una distribución de frecuencias si el porcentaje de datos inferiores a él es igual a **p** (y los superiores 100-p).

Por ejemplo, el cuantil C_{85} será un punto que cumple que el 85% de los datos sea inferior o igual a él.

Cálculo con datos aislados

Si los datos están aislados, se calcula el **p%** del total de datos y se va contando de menor a mayor hasta llegar a ese número de datos:

En el conjunto 1 1 1 2 2 3 3 3 3 4 4 5 6 6 7 8 9 9 de 18 datos ¿dónde se encuentra el cuantil C_{60}? Calculamos el 60% de 18, que son 10,8 datos, redondeando, 11 datos. Contamos, pues, los 11 primeros y llegamos al 4, luego **el número 4 es el cuantil del 60%**.

Cálculo con datos agrupados

Su fórmula es similar a la de la mediana:

$$C_p = L_i + \frac{N.p - n_{ant}}{n_{interv}}.Ampl.$$

En la fórmula los símbolos se interpretan así:

N.p es el producto del número de datos por la proporción que define el cuantil (si es porcentaje, se deberá después dividir entre 100)

Ampl. es la amplitud del intervalo contiene la frecuencia acumulada **p**

n_{ant} es la frecuencia acumulada anterior al intervalo del cuantil

$n_{interv.}$ es la frecuencia absoluta de dicho intervalo.

L_i representa el límite inferior verdadero del intervalo del cuantil.

Cuartiles

Los cuantiles que dividen a la distribución **en cuatro partes iguales**, es decir, C_{25}, C_{50} y C_{75}, reciben el nombre de **cuartiles**, y se representan por

Q1 o primer cuartil es el número que deja inferiores a él un 25% de los datos.

Q2 o segundo cuartil o mediana es el número que deja inferiores a él un 50% de los datos.

Q3 o tercer cuartil es el número que deja inferiores a él un 75% de los datos.

Su cálculo se efectúa como en los demás cuantiles.

Los cuartiles son muy interesantes como alternativa a las medidas de variabilidad y asimetría:

La variabilidad de una distribución (ver sesión 2) se puede medir también mediante el

Semiintervalo intercuartílico

Esta medida se obtiene mediante la fórmula

$$I = \frac{Q_3 - Q_1}{2}$$

y es una buena alternativa a la desviación típica si los datos son de tipo ordinal, si existen intervalos abiertos o si la distribución es muy asimétrica.

Deciles

Se suelen definir **9 deciles D1, D2, ... D9**, que son los puntos que dividen al intervalo en **diez partes iguales**, correspondientes a los cuantiles de porcentajes 10%, 20%, ...90% respectivamente.

Percentiles (o centiles)

Similares a los anteriores, **P1, P2, P3, ...P99**, son 99 números que dividen la distribución en 100 pàrtes iguales. Son muy usados en Ciencias del Comportamiento.

Quintiles

Son menos usados: equivalen a los percentiles 20, 40, 60 y 80, y dividen el conjunto en cinco partes iguales.

Rango percentil

Es la medida inversa del **percentil**. Dada una medida concreta, como puede ser la calificación de una alumna en Música, su rango percentil equivale al **percentil más cercano** a esa calificación. Un alumno que tenga rango percentil de 78 es aquel en el que el 78% de sus compañeros tiene una puntuación inferior a él.

Para calcular el rango percentil de una medida cualquiera en datos aislados basta contar los inferiores a él, dividir ese número entre el total, multiplicar por 100 y redondear:

Si los datos están agrupados, la fórmula para calcular el rango percentil es:

$$RP = F_{ant.} + \frac{X - L_i}{Ampl.} \cdot f_{int,}$$

donde ***RP*** es el rango percentil, ***X*** la medida directa***,*** $f_{ant.}$ la frecuencia relativa anterior al intervalo que

contiene a X, L_i el límite inferior verdadero de ese intervalo, ***Ampl.*** la amplitud del mismo y f_{int}. su frecuencia relativa

NÚMEROS ÍNDICES

ESTUDIO DE SERIES TEMPORALES O ESPACIALES

En muchos estudios estadísticos nos encontramos con series de muchos datos consecutivos, que suelen estar repartidos temporalmente (por días, meses o años) o, más raramente, de forma espacial (los distintos grupos de 2º de ESO que existen en un Centro), de los que deseamos averiguar qué variación suponen unos respecto a otros. Por ejemplo:

¿Con 100 ptas. del año 1980, qué compraría hoy? ¿Cómo ha ido creciendo la población de mi pueblo en estos últimos años?

¿Ha habido algún progreso en este grupo de alumnos desde Octubre hasta Mayo?

Los números índices se usan por varias razones:

Permiten comparar unos valores con otros en un serie de forma porcentual.

Añaden claridad a algunos procesos.

Con ellos se pueden comparar series de distinta índole.

Su propiedad multiplicativa simplifica algunos cálculos.

Índice simple de base fija

Un término de la serie se identifica (convencionalmente) con el número 1, o el 100%. Diremos que este valor $\mathbf{y_0}$ posee el índice 1. Para el resto de valores se define el índice como **el cociente entre su propio valor $\mathbf{y_i}$ y el valor $\mathbf{y_0}$ identificado como de índice 1**. Eventualmente, multiplicaremos por 100 si deseamos expresarlo como porcentaje, pero en los cálculos se deja así.

$$I_0 = \frac{y_i}{y_0}.100$$

Índice simple de base variable (o en cadena)

Tiene la misma definición que el anterior, pero en lugar de elegir un valor $\mathbf{y_0}$ como base, en el cociente se toma el término anterior y_{i-1}.

$$I_i = \frac{y_i}{y_{i-1}}.100$$

La propiedad fundamental es la siguiente:

El índice definido entre dos términos de una serie no consecutivos equivale al producto de todos los índices en cadena intermedios.

Así, si el coste de vida en el 1993 fue en un país del 3%, en 1994 del 5% y en 1995 del 2%, los índices en cadena serían 1.03, 1.05 y 1.02 respectivamente. Para calcular el incremento habido entre 1993 y 1996, deberíamos multiplicar los tres índices, y nos resultaría 1,10313, es decir, un 10,31%. Si hubiéramos sumado 3+5+2 = 10, se hubiera producido un error.

Los índices en cadena son multiplicativos y no aditivos.

Índice compuesto

Cuando se desea comparar la evolución de varios conjuntos a la vez, se definen *índices compuestos*, obtenidos combinando los índices simples. En general es una operación compleja (recuérdese el índice de coste de la vida). Una técnica sencilla es sustituir los múltiples valores de cada término por su media ponderada.

CONCENTRACIÓN. ÍNDICE DE GINI

Cuando una serie de datos representa una magnitud acumulable (dinero, producción, etc.) nos podemos plantear si esa magnitud está bien repartida o no. Casos típicos son el reparto de salarios en una empresa, la producción industrial de los países o el reparto de la riqueza en el mundo. La mayor o menor equidad en el reparto viene representada por la **concentración**. Esta es una magnitud convencional que se mide por el índice de Gini. No vamos a profundizar en este concepto, pero es bueno incluirlo en este curso simplemente para que se conozca su

existencia y facilitar la profundización en el tema a quien lo desee.

Para estudiar la concentración o equidad debemos definir dos índices P y Q.

P representa las frecuencias relativas acumuladas de la tabla que estemos estudiando.

Q equivale a la acumulación relativa de los productos de cada dato X por su frecuencia. Si P es el número de individuos, Q es la acumulación de la magnitud entre todos los individuos. Por ejemplo:

Dos profesores de un colegio presentan las siguientes distribuciones de notas, si se traducen INS,SUF,BIEN, etc por 1,2,3,4,5

Calificación	Profesora A	Profesor B
1	5	10
2	15	5
3	17	5
4	12	5
5	1	25

¿Cuál de los dos profesores ha repartido sus calificaciones de manera más uniforme?

En la profesora A las P y Q tienen los valores

Profesora A	P (individuos)	Q (indiv. por notas)
5	0,10	0,04
15	0,40	0,25
17	0,74	0,62
12	0,98	0,96
1	1	1

y en el profesor B

Profesor B	P (individuos)	Q (indiv. por notas)
10	0,20	0,06
5	0,30	0,11
5	0,40	0,19
5	0,50	0,31
25	1	1

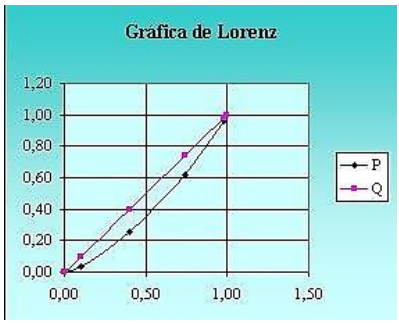
Se ve que en el profesor B la masa de notas Q crece muy poco a poco, mucho menos que en A. Además, en esta profesora las P y Q se parecen más. Parece que ella reparte mejor las calificaciones, con más equidad. En el profesor B los sobresalientes son demasiados.

Estadísticamente esta propiedad de concentración se mide con el índice de Gini, definido por

$$IG = \frac{\sum(p - q)}{\sum p}$$

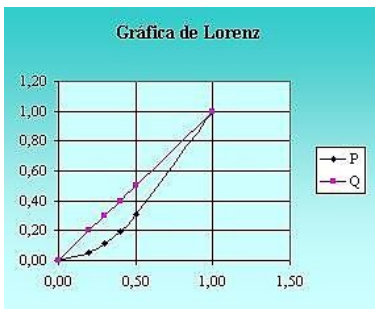
Este índice está comprendido entre 0 (distribución equitativa) y 1 (distribución totalmente desequilibrada).

Abre el modelo ***concentra.ods*** y rellena las dos primeras columnas con las frecuencias de la profesora A: 5, 15, 17... y borra lo sobrante por abajo con la tecla **Supr**. Pasa a la segunda hoja de Resultados y leerás el índice de Gini para ella, que será de 0,157, bastante bajo. Si observas la curva de Lorenz de esa hoja, observarás que compara P y Q. La P viene representada por la línea recta y la Q por la curva. No hay demasiadas diferencias entre ellas.



Cambia ahora las frecuencias por las del profesor B.

Ahora el índice es mucho mayor: 0,524 y la curva de Lorenz más curvada.



Luego el profesor reparte sus calificaciones de una forma menos equitativa que la profesora.

UN CASO PRÁCTICO: CREACIÓN DE UN PERFIL

Una profesora desea estudiar las capacidades de su alumnado ante la formación de grupos de trabajo. Desea crear perfiles respecto a las variables de extraversión, adaptabilidad, capacidad de liderazgo, insatisfacción y sociabilidad. Para ello realiza cinco pruebas a todos los alumnos y alumnas de la clase, pero se encuentra con el inconveniente de que cada prueba usa una escala distinta. ¿Cómo podría unificarlas todas a fin de crear gráficos que representen los perfiles que ella desea?

Para mayor simplicidad la profesora ha rotulado las variables con las letras A, B, C, D y E. Después de pasadas las pruebas ha eliminado a quienes hubieran faltado a algunas de ellas y por motivos de privacidad ha representado a sus alumnos y alumnas con un número de orden. La tabla resultante, de la que hemos añadido aquí una copia parcial, la ha almacenado en un archivo.

	A	B	C	D	E
1	5	7	10	11	4
2	4	7	25	14	4
3	4	6	25	17	3
4	5	6	30	9	5
5	3	4	48	9	2

Observando la tabla se descubre que unas variables se han medido según una escala del 1 al 5, mientras otras llegan al 10, 20 e incluso 100

¿Cómo tratar estadísticamente estos datos?

Debemos unificar las medidas, a fin de que el perfil sea representativo del grado que cada persona alcanza en las cinco variables. No podemos usar las medidas típicas, porque no nos consta que las variables sean de intervalo. Es más, dado su carácter de variables de tipo psicológico, hemos de suponer que las medidas son de tipo ordinal.

Según las consideraciones anteriores, lo más adecuado sería acudir al **Rango Percentil**, que sitúa cada medida

dentro de su grupo. Esto es importante, porque los perfiles que obtenga la profesora no se podrán exportar fuera del grupo, pues perderían su sentido. No son magnitudes absolutas de los individuos, sino relativas a su grupo.

Para calcular los rangos percentiles de los distintos sujetos se ha usado la variable **BUSCARV** de LibreOffice Calc y Excel, que permite elegir unos datos que estén situados en la misma fila que uno dado. Así, en su hoja de cálculo se pregunta por el número de orden (se podría haber organizado por apellidos o por otras claves) y a partir de él se consigue la lista de las cinco puntuaciones del sujeto.

Escribe el número de orden				
	8			
Medidas directas				
A	B	C	D	E
2	3	45	14	1

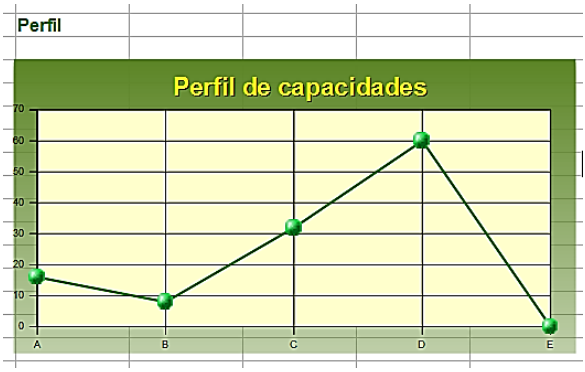
Una vez obtenidas las medidas directas, se aplica sobre cada una la función RANGO.PERCENTIL. Aunque esta función de valores 0 y 100 en algunos casos, en contra de lo que es usual, para nuestro caso cumple perfectamente, pues sitúa en una escala relativa del 0

al 100 las medidas de los sujetos, según el porcentaje de medidas inferiores a la dada.

Rangos percentiles

A	B	C	D	E
16	8	32	60	0

Por último, a partir de esta tabla de rangos percentiles se construye un gráfico que representa el perfil de variables de cada sujeto respecto a su integración en los grupos. Normalmente la profesora, a la vista de los 26 perfiles podrá diseñar la composición de los grupos.



DISTRIBUCIONES BIDIMENSIONALES. CORRELACIÓN.

CUESTIÓN - EJEMPLO

¿Influye la primavera en estos chavales?

La Directora de un centro está preocupada por el incremento de faltas graves que se ha producido en los primeros meses del año. Tiene a su cargo tres niveles de enseñanza A, B y C, y el seguimiento de las faltas graves lo ha resumido en la siguiente tabla:

Meses Niveles	Enero	Febrero	Marzo	Abril	Mayo
A	4	6	7	8	8
B	3	3	6	5	9
C	9	7	7	13	14

¿Cómo podría estudiar bien estos datos?

¿Son independientes la distribución de faltas, el nivel de enseñanza y los meses?

¿Qué medidas podríamos usar?

Estamos ante un caso de distribución bidimensional, porque cada falta grave se representa por dos medidas: el mes y el nivel. Siempre que las observaciones comporten dos medidas distintas para cada sujeto estaremos ante una variable bidimensional. Cada medida pertenecerá a una variable distinta, y esta puede ser nominal, cuantitativa, etc. Por ejemplo, en este caso se trata dos variables nominales, los meses y las letras A, B y C.

En esta tabla figuran frecuencias absolutas, pero veremos que también puede haber tablas con frecuencias relativas o porcentajes. Por ser una tabla de doble entrada, se le suele llamar también **Tabla conjunta de frecuencias**

A continuación se explican los distintos tipos de tablas de frecuencias que se pueden usar en estos casos.

DISTRIBUCIONES BIDIMENSIONALES

En algunos experimentos las medidas que se obtienen son dobles, pertenecientes a dos variables distintas, a las que llamaremos X e Y respectivamente.

Este tipo de estudios es muy frecuente. Daremos algunos ejemplos:

- Comparación entre mortalidad y natalidad
- Ídem entre extensión y población de diversos países.
- Diferencias de renta entre la población en general y los titulados universitarios.
- Pruebas *pretest* y *postest*.
- Influencia de la latitud en la temperatura media.
- Ídem de las horas de estudio en la calificación en una asignatura.
- Etc.

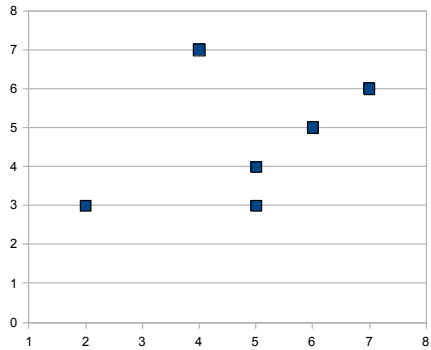
Tipos de variables

Las dos variables que se comparan pueden ser de igual naturaleza, ambas nominales u ordinales o de intervalo, o de distinta, lo que da lugar a muchos casos posibles, que es imposible estudiarlos todos en este curso.
Incluimos algunos ejemplos:

Tablas simples de comparación de dos datos cuantitativos

Alumnos	X: Examen de Geografía	Y: Examen de Matemáticas
Julia	4	7
Pedro	6	5
Miguel	5	4
Marta	2	3
.....	.....	.....

En estos casos cada par de valores representa a un sujeto o medición. Se representan mediante gráficos de dispersión XY



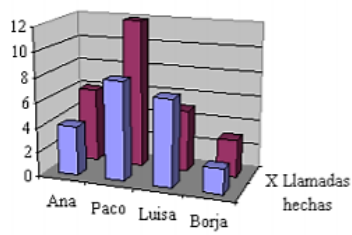
Tablas de doble entrada

En ellas la X y la Y pueden ser de naturaleza muy distinta, por lo que se disponen en tabla de doble entrada. Cuando existen frecuencias, es el mejor método, pues permite tratar una variable por columnas y otra por filas.

La siguiente tabla muestra la distribución de las llamadas telefónicas con origen o destino en los cuatro hijos de una pareja.

	X Llamadas hechas	Y Llamadas recibidas
Ana	4	6
Paco	8	12
Luisa	7	5
Borja	2	3

Estas tablas de doble entrada con frecuencias admiten una representación gráfica muy intuitiva mediante barras (columnas) ordenadas en varios conjuntos mediante tres ejes.



TIPOS DE FRECUENCIAS

Para aclarar las definiciones de los tipos de frecuencias usaremos la siguiente tabla:

Meses Niveles	Enero	Febrero	Marzo	Abril	Mayo
A	4	6	7	8	8
B	3	3	6	5	9
C	9	7	7	13	14

FRECUENCIAS CONJUNTAS

Se representan por n_{ij}, y son las frecuencias incluidas en la tabla primitiva de entrada. Los subíndices ***i*** y ***j*** representan la fila y columna en la que está situada la frecuencia.Así, en la tabla $n_{13} = 7$ y $n_{34} = 13$

Llamaremos N a la suma total de estas frecuencias. En el ejemplo, N es 109.

Representaremos este hecho mediante un sumatorio doble sin índices, para no complicar las fórmulas:

$$\sum\sum n_{ij} = N$$

Al conjunto de las frecuencias conjuntas lo denominaremos como **Distribución conjunta de las dos variables**.

FRECUENCIAS MARGINALES

Llamaremos ***frecuencia marginal*** de un valor de X, a la que le corresponde a ese valor si no tenemos en cuenta la existencia de Y. En la práctica coincide con la suma de todas las frecuencias contenidas en **la fila correspondiente a ese valor**.

En la tabla del ejemplo, la frecuencia marginal de B es 26, suma de las frecuencias de la segunda fila. La frecuencia marginal de la fila ***i*** se representará por n_{i*}
De la misma forma se define la frecuencia marginal en la variable Y, como la que tendría si no se tuviera en cuenta la X, o la suma de la columna correspondiente. En el ejemplo, la frecuencia marginal de Marzo es n_{*3} = 20

FRECUENCIAS CONDICIONADAS

Son las frecuencias que posee una variable si sólo consideramos **un valor** (o varios) de la otra variable. En la práctica se traduce a considerar sólo una fila o sólo una columna, según el valor elegido.

Las frecuencias condicionadas se representan con este símbolo: $\mathbf{n_{x/y}}$, que se puede leer como *Frecuencia de x condicionada por y*.

En la tabla del ejemplo, la distribución de X condicionada a Marzo es la columna A=7, B=6, C=7. Las frecuencias condicionadas son más representativas si se convierten en proporciones o porcentajes.

MEDIDAS EN UNA DISTRIBUCIÓN BIDIMENSIONAL

Al existir dos variables X e Y, las medidas también son dobles. Así, consideraremos las siguientes:

Media de X

Tiene la misma definición que en el caso unidimensional. Viene dada por la fórmula

$$\bar{x} = \frac{\sum x}{N}$$

si los datos están aislados y por esta otra

$$\bar{x} = \frac{\sum x \cdot n}{N} = \sum x \cdot f$$

si están agrupados.

Media de la Y

Se define de forma similar:

$$\bar{y} = \frac{\sum y}{N}$$

y para agrupados

$$\bar{y} = \frac{\sum y \cdot n}{N} = \sum y \cdot f$$

(Las siguientes definiciones las desarrollaremos sólo para aislados, pues su traducción es fácil)

Varianzas y desviaciones típicas

También serán dobles:
La varianza de X será

$$s_x^2 = \frac{\sum(x - \bar{x})^2}{N} = \frac{\sum x^2}{N} - \bar{x}^2$$

y su desviación típica $\mathbf{s_x}$ será la raíz cuadrada de esa expresión.

En el caso de Y la definición es similar:

$$s_y^2 = \frac{\sum(y - \bar{y})^2}{N} = \frac{\sum y^2}{N} - \bar{y}^2$$

Covarianza

Esta medida es muy interesante. Mide el *paralelismo* existente entre ambas variables (en función **sólo** de los datos presentes en la tabla). Si la covarianza es grande, manifestará la existencia de un cierto paralelismo o dependencia (en sentido estadístico) entre X e Y. Si es pequeña, indicará que ambas variables se comportan de manera más independiente.

Su definición es:

$$s_{xy} = \frac{\sum(x - \bar{x})(y - \bar{y})}{\boldsymbol{N}} = \frac{\sum xy}{\boldsymbol{N}} - \overline{xy}$$

y puede ser positiva, cero o negativa.

El significado de la varianza es el siguiente:

Si en el numerador la mayoría de los productos son positivos, será porque las diferencias de X y de Y tienen el mismo signo. Eso significa que para X mayor que la media, la Y también lo es, y al contrario, a valores pequeños de X le corresponden pequeños en Y. Por tanto, los productos serán mayoritariamente positivos y la varianza crecerá.

Una varianza positiva y alejada del valor cero indica un cierto paralelismo entre X e Y, en el que a valores mayores de X le corresponden los mayores en Y.

Si los productos son mayoritariamente negativos, es que las diferencias tienen distintos signos, por lo que
Una varianza negativa y alejada del cero indica un paralelismo inverso, en el que a valores pequeños de X le corresponden valores grandes de Y, y a la inversa.

Por último, si están muy repartidos los productos positivos y negativos, es que apenas existe paralelismo, y la varianza se acercará a cero.

El problema de la varianza es que carece de un valor máximo, por lo que es difícil juzgar si la correspondencia entre las dos variables es la mejor posible.

Coeficiente de correlación

Como en el caso de una variable, la *covarianza* no es adecuada para establecer comparaciones entre medidas muy diferentes, además del inconveniente de no tener un valor máximo, lo que impide valorar el grado de paralelismo existente en los datos.

Para normalizar la covarianza procederemos como en el Coeficiente de Variación: dividiremos dicha covarianza entre las dos desviaciones típicas (de X y de Y respectivamente). Al resultado le daremos el nombre de *Coeficiente de correlación* y lo representaremos por **r**.

$$r = \frac{s_{xy}}{s_x s_y}$$

El coeficiente **r** también recibe el nombre de ***Coeficiente de Pearson*** o también ***Coeficiente de correlación producto-momento***.

También se puede demostrar que este coeficiente es en realidad la covarianza del conjunto si expresamos los datos en medidas típicas ***z*** (ver sesión 3).

El valor de **r** oscila entre -1 y +1, y mide el paralelismo o *correlación* entre X e Y. Si sus valores se acercan a 1 o a -1, diremos que existe correlación **fuerte**, y está cerca del cero, **débil**.

Podemos desarrollar más estos comentarios mediante una tabla:

Valor de r	Comentario
+1	Dependencia funcional positiva (función creciente entre ambas)
Cercana al 1	Correlación fuerte positiva
Cercana al 0	Correlación débil o independencia

Cercana al -1	Correlación fuerte negativa
-1	Dependencia funcional negativa (función decreciente)

Se deben evitar interpretaciones erróneas del coeficiente **r**. Seleccionamos las más frecuentes:

La dependencia es sólo matemática: no supone relación causa-efecto. Las causas nunca son tan simples y pueden existir, pero respecto a una tercera variable.

Se deben evitar demasiados adjetivos como *correlación regular, media, ...* pues el significado exacto de **r** depende de cada experimento en concreto.

Si la relación entre datos es de tipo curvilíneo, el coeficiente **r** pierde representatividad.

A veces, si existe asimetría, **r** no puede acercarse al 1.

OTRAS MEDIDAS DE CORRELACIÓN

El coeficiente de correlación de Pearson exige que la escala de medida sea de intervalo o razón. Cuando este supuesto no se cumple, deberemos usar otros coeficientes, aunque muchos de ellos equivalen, en sus cálculos, al coeficiente de Pearson.

Coeficiente de Spearman o de rangos

Si la variable es de tipo ordinal, podemos usar los rangos (número de orden de cada dato) para evaluar la correlación.

Representaremos por **d** a la diferencia entre rangos que presenta un dato en dos ordenaciones distintas. Por ejemplo, supongamos que diez individuos han sido ordenados de forma diferente por dos evaluadores A y B:

Individuos	1	2	3	4	5	6	7	8	9	10
A	2	3	4	1	5	9	10	8	7	6
B	3	5	1	4	2	6	8	10	9	7
D	+1	+2	-3	+3	-3	-3	-2	+2	+2	+1

La suma de todas las diferencias será cero.

La fórmula del coeficiente de Spearman es

$$\rho = 1 - \frac{6 \sum d^2}{N(N^2 - 1)}$$

Si existen empates entre ordenaciones se resuelven asignando el rango promedio.

Equivale al coeficiente de Pearson, aunque se calcule mediante otras técnicas. Un coeficiente positivo significará que rangos altos en una de las variables se corresponderán con rangos también altos en la otra, y negativo cuando a los altos en una correspondan bajos en otra.

Coeficiente biserial puntual

Se utiliza cuando X es cuantitativa y la Y dicotómica (variable con sólo dos valores). Por ejemplo, X puede ser la calificación en un examen de Ciencias Sociales, y la Y el hecho de que los alumnos examinados tengan o

no una habitación para estudiar ellos solos, sin compartirla con los hermanos.

Los valores de la variable Y se suelen representar por 1 y 0. Puede ser dicotómica en su definición (tener o no tener, aprobar o suspender, ...), o bien haber sido *dicotomizada*, si, por ejemplo, asignamos un 1 a los individuos que superen un valor y 0 a los que no lo superen.

La fórmula de este coeficiente es:

$$r_{bp} = \frac{\overline{X_1} - \bar{X}_0}{s_x} \cdot \sqrt{pq}$$

donde las medias del numerador corresponden a los valores correspondientes a Y=1 e Y=0 respectivamente, la desviación típica del denominador a la de **todas las X**, y los valores **p** y **q** a las proporciones de sujetos con Y=1 e Y=0 respectivamente.

En la siguiente tabla presentamos un ejemplo de situación en la que es aplicable este coeficiente:
X: Notas en el examen de Ciencias Sociales.

Y: Disposición de habitación de estudio individual, representada por 0 y 1.

X	9	5	4	8	6	9	8	6	6	7
Y	1	1	0	0	0	1	1	0	1	0

Coeficiente de contingencia

Se utiliza para tablas de doble entrada que contengan frecuencias correspondientes a dos variables de cualquier tipo de escala, desde nominal hasta cuantitativa de razón.

Usa la distribución chi-cuadrado χ^2, que se estudia en otra sesión del curso.

Su fórmula es

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

PRUEBA DE INDEPENDENCIA

La Prueba de Independencia o Test de Homogeneidad investiga si existe un buen grado de asociación entre dos variables que se estudian conjuntamente, o bien son independientes. Usa la distribución chi-cuadrado.

La idea en la que se basa es muy simple, pero no la desarrollaremos aquí. Se considera que si dos variables son independientes, los valores de una no influirán en los de la otra, es decir, las probabilidades condicionadas serán siempre las mismas. Eso se traduce en la práctica en una proporcionalidad en las frecuencias de la tabla. Sobre esa idea se construyen unas frecuencias teóricas y se comparan con las reales. En los textos de Estadística podrás leer todo el desarrollo completo.

Volvemos al ejemplo de las faltas graves de los alumnos, según la tabla

Meses Niveles	Enero	Febrero	Marzo	Abril	Mayo
A	4	6	7	8	8
B	3	3	6	5	9
C	9	7	7	13	14

¿Que querría decir la afirmación de que *los factores nivel y mes son independientes?* Pues que la distribución de faltas (por ejemplo, en porcentaje) debería ser similar en todas las columnas, *independientemente* del mes en el que se tomen. Con más claridad, los porcentajes que obtuvimos

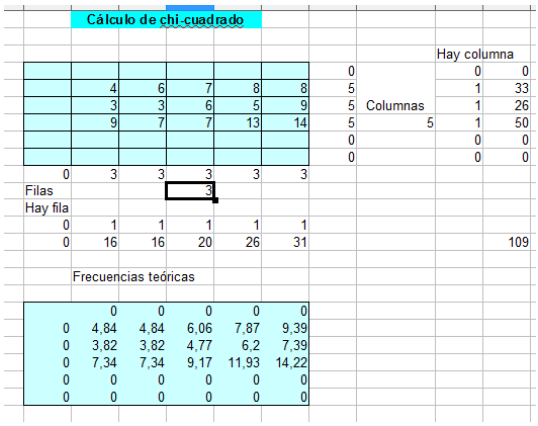
Meses Niveles	Enero	Febrero	Marzo	Abril	Mayo
A	12,1%	18,2%	21,2%	24,2%	24,2%
B	11,5%	11,5%	23,1%	19,2%	34,6%
C	18,0%	14,0%	14,0%	26,0%	28,0%

deberían haber sido los mismos independientemente del mes que consideremos.

Si esto fuera así, ***todas las frecuencias deberían ser proporcionales*** y cada una debería poderse calcular mediante proporciones, o la clásica regla de tres. Por ejemplo, la frecuencia del nivel A en el mes de Febrero debería poder calcularse como el producto del total de A por el total de Febrero, dividido después entre el total de todos los alumnos. Comprueba que resultaría esta tabla *teórica*.

Meses Niveles	Enero	Febrero	Marzo	Abril	Mayo	
A	4,84	4,84	6,06	7,87	9,39	33
B	3,82	3,82	4,77	6,20	7,39	26
C	7,34	7,34	9,17	11,93	14,22	50
	16	16	20	26	31	109

Comprueba algún valor: la frecuencia 6,06 ha resultado de multiplicar 33 por 20 y dividir después entre 109, la frecuencia 11,93 es el resultado de 50*26/109.



Cálculo de chi-cuadrado

						0		Hay columna	
								0	0
	4	6	7	8	8	5		1	33
	3	3	6	5	9	5	Columnas	1	26
	9	7	7	13	14	5	5	1	50
						0		0	0
						0		0	0
0	3	3	3	3	3				
Filas			3						
Hay fila									
0	1	1	1	1	1				
0	16	16	20	26	31				109

Frecuencias teóricas

	0	0	0	0	0
0	4,84	4,84	6,06	7,87	9,39
0	3,82	3,82	4,77	6,2	7,39
0	7,34	7,34	9,17	11,93	14,22
0	0	0	0	0	0
0	0	0	0	0	0

Resultado de la prueba de homogeneidad

Los datos forman un rango de	3	filas
	5	columnas
Sus grados de libertad son:	8	
Valor correspondiente de chi-cuadrado:	3,03	
Probabilidad correspondiente:	0,93	
Decisión al 5%	**Muestras homogéneas**	
Valor crítico al 5%	15,51	
Ídem al 1%	20,09	

Si trabajamos al nivel de significación del 5%, no podemos rechazar la homogeneidad de las frecuencias, no tenemos motivos para pensar que la influencia de la primavera es distinta en cada nivel. Es muy fácil que se presenten estas diferencias: tienen un 93% de posibilidades. Así que, aunque en la primavera se han incrementado las faltas, lo han hecho por igual en los tres niveles, de forma que las observadas se pueden deber al azar.

CUESTIÓN – EJEMPLO

¿Tendré que estudiar mucho para sacar notable?

Un grupo de Enseñanza Secundaria ha elaborado una encuesta sobre las horas diarias que emplean en el estudio y la calificación obtenida en Matemáticas en el último examen.

Han recogido los resultados en la siguiente tabla:

Horas de estudio	0	0	1	1	1	1	1	2	2	2	2	2	3	4	4	5
Calificación	2	1	3	4	3	2	2	4	5	7	8	6	5	8	10	7

Además de estudiar el grado de asociación entre las dos variables, que ya se explicó en el tema anterior mediante el coeficiente de correlación, nos puede interesar hacer pronósticos: *¿Qué nota puedo esperar si estudio 2 horas y meda?* Para realizar esos pronósticos usaremos las técnicas de Regresión.

RECTA DE REGRESIÓN

El problema de la Regresión Lineal es el más importante, junto con la Correlación, que podemos considerar en las distribuciones bidimensionales. Nos ceñiremos a los casos de tablas simples que contengan pares de valores de X e Y, sin consideración de frecuencias. Supondremos también que los datos son de tipo cuantitativo.

Quizás debas repasar los conceptos del Tema 4, en el que se explican la variables bidimensionales.

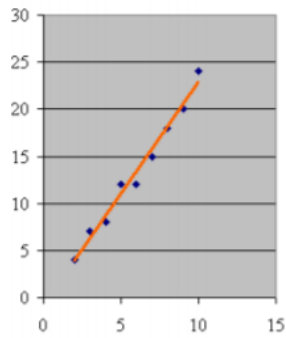
Llamaremos **Recta de Regresión de Y sobre X** a aquella que *mejor se adapta* al diagrama de dispersión XY, también llamado *Nube de puntos*. Este acercamiento se define de forma rigurosa como

La recta de regresión de Y sobre X es aquella que minimiza la suma de cuadrados de las diferencias entre los valores de Y y los correspondientes Y' medidos en dicha recta.

Así, la tabla

X	2	3	4	5	6	7	8	9	10
Y	4	7	8	12	12	15	18	20	24

da lugar a una nube de puntos en cuya gráfica hemos añadido la recta de regresión:



Efectivamente, esta recta sigue *lo mejor posible* la tendencia de los puntos. Matemáticamente, las diferencias al cuadrado de los valores verdaderos de Y y los incluidos en la recta, suman lo mínimo posible.

A la variable X se le suele llamar *predictora*, y a la Y, *criterio*.

La recta de regresión es un instrumento para efectuar predicciones, ya sea en el rango de datos como fuera de él. Llamaremos ***pronóstico o predicción*** para un valor de X a su imagen Y´en la recta de regresión.

La recta de regresión tiene una validez limitada. No debemos efectuar predicciones en valores de X muy alejados del rango considerado. Además, no todas las relaciones son de tipo ***lineal***.

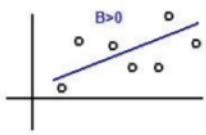
El origen de la palabra *regresión* es histórico. Cuando Galton estudió estas cuestiones, descubrió que los hijos de padres muy altos o muy bajos no lo eran tanto como sus padres, *regresaban* a valores medios. Después se vio que este fenómeno no siempre se daba.

Recordemos que la ecuación de una línea recta en dos dimensiones tiene la forma:

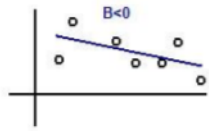
$Y' = A + BX$

donde el coeficiente B representa la tasa de cambio o ***pendiente*** y el coeficiente A es el valor correspondiente a X=0, y la llamaremos ***ordenada en el origen.***

Según el signo de la pendiente, hablaremos de relación ***positiva o creciente***



y de relación ***negativa o decreciente***.



Mediante las técnicas de búsqueda de mínimos podemos demostrar que la recta que minimiza los cuadrados de los errores es la que viene dada por estas fórmulas:

$$B = \frac{S_{xy}}{S_x^2}$$

es decir, la *covarianza* dividida entre la ***varianza de X***

$$A = \bar{Y} - B\bar{X}$$

que podemos expresar como la diferencia entre la media de Y y la de X multiplicada por B

Existen desarrollos simplificados de estas fórmulas para facilitar su cálculo, que no consideraremos aquí.

También se puede considerar la recta de X sobre Y, pero no lo haremos aquí.

Sí puede ser interesante estudiar la recta de regresión para puntuaciones típicas, porque en ese caso su fórmula es muy sencilla: $Z_y = R_{xy}. Z_x$, donde R es el coeficiente de correlación.

PREDICCIONES

Llamaremos ***predicción o pronóstico*** para un valor de X al dado por la expresión Y´= A + BX.

En los gráficos de dispersión XY las predicciones pertenecerán a la línea recta, mientras los valores reales Y figurarán más arriba o abajo de ella.

Llamaremos ***error de predicción*** a la diferencia Y – Y´

El promedio de las predicciones Y´ coincide con el de los valores reales Y.

El promedio de los errores de predicción cometidos es cero.

VARIANZAS EN LA REGRESIÓN

En el problema de la regresión es conveniente considerar distintas sumas de cuadrados para calcular varianzas también distintas:

Varianza total de Y: Si no consideramos la recta de regresión y deseamos medir la variabilidad del conjunto de datos que estamos usando, acudiremos a la varianza de Y, o ***varianza total***.

$$S_y^2 = \frac{\sum(Y - \bar{Y})^2}{N}$$

que es la varianza en el sentido general.

Si sólo consideramos la variabilidad que presentan las predicciones (los valores situados en la recta), deberemos usar en la fórmula anterior los datos Y´ en lugar de Y (la media no cambia, según se indicó más arriba). Al resultado le llamaremos ***varianza explicada***.

$$S_{exp}^2 = \frac{\sum(Y' - \bar{Y})^2}{N}$$

Un resultado fundamental es el siguiente:

$$S_{exp}^2 = S_y^2 \cdot r^2$$

siendo ***r*** el ***coeficiente de correlación de Pearson***. Esto significa que la relación entre la varianza explicada y la total es el cuadrado del coeficiente ***r***. A este cuadrado lo conocemos como ***Coeficiente de determinación*** y expresa el porcentaje de varianza que explica la línea recta.

Por último, llamaremos ***varianza de error o residual*** a la que presentan los valores de Y comparados con sus pronósticos:

$$S_e^2 = \frac{\sum(Y - Y')^2}{N}$$

Se puede demostrar la relación:

$$S_e^2 = S_y^2 - S_{exp}^2 = S_y^2(1 - r^2)$$

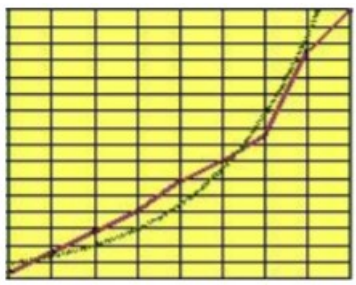
A la raíz cuadrada de la varianza residual la llamaremos ***error típico de estimación***, que es importante en la teoría de la Regresión.

REGRESIÓN NO LINEAL

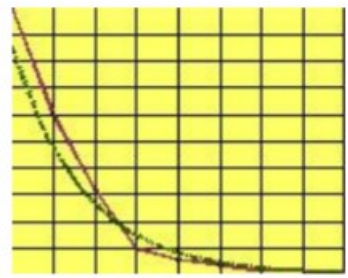
Cuando no está clara o bien fundamentada una tendencia lineal, tendremos que buscar otras formas de gráficas que se ajusten mejor a la distribución bidimensional que estemos estudiando. Por su facilidad de manejo y cálculo, se suelen estudiar las siguientes:

Función exponencial: Se usa para crecimientos y decrecimientos en los que la tasa es proporcional al valor actual (de forma aproximada). Cuanto mayor es el valor actual, mayor es el incremento que sufre. Según ese incremento sea positivo o negativo, la gráfica puede presentar una de estas variantes:

Exponencial creciente



Exponencial decreciente



La gráfica de color rojo corresponde a los datos reales y la de color verde al ajuste exponencial

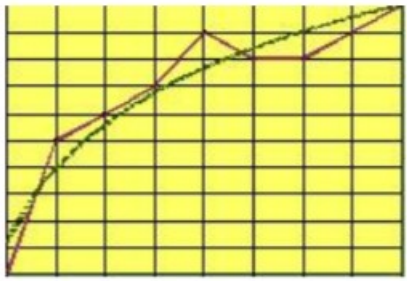
Su expresión es $\mathbf{y = ae^{bx}}$, en la que a y b son dos parámetros a determinar, y e es el número trascendente 2,71828...

Su ajuste se logra transformando la anterior expresión mediante logaritmos neperianos, con lo que queda de la forma Lny = bx + Lna, que al ser lineal, admite los cálculos de regresión explicados hasta ahora. No desarrollaremos esta técnica. Tan sólo hay que recordar que se sustituye Y por su logaritmo.

Función logarítmica: Si se da la proporcionalidad anterior entre el valor actual y la tasa, pero de forma inversa, es decir, que la tasa de variación sea proporcional al valor inverso del actual (1/X), el mejor ajuste es el logarítmico.

Su gráfica suele presentar este aspecto

Se observa un crecimiento de los valores, pero un decrecimiento progresivo de la tasa de variación:



La gráfica de color rojo corresponde a los datos reales y la de color verde al ajuste exponencial

Su expresión es **y = a + b.log(x)**

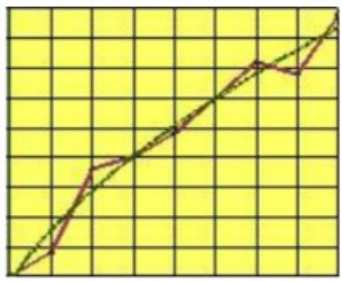
Su ajuste se realiza sustituyendo los valores de x por sus logaritmos.

Función potencial: Es la más potente, pues permite encontrar un exponente fraccionario, lo que abarca las potencias y raíces de todo tipo de exponentes. Su expresión es $\mathbf{y = a.x^b}$

Presenta múltiples formas de gráfica, pues depende del valor de b.

Para ajustar con este procedimiento deberemos tomar logaritmos tanto en la X como en la Y.

La gráfica siguiente corresponde a un ajuste a una raíz cuadrada.



Función polinómica: Suelen ajustarse bien a los datos, pero sus fórmulas pueden complicarse, ya que presentan forma de polinomios, que, a partir del tercer grado son muy complicados. Se usa a menudo el ajuste a un polinomio de segundo grado, o ajuste cuadrático, especialmente en ámbitos científicos.

Para elegir el mejor ajuste a los datos disponemos del coeficiente **R^2 de Determinación**, que es el cuadrado del coeficiente de correlación de Pearson, y nos informa del porcentaje de varianza que está explicado por la función que ajusta los datos. Así, basta elegir la modalidad que tenga el coeficiente mayor, o bien, que su fórmula sea sencilla y el coeficiente apreciable.

El significado más sencillo del valor de **R^2** es el de que representa el cociente entre la varianza explicada por el

modelo y la varianza total (ver atrás para el caso lineal). Así, cuanto mayor sea su valor, más varianza explica el modelo y menor es la varianza de error.

DOS EJEMPLOS DE REGRESIÓN

RELACIONES ALOMÉTRICAS

En Ecología se llama relación alométrica a la existente entre la velocidad de un proceso biológico y cualquier medida (volumen, masa, altura, etc.) de los organismos en los que ocurre, o bien entre dos medidas tomadas en el mismo organismo. Por ejemplo, existe una relación entre la frecuencia cardiaca y la masa total de un animal, o entre áreas de hojas y tamaños de tallos, etc. ¿Cómo encontraríamos la relación alométrica existente entre dos variables si partimos de una tabla bidimensional?

Las relaciones alométricas suelen seguir la fórmula

$$Y = aX^b$$

en la que Y es ua característica de un proceso, X una medida de la que depende Y, a y b constantes, y a esta última b se le suele llamar constante alométrica. Según sea positiva o negativa, mayor o menor que 1, cambiará totalmente el comportamiento creciente o decreciente de Y, así como la relación de su velocidad de cambio respecto a la de X.

La relación alométrica es, pues, de tipo potencial, luego no es correcto aplicar una regresión de tipo lineal. Deberemos acudir en este caso a la hoja de cálculo http://www.hojamat.es/estadistica/tema5/open/tendencias.ods

Supongamos que se han realizado unas medidas en la clase de Biología, en distintos vegetales y se han obtenido estos datos:

X	5	8	10	26	27	31	34	40	67	89
Y	5	7	7	11	13	14	16	15	18	22

Si estas dos medidas siguen el modelo alométrico, deberemos copiar los datos en la hoja tendencias.ods y estudiar su comportamiento como función potencial.

Abre ***tendencias.ods*** y escribe estos datos en su zona de entrada. También puedes seguir el procedimiento usado en otras ocasiones.

Copia esta tabla a la hoja **Borrador** de ***tendencias.ods***. Desde allí usar **Copiar**, pasar a la hoja de **Entrada de Datos** y usar **Pegado Especial**, activando **Transpone**r y eligiendo ***copiar sólo Números***.

Si recorres las distintas tendencias, podrás confirmar que el ajuste potencial es el que presenta un coeficiente R^2 mayor (no tendría que ser necesariamente así)

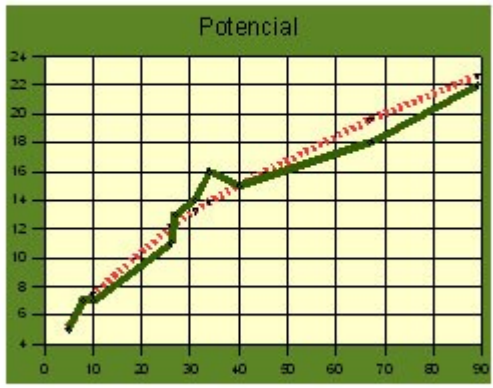
Según la fórmula del ajuste, la constante alométrica tiene un valor de 0,509, coeficiente que equivale aproximadamente a la raíz cuadrada, luego podemos afirmar:

Los valores de Y son proporcionales a las raíces cuadradas de los valores de X

Esto ocurriría, por ejemplo, si X dependiera de una superficie e Y de una longitud.

Potencial

Tipo $y=Ax^B$

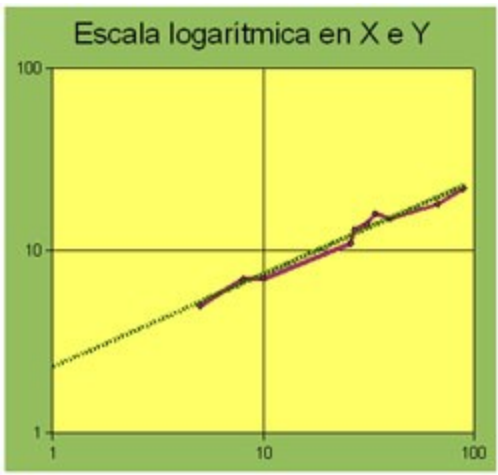


Ajuste: Y´=2,299X^0,509

R^2= 0,9746

Si pasas ahora a la hoja ***Potencial*** de **tendencias.ods**, descubrirás el gráfico en escala logarítmica, mediante la cual los procesos potenciales se representan en línea recta. En este caso el proceso, salvo un pequeño desajuste, sigue esa tendencia. Esta sería una buena motivación para que el alumnado se interesase por los logaritmos, que resultan ser conceptos que tradicionalmente producen más indiferencia.

Si repasas la tabla de esa hoja, en especial la columna DIF, que da los errores cometidos, verás claramente que sólo tres medidas, que producen errores superiores a 1, se apartan de la tendencia general.



RESISTIVIDAD DE UN CONDUCTOR

Para averiguar el coeficiente de incremento de la resistividad de un conductor metálico con la temperatura, se ha sumergido una resistencia en aceite de transformador (aislante) cuya temperatura, medida por un termómetro también sumergido, se puede alterar

a voluntad. Se sabe que la relación entre resistividad y temperatura es prácticamente lineal. Se toman medidas de ambas magnitudes y se desea, a partir de ellas, descubrir el valor de dicho coeficiente.

Imaginemos que los datos obtenidos están recogidos en esta tabla (son datos imaginados, sin base experimental)

Temperatura en ºC	Resistencia en Ω
20	5,27
25	5,22
40	5,53
60	5,65
75	6,3
80	5,86
110	6,6
150	6,53
175	7,21
220	7,74
240	7,87
250	7,66
300	8,23

310	8,65
340	8,96

Para averiguar el coeficiente de incremento de resistencia, bastará usar la fórmula correspondiente

$$\mathbf{R_t = R_0*(1+a\Delta t)}$$

y al quitar paréntesis en ella se nos transformará en

$$\mathbf{R_t = R_0+R_{0*}a\Delta t}$$

Si llamamos X a la temperatura en º C e Y a la resistencia correspondiente, podremos ajustar los datos a una recta de regresión, en la que la pendiente equivaldrá al producto $\mathbf{R_{0*}a}$ y la ordenada en el origen a la medida $\mathbf{R_0}$

Copiamos los datos en el modelo ***regresion.ods*****,** con lo que obtenemos los siguientes valores:

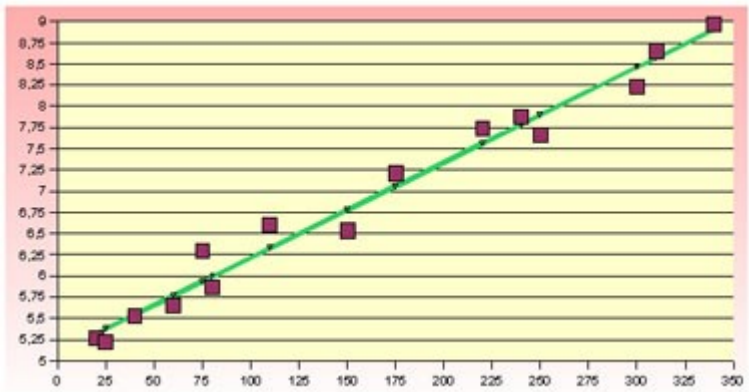
Pendiente = 0,01121

Ordenada en el origen = 5,09578

Coeficiente de correlación = 0,989

La regresión, según el coeficiente de correlación (y su cuadrado el coeficiente $\mathbf{R^2 = 0,977}$) es altamente significativa, por lo que podemos afirmar que las

resistencias aumentan según una relación lineal con las temperaturas. Se ve claramente en el gráfico



Para concretar más, interpretamos los datos:

El valor **5,09578** es una estimación de **la resistencia a cero grados** (no hemos usado incrementos, sino valores totales, luego la escala comienza en cero)

Para hallar el coeficiente deberemos dividir la pendiente entre ese valor 5,09578:

a = 0,00220

Este valor, al ser los datos inventados, no ha de coincidir con el de ningún material conductor conocido.

DISTRIBUCIONES ESTADÍSTICAS TEÓRICAS.

CUESTIÓN-EJEMPLO

Este alumno, ¿está respondiendo al azar?

Un profesor suele plantear a sus alumnos cuestionarios de veinte preguntas, en las que hay que elegir la verdadera entre tres posibilidades. Quiere evaluar con justicia, pero piensa que algunos de sus alumnos y alumnas pueden superar el cuestionario respondiendo al azar. Desearía averiguar, por ejemplo, qué probabilidad se tendría de acertar 10 o más preguntas sin saber nada del tema.

En estos casos, la Estadística puede orientar mediante la comparación de los resultados empíricos con los que se esperarían según unos **Modelos Estadísticos Teóricos**, elaborados mediante técnicas derivadas de la probabilidad. Para conseguir esto, debemos tener muy claras las propiedades del estudio que estamos efectuando, para ver si coinciden con las propias de los modelos teóricos.

En el caso de la cuestión anterior se dan tres condiciones:

- La posible elección de respuestas al azar se repite varias veces, son "***medidas repetidas***".
- En cada intento, si quien responde no sabe nada, tendrá siempre la misma probabilidad de acertar: 1/3, **que no cambiará en todo el experimento.**
- Cada intento es independiente del anterior. **Son fenómenos independientes**.

Resulta que estas tres condiciones caracterizan a un modelo teórico muy popular, que es la Distribución binomial. En este tema nos ocurrirá esto a menudo, que si se cumplen unas condiciones, podrá existir **un modelo** que nos resuelva algunos cálculos.

Distribución binomial

Efectivamente, el ejemplo citado es un caso típico de Distribución Binomial, pues el alumno que no sabe nada siempre tiene la probabilidad p=1/3 de acertar una pregunta por casualidad. Las preguntas son independientes (salvo algunas pautas inconscientes que pueden seguir) y se trata de estudiar los éxitos obtenidos en 20 intentos.

Si repasas la teoría que sigue, reconocerás en este caso la ley Binomial, que será la que apliquemos.

VARIABLE ALEATORIA

El estudio de la probabilidad no entra en los objetivos de estos temas. Por esta razón, de aquí en adelante usaremos la probabilidad como límite de las frecuencias obtenidas en las muestras cuando el número total de datos tiende al infinito*. La Ley débil de los grandes números*, afirma, en efecto, con lenguaje más matemático, que

$$\lim_{n\to\infty} f = p$$

En este sentido usaremos en este curso la probabilidad, aunque no habrá inconveniente en usar hechos derivados de la teoría axiomática correspondiente y quien los conozca podrá seguir este tema con más comodidad.

Llamaremos *Variable aleatoria simple* (discreta) a un conjunto de valores X_1, X_2, X_3, ...,X_n (llamados también *sucesos*) a los que les corresponden unos números (llamados *probabilidades*) , p_1, p_2, p_3, ..., p_n que cumplen:

a) Todas las probabilidades son positivas o nulas.

b) La suma de todas ellas es igual a la unidad

Como consecuencia de las dos propiedades anteriores se deduce que todas las probabilidades están contenidas entre 0 y 1. En lenguaje menos matemático, diremos que estas probabilidades miden las expectativas que podemos tener o las posibilidades que existen de que ocurra un suceso.

A las variables aleatorias también podemos designarlas con el nombre de *Distribuciones teóricas*.

La media en una distribución teórica viene dada por

$$E(X) = \sum X_i \cdot p_i$$

(en la teoría, la palabra media se sustituye por la de ***Esperanza matemática***)

La varianza, a su vez, viene dada por

$$V(X) = \sum {X_i}^2 \cdot p_i - E(X)^2$$

Una distribución de este tipo se representa mediante una tabla en la que estarán contenidos los valores de X y sus probabilidades. Por ejemplo, la distribución de una tirada de dado viene dada por

X	P
1	1/6
2	1/6
3	1/6
4	1/6
5	1/6
6	1/6

Llamaremos ***función de distribución F(x)*** de una variable aleatoria, a la formada por las probabilidades acumuladas, es decir:

F(m) = Prob(x≤m)

(El símbolo ***Prob*** designa a la probabilidad de que sea cierta la comparación del paréntesis)

En una hoja de cálculo es imposible distinguir entre frecuencia y probabilidad, por lo que las usaremos de igual forma.

A continuación repasaremos las distribuciones teóricas más importantes usadas en la Estadística. Existen muchas más, cuya inclusión extendería demasiado este documento.

DISTRIBUCIONES DISCRETAS TEÓRICAS MÁS USADAS

UNIFORME

Una distribución se llama **uniforme** cuando todas las probabilidades son iguales. Como todas suman 1, cada una será igual a 1/n. La distribución del dado incluida en el apartado anterior es un caso típico de esta distribución. Otros ejemplos son el modelo de la tirada de una moneda equilibrada:

X	P
Cara	1/2
Cruz	1/2

Todas las extracciones equilibradas en los juegos de azar son de este tipo.

La media y la varianza de esta distribución se calculan del mismo modo que en una distribución de frecuencias relativas.

En el caso particular de una distribución uniforme discreta en la que X abarca el conjunto de números naturales de 1 a n (como las caras de un dado), la media coincide con **(n+1)/2**, y la varianza con **$(n^2-1)/12$.**

DISTRIBUCIÓN DE BERNOUILLI

Una distribución de Bernouilli se compone de dos sucesos contrarios A y B, a los que se les suele llamar *éxito* y *fracaso*, con probabilidades **p** y **q**

respectivamente. Es evidente que **q=1-p**. Si a **p** la llamamos probabilidad **a favor**, a **q** la designaremos por probabilidad **en contra**. Estas palabras son convencionales, pues si se estudia una epidemia, el *éxito* lo constituiría el ver aparecer un nuevo caso de infección.

Su distribución de probabilidad es:

X	P
Éxito	P
Fracaso	q

Todos los trabajos estadísticos efectuados sobre una *variable dicotómica*, con dos resultados A y B dan lugar a una distribución de Bernouilli: Tener o no un accidente en carretera, ganar o perder en el tenis, contraer o no una enfermedad, etc.

La media de una distribución de este tipo coincide con ***p***:

$$E(X) = p$$

y la varianza con

$$V(X) = pq$$

Un hecho que usaremos más adelante es el de que la máxima varianza se obtiene cuando **p** y **q** son iguales.

DISTRIBUCIÓN BINOMIAL

Esta importante distribución se aplica a pruebas repetidas de la ley de Bernouilli, con las siguientes condiciones:

a) Se realizan experimentos repetidos del tipo Bernouilli, **n** en total.

b) La probabilidad **p** permanece constante en todos ellos

c) Cada experimento es independiente del resultado anterior.

Llamamos a **n** el ***número de intentos***. Estamos interesados en estudiar el número de veces que aparece el suceso A (éxito). A su número de ocurrencias le llamaremos ***número de éxitos***.

Por tanto la ley binomial se aplicará cuando repetimos un experimento cumpliendo las condiciones a), b) y c) establecidas y deseamos estudiar el número de éxitos

que obtendremos. Son de este tipo las tiradas múltiples de monedas, de dados, de ruleta, etc.

La probabilidad de obtener **r** éxitos en **n** intentos se demuestra que equivale a

$$B(r) = \binom{n}{r} p^r q^{n-r}$$

En la que el paréntesis es el número combinatorio **n sobre r**. Del hecho de que esta fórmula sea muy similar a la del Binomio de Newton proviene el nombre de ***binomial***.

La media (esperanza matemática) de esta distribución viene dada por

$$E(r) = np$$

y su varianza por

$$V(r) = npq$$

Consecuencia de esta es una fórmula que nos será muy útil, y es la de su desviación típica, que viene dada por

$$DESV(r) = \sqrt{npq}$$

La distribución binomial de probabilidad **p** y número de intentos **n** se representa generalmente por ***B(n,p)***

DISTRIBUCIÓN DE POISSON

Esta distribución, llamada de *los sucesos raros*, es el caso límite de la binomial, con las siguientes condiciones:

a) El número de intentos **n** debe tender a infinito.

b) La propiedad **p** debe ser muy pequeña (de ahí el nombre de *suceso raro*)

c) El producto de **n.p** ha de ser constante, y al que llamaremos **m**.

Siguen esta distribución el reparto de estrellas en el firmamento, el cómo cayeron sobre Londres los bombardeos en la Segunda Guerra Mundial, las llamadas a urgencias, las averías de las máquinas de una fábrica, etc.

En general la siguen procesos estables, cuyo promedio de ocurrencias por unidad **m** se mantenga constante.

Han de ser procesos aleatorios y las distintas ocurrencias deben ser independientes.

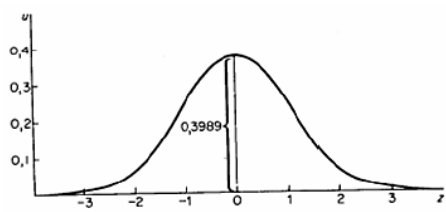
La fórmula de la probabilidad de que aparezcan **x** éxitos viene dada por la fórmula

$$p(x) = \frac{e^{-m}}{x!} m^x$$

La media de esta distribución es **m** y su varianza también vale **m**.

DISTRIBUCIÓN NORMAL

La distribución ***Normal*** o ***ley de Gauss*** es la más usada de las distribuciones teóricas **continuas**. La popularizaron Gauss, en el estudio de los errores de las medidas, y también Laplace, pero ya la había usado Moivre como límite de la binomial.



Por su característica forma, se la conoce también como *campana de Gauss*. Aquí sólo nos interesa su definición y uso dentro de la Estadística.

La expresión de su función de densidad tiene dos versiones:

1) Normal de media μ y desviación típica σ

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{1}{2\sigma^2}(x-\mu)^2)$$

A esta distribución la denominaremos con el símbolo N(μ,σ)

2) Normal tipificada, que se aplica a una variable tipificada z =(x-media)/Desv. típ.

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}z^2)$$

La distribución tipificada se representa por N(0,1)

La distribución normal aparece en muchos fenómenos y estudios. Podemos destacar:

- Magnitudes que dependen de muchas causas independientes, cuyos efectos se suman y cualquiera de ellas aislada tenga efectos despreciables.

- Distribuciones de errores en las medidas

- Medidas de tipo antropológico (estaturas, pesos, inteligencia...) y biológico (glucemia, nivel de colesterol...)

- Límite de otras distribuciones estadísticas cuando **n** aumenta.

Tiene características matemáticas importantes:

Su media, mediana y moda coinciden en el valor cero.

Es simétrica y mesocúrtica.

Posee un valor máximo en la media, y unos puntos de inflexión en $\mu \pm \sigma$

Es asintótica, es decir, que si x tiende a infinito, su densidad de probabilidad tiende a cero.

El uso generalizado de esta distribución proviene de ser el límite de la **binomial** en virtud del Teorema de Moivre:

Si la variable **x** sigue una ley binomial de probabilidad **p**, entonces se cumple:

$$\lim_{n\to\infty} \frac{x - np}{\sqrt{npq}} = z$$

donde **z** sigue la ley normal N(0,1)

Es decir, que si obtenemos una medida tipificada **z** de una distribución binomial con **n** grande, la distribución de **z** se aproximará a la normal. Esta operación se suele efectuar también en procesos no binomiales: Para ajustar sus datos a una distribución normal, se tipifican en primer lugar y después se tratan como valores en la curva normal N(0,1).

Muchos autores han estudiado en qué circunstancias el *ajuste* entre binomial y normal funciona en la práctica. Algunos consejos son:

- Los productos **np** y **nq** deben ser ambos mayores que 3

- Si p<0,1, debe ser np>5

- Si p>=0,1, aunque np<5, el ajuste es aceptable.

OTRAS DISTRIBUCIONES CONTINUAS

En textos universitarios puedes encontrar muchas más distribuciones derivadas de la Normal. Tres de ellas, la ***chi-cuadrado***, ***T de Student*** y ***F de Snedecor***, son muy importantes en la Inferencia Estadística. Están defibidas, además, la *geométrica, la binomial negativa, las distribuciones Alfa, Beta y Gamma*, etc.

RELACIÓN ENTRE FRECUENCIA Y PROBABILIDAD

El problema más importante que hay que considerar cuando se estudian las distribuciones teóricas es la relación que existe entre la probabilidad definida de forma teórica y las frecuencias observadas. Existe un criterio pragmático, y es que si se define una variable aleatoria y se asignan unas probabilidades, las observaciones posteriores de esa variable han de tener un cierto acuerdo con lo definido. Si se asignan los valores de 1/2 a las probabilidades en una tirada de moneda, sospecharemos que es defectuosa si después las frecuencias se alejan del 50%.

Hay dos metodologías para asignar probabilidades:

A) Se estudian muchas muestras aleatorias de una variable, y se asigna la probabilidad como límite de las frecuencias observadas. Podíamos llamarla probabilidad a posteriori, y se basa en la creencia en que las condiciones del experimento no cambian.

B) Se diseña un modelo teórico, basado generalmente en consideraciones de simetría e igualdad de oportunidades, y después se somete ese modelo a pruebas reales para ver si coinciden con lo previsto.

Podemos especificar más esta relación entre frecuencia y probabilidad mediante **los teoremas de los grandes números**, que aquí incluimos en la versión menos rigurosa.

TEOREMA CENTRAL DE LA ESTADÍSTICA

Dada una variable aleatoria **x**, cuya función de distribución es F(x), en la que se han efectuado **n** observaciones, si se designa como FR(x) a las frecuencias acumuladas de dichas observaciones, se

tendrá, para **n** tendiendo a infinito, que será 1 la probabilidad de que la diferencia F(x)-FR(x) sea cero.

De forma más sencilla: *Las frecuencias observadas tienen como límite las probabilidades cuando **n** tiende al infinito*.

Solemos llamar a este hecho la ***Ley de los grandes números***.

Si esta ley falla, es un indicio inequívoco de la probabilidad está mal definida.

TEOREMA CENTRAL DEL LÍMITE

Podemos precisar aún más el carácter de límite de frecuencias que posee la probabilidad:

*Si las variables x_1, x_2, x_3,... x_n, tienen todas la misma distribución, con los mismos valores **m** para la media y **s** para la desviación típica, la variable*

$$\frac{x_{1+}x_{2+}x_{3+} \ldots + x_m - nm}{s\sqrt{n}}$$

sigue asintóticamente la distribución normal N(0,1).

Con la palabra asintótica queremos expresar su coincidencia para n tendiendo a infinito.

Consecuencia importante de esto es:

En toda muestra aleatoria de una población de media μ y desviación típica σ, si llamamos ***m*** *a la media de la muestra, se verifica que la variable*

$$\frac{m-\mu}{\sigma/\sqrt{n}}$$

es asintóticamente normal N(0,1)

Esta convergencia es aceptable a partir de n=30, por lo que este límite se toma para distinguir entre *pequeñas muestras*, en las que la media no se comporta de forma aproximadamente normal, y *grandes muestras*, en las que se sí se puede usar la distribución normal para describir el comportamiento de la media de la muestra.

BONDAD DE AJUSTE

Una profesora de inglés ha ajustado sus datos a una distribución normal, con el siguiente resultado:

Calificación	Frecuencia	Frec. Esperada
0 a 2	6	7,2
2 a 4	24	22,4
4 a 6	34	31,1
6 a 8	12	17,2
8 a 10	6	3,8

Deseamos investigar el grado de confianza que nos proporciona este ajuste de frecuencias.

Existe una distribución, llamada **chi-cuadrado** χ^2, que nos ayuda a medir la aproximación. Su fórmula es la siguiente:

$$\chi^2 = \sum \frac{(O-T)^2}{T^2}$$

en la que O representa a las frecuencias observadas y T a las teóricas. El resultado es un número positivo, la chi-cuadrado, que en sí mismo, apenas nos informa: si es grande, la discrepancia entre ambos conjuntos también lo será, y si es muy pequeño, el ajuste será bueno.

Para medir mejor el ajuste disponemos de las técnicas de estimación que estudiaremos en las últimas sesiones. Adelantando un poco, veremos que se puede medir la probabilidad de la discrepancia que observamos. De esta forma, si nos da una probabilidad muy pequeña, es poco probable que nuestra distribución se ajuste a la teoría.

Se suele marcar como límite el 5%: Si la probabilidad de encontrar una distribución como la nuestra es menor que el 5%, debemos pensar que no existe un buen ajuste, y admitimos que existe en caso contrario, si la probabilidad es mayor que el 5%.

Para entender mejor esto, abre el modelo ***chicuad.ods*** (http://www.hojamat.es/estadistica/tema6/open/chicuad.ods). Copia en él, en la zona que se te indica, las frecuencias reales y teóricas que obtuvo la profesora. Quizás te convenga Pegado Especial como HTM. Es probable que el formato no se conserve.

	Frecuencias observadas	Frecuencias teóricas
1	6	7,2
2	24	22,4
3	34	31,1
4	12	17,2
5	6	3,8
6		
7		

Consulta los resultados en la parte inferior de la hoja y comprobarás que el valor de la chi-cuadrado es de 3,43, que no nos dice nada. Sigue leyendo: como probabilidad de que los resultandos se aparten en este grado de la normal figura el valor 0,3299 (a esta probabilidad la llamaremos ***p-valor***), un 33%, que al ser tan alta, nos permite aceptar que las calificaciones se pueden considerar normales, y las discrepancias fruto del azar.

	Número elementos	**5**	
	Grados de libertad	**3**	
	Valor de chi-cuadrado	**3,43**	
	P-valor	**0,3299**	
	A nivel del 5%	**Se ajusta bien a la teoría**	
Valor crítico al 5%	7,81	Ídem al 1%	11,34

Más abajo figuran los valores críticos: 7,84 si trabajamos al 5% y 11,34 al 1%, claramente superiores al obtenido de 3,43, que **entra dentro de lo esperado** y nos confirma la idea del buen ajuste existente entre los datos empíricos y los teóricos.

Resuelve tú esta otra cuestión:

¿Se puede considerar bien construido un dado que presenta estas frecuencias en 300 tiradas?

Cara del dado	1	2	3	4	5	6
Frecuencia	55	45	50	40	60	50

Las frecuencias teóricas de un dado te las da el sentido común.

Solución: Su p-valor es 0,28, superior al 5%, luego se ajusta a la teoría. No hay sospecha de que esté mal construido, a pesar de las diferencias que se observan.

Cambia las dos primeras frecuencias por 70 y 30 y verás la diferencia.

CASO PRÁCTICO

Una empresa suministradora de prendas profesionales tiene el dato, procedente de pedidos anteriores, de que cierta clase de pantalones presenta una media de talla

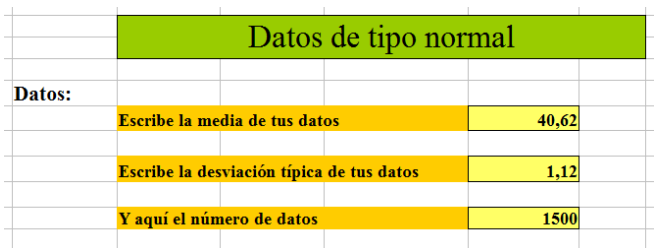
de 40,62 y una desviación típica de 1,12. Desea confeccionar estas prendas para atender los pedidos del próximo año, que calcula serán de unos 1.500 pantalones aproximadamente. ¿Qué número de prendas por talla debe tener preparados?

Los datos antropométricos suelen seguir la distribución normal con bastante exactitud. En este caso podemos suponer que sí se trata de datos de tipo normal, de media 40,62 y desviación típica 1,12.

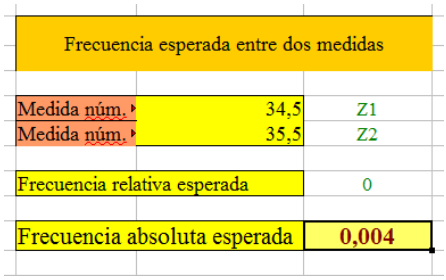
Imaginemos que las tallas que suele ofrecer la empresa van desde la 35 a la 45. Si después se necesitaran más o menos se procedería a cambiar la lista.

La hoja de cálculo ***tablanorm.ods*** (http://www.hojamat.es/estadistica/tema6/open/tablanorm.ods) nos permite traducir los datos a frecuencias.

Para ello debemos rellenar los datos de media, desviación típica y número total



Una vez rellenos los datos fundamentales, podemos escribir cada talla, creando un intervalo entre media unidad antes y media después. Por ejemplo, la talla 35 la representaríamos por el intervalo (34,5 , 35,5). En la siguiente imagen podemos observar que para la talla 35 espera frecuencia 0, por lo que la empresa no tendría, en principio, que ofrecer talla 35.



Frecuencia esperada entre dos medidas		
Medida núm.	34,5	Z1
Medida núm.	35,5	Z2
Frecuencia relativa esperada		0
Frecuencia absoluta esperada		0,004

Si se procede de esta forma se podrá construir una tabla con los datos de las tallas consideradas:

Talla	Número de prendas
35	0
36	1
37	14
38	97

39	324
40	514
41	387
42	138
43	23
44	2
45	0

Suma 1500

Resulta una suma de 1500, pero, a causa de los redondeos, no siempre ha de coincidir el total de previsiones con el número decidido en principio.

Como era de esperar, la gran mayoría de prendas estarían entre las tallas 38 y 42. Sería una decisión empresarial qué tallas se ofrecerán con un carácter general y cuáles pasarían a tallas especiales o fabricadas sólo bajo pedido.

También se observa en la tabla que, en principio, se puede prescindir de la talla 35, e incluso la 36. Lo

mismo ocurriría con las tallas 43 a 45. También constituiría una decisión posterior.

Estas decisiones las han tomado siempre los pequeños comercios e industrias según su experiencia, pero la Estadística ayuda a afinar las previsiones.

CUESTIÓN – EJEMPLO

¿Cómo votarán estos jóvenes en el referendum?

Una profesora de Historia, ante la proximidad de un Referendum en la Unión Europea, decide realizar una encuesta entre 200 de sus alumnos. Obtiene los siguientes resultados:

	Sentido de la votación
SI	95
NO	70
No sabe/No contesta	35

Según estos resultados, ¿se puede inferir que en el Referendum el SI obtendrá el 50% de los votos, o más?

Esta situación es un ejemplo claro de la necesidad de efectuar una operación estadística llamada *Estimación*.

La encuesta que realiza la profesora abarca una muestra de alumnos y lo que le interesa a ella es qué puede ocurrir en la población. Cada vez que tenemos que comparar un colectivo (llamado *población*) con una

de sus partes (llamada *muestra*), debemos realizar una operación llamada *estimación*.

Concretamos

Población

Es el conjunto de referencia que pretendemos estudiar, formado por elementos que comparten una misma propiedad: Españoles adultos, alumnos de la Enseñanza Privada de Méjico, fresnos existentes en la Sierra de Guadarrama.

Censo

Si es posible estudiar toda la población, por ejemplo, los alumnos de un colegio, a este estudio le llamaremos ***censo.*** Un censo no siempre es posible, especialmente por motivos económicos.

Muestra

Una ***muestra*** es un subconjunto de la población, y es el que verdaderamente se estudia en la inmensa mayoría de los experimentos y estudios. Se debe acudir a muestras cuando la población es demasiado numerosa (*población infinita*), o bien resulta muy caro un estudio

exhaustivo. Otro motivo suele ser que el experimento requiera pruebas destructivas, y no es caso de destruir la población.

Una muestra es ***representativa*** cuando tiene una estructura y unos parámetros muy parecidos a la población. Desgraciadamente, esta definición no es útil, pues generalmente no se conoce con seguridad la población, o existe la sospecha de que sus características hayan cambiado.

Llamaremos ***muestreo*** al conjunto de técnicas que nos ayudan a elegir una muestra representativa.

Muestreo

La operación de elegir una muestra puede ser tan compleja que llena libros enteros. Aquí sólo repasaremos las cuestiones de muestreo más frecuentes.

La parte de la Estadística que estudia las estimaciones (relaciones entre poblaciones y muestras) se llama Estadística Inferencial, y es la que comenzamos a estudiar en esta sesión.

Estimación de parámetros

Como habrás visto en la Teoría, se pueden estimar unas características numéricas de la población, llamadas ***parámetros***, mediante unas medidas efectuadas en la muestra, a las que llamaremos ***estadísticos***. Los más populares son:

La media: Mediante el promedio de los datos de una muestra se intenta inferir qué media tendrá la población. Por ejemplo, se mide la resistencia de unos tornillos y se desea con ellos estimar qué resistencia ofrecerán los tornillos fabricados en un largo periodo de tiempo.

La proporción: Es la estimación propia de las encuestas, y por tanto de la de nuestro ejemplo. Se calculan porcentajes en la muestra y con ellos se estiman las proporciones en la población.

La varianza: Se mide la variabilidad de la muestra y con ella se estima la de la población. En este caso no se usa la desviación típica, sino un estadístico muy parecido, la cuasidesviación típica o desviación estándar. Por ejemplo, midiendo las varianzas de varios exámenes de una asignatura en varios cursos se puede inferir la que esperaremos en el próximo curso.

Iremos viendo ejemplos de cada caso. No es necesario que memorices o estudies a fondo la teoría, sino más bien observa cómo trabajan los modelos de esta sesión.

Hay dos clases de estimación:

Puntual: Consiste en asignar al parámetro de la población el mismo valor que su correspondiente estadístico en la muestra. Es una operación muy arriesgada, porque normalmente no coinciden los dos valores. Si así fuera, acertarían todos los sondeos previos a las elecciones.

Por intervalos: En esta modalidad se rodea el valor de la estimación de todo un intervalo de tolerancia, llamado ***intervalo de confianza*** (*horquilla*), en el que se puede evaluar la probabilidad de que figure el verdadero valor del parámetro. Así, si afirmamos que (8,22 , 9,40) es un intervalo de confianza al 96% para la media de una población, queremos indicar que en un 96% de las estimaciones similares que se realizaran, en un 96% de los casos la media pertenecería a ese intervalo, y sólo en un 4% caería fuera.

DEFINICIONES

Cuando el colectivo que se pretende estudiar es muy extenso o inaccesible, se recurre a un subconjunto del mismo llamado ***muestra***, y al conjunto de técnicas usadas se le denomina ***muestreo***.

Población

Es el conjunto de referencia que pretendemos estudiar, formado por elementos que comparten una misma propiedad: Españoles adultos, alumnos de la Enseñanza Privada de Madrid, fresnos existentes en la Sierra de Guadarrama.

Censo

Si es posible estudiar toda la población, por ejemplo, los alumnos de un colegio, a este estudio le llamaremos ***censo***. Un censo no siempre es posible, especialmente por motivos económicos.

Muestra

Una ***muestra*** es un subconjunto de la población, y es el que verdaderamente se estudia en la inmensa mayoría de los experimentos y estudios. Se debe acudir a muestras cuando la población es demasiado numerosa (*población infinita*), o bien resulta muy caro un estudio exhaustivo. Otro motivo suele ser que el experimento requiera pruebas destructivas, y no es caso de destruir la población.

Una muestra es ***representativa*** cuando tiene una estructura y unos parámetros muy parecidos a la población. Desgraciadamente, esta definición no es útil, pues generalmente no se conoce con seguridad la población, o existe la sospecha de que sus características hayan cambiado. Llamaremos ***muestreo*** al conjunto de técnicas que nos ayudan a elegir una muestra representativa.

Muestreo

La operación de elegir una muestra puede ser tan

compleja que llena libros enteros. Aquí sólo repasaremos las técnicas de muestreo más frecuentes;

Aleatorio: Una muestra es aleatoria cuando su elección se hace depender del azar. En concreto, si todos los elementos de la muestra han tenido las mismas oportunidades de ser elegidos, diremos que constituye una ***muestra aleatoria simple (m.a.s.)***. Esta es la muestra que consideraremos aquí.

Intencional: Se llama así cualquier técnica que dependa de la libre voluntad del experimentador, sin recurso al azar.

Errática: Una muestra errática es la que nos encontramos ya formada, sin intervención nuestra, como puede ser el conjunto de alumnos asignados al principio de curso.

DISTRIBUCIONES EN EL MUESTREO

Es fácil confundir las distintas distribuciones estadísticas que concurren en el muestreo. Fundamentalmente son tres:

Distribución en la población: Es el conjunto de frecuencias y medidas que se dan en la población. Salvo mediante un censo, esta distribución sólo se conoce aproximadamente. Las medidas tomadas en la población se llaman ***parámetros***. Los más importantes son

* la media μ

* la desviación típica σ

* cualquier proporción P

* su tamaño N

Distribución en la muestra

Es el conjunto de características de la **muestra concreta que hemos elegido**. Su parecido a la de la población depende totalmente del azar: podemos elegir una muestra representativa sin saberlo, o elegir una muestra sesgada por pura mala suerte. Sus medidas se llaman ***estadísticos***. Los más importantes son

* la media $\bar{x}$

* la desviación típica S

* cualquier proporción p

* su tamaño n

Distribución muestral

Es la resultante de considerar, de forma teórica, **todas las posibles muestras que se puedan elegir**. Es una distribución teórica, construida sobre variables aleatorias, y sus elementos se obtienen mediante técnicas matemáticas. A la media de cualquier estadístico teórico D la representaremos por m_D y a su desviación típica $\mathbf{s_D}$. También usaremos el lenguaje de las variables aleatorias: E(D) representa la media, VAR(D) a la varianza y DESV(D) a la desviación típica.

PRINCIPALES DISTRIBUCIONES MUESTRALES

La teoría que sigue no contiene justificaciones matemáticas de las propiedades que figuran en ella. Todas se pueden demostrar, algunas con técnicas elementales y otras mediante teoremas del límite. Remitimos a textos especializados en Estadística Inferencial.

DISTRIBUCIÓN MUESTRAL DE LA MEDIA

Media: La media de todas las medias muestrales cocincide con la de la población. Es decir, si elegimos muchas muestras distintas, no todas tendrán la misma media que la población; incluso muchas de ellas la tendrán muy alejada. No obstante, si pudiéramos considerar **todas las muestras**, el promedio de todas las medias coincidiría con la media de la población:

$$E(\overline{X}) = \mu$$

por tener esta propiedad, diremos que la media es un estimador insesgado.

Varianza: La varianza de la media tiene, en principio, una distribución más complicada;

$$VAR(\overline{X}) = \frac{\sigma^2}{n}\sqrt{\frac{N-n}{N-1}}$$

La expresión se simplifica si la población es infinita, pues en ese caso la raíz cuadrada tiende a 1, y nos queda una expresión más simple.

$$VAR(\overline{X}) = \frac{\sigma^2}{n}$$

Este resultado es muy interesante: **Cuanto mayor sea el tamaño de la muestra, más pequeña será la varianza de la media, lo que minimizará los errores.**

Podemos deducir de la fórmula anterior la expresión de la desviación típica del estimador media, y obtendríamos

$$e = \sqrt{\frac{\sum (x - \bar{x})^2}{N \cdot (N - 1)}}$$

también llamado ***error muestral*** o ***error de estimación.***

Distribución muestral: Para saber cómo se distribuye la media deberemos distinguir varios casos:

* Si la distribución de la población es **Normal**, y se conoce la **s** de la población, la de la media muestral también será **normal**.

* Si la muestra es **de tamaño mayor o igual que 30**, y se conoce la **s** de la población, aunque la población no sea normal, la media de la muestra sí se comportará como **normal**. Este hecho fundamental se conoce por el nombre de **Teorema Central del Límite**.

* Si la población es aproximadamente normal, y **no se conoce la s** de la población, en muestras grandes (n>120) puede usarse la distribución normal, de forma aproximada, pero en muestras más pequeñas hay que acudir a la **Distribución T de Student**.

DISTRIBUCIÓN MUESTRAL DE LA PROPORCIÓN

La proporciones **p** en las muestras forman una distribución binomial. Si llamamos **P** a la proporción equivalente en la población, la distribución muestral, para poblaciones infinitas, queda:

E(p) = P

por tener esta propiedad, diremos que la proporción es un estimador insesgado.

Es decir, la media de la proporción de las muestras coincide con la proporción en la población.

VAR(p) = PQ/n , llamando **Q** a **1-P**

Como en la media, el aumento del tamaño disminuye los errores.

Si n<30, la proporción sigue la distribución binomial.

Si n>=30, se puede aproximar a la normal.

Si P no se conoce, en la fórmula de la varianza PQ/n podemos sustituir P y Q por p y q, con un pequeño error. Más aún, en la práctica se puede tomar como p y q el valor 1/2, que se puede demostrar daría el error máximo. Así, la varianza quedaría como **VAR(p) < 1/(4n)**. Esta fórmula es muy útil en la práctica.

DISTRIBUCIÓN MUESTRAL DE LA VARIANZA

La varianza de las muestras sigue un proceso distinto a los de la media y proporción. La causa es que el promedio de todas las varianzas de las muestras no coincide con la varianza de la población σ^2. Se queda un poco por debajo. En concreto, se verifica que

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

Hemos usado el subíndice n para recordar que en la varianza se divide entre n.

Si deseamos que la media de la varianza coincida con la varianza de la población, tenemos que acudir a la

cuasivarianza o varianza insesgada, que es similar a la varianza, pero dividiendo las sumas de cuadrados entre n-1.

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

Su raíz cuadrada es la cuasidesviación típica o desviación estándar.

Si se usa esta varianza, si coinciden su media y la varianza de la población

$$E(S_{n-1}^2) = \sigma^2$$

lo que nos indica que la cuasivarianza es un estimador insesgado, y la varianza lo es sesgado.

Distribución muestral de la varianza

La suma de cuadrados de la varianza, dividida entre la varianza de la población

$$\frac{\sum (x-\bar{x})^2}{\sigma^2} = \frac{nS_n^2}{\sigma^2} = \frac{(n-1)S_{n-1}^2}{\sigma^2}$$

se distribuye según una **chi-cuadrado χ^2 con n-1 grados de libertad**

Estimación

Es la operación mediante la cual identificamos el valor de un **parámetro** de la población con el valor de un **estadístico** de la muestra. Es como un acto de confianza: suponemos que la estructura de la muestra permite que sus medidas sean también las de la población. Puede ser una operación arriesgada.

Estimación puntual

La estimación se llama **puntual** cuando identificamos, sin más, el parámetro con el estadístico. En ese caso añadiremos un acento circunflejo al parámetro para representar que estamos estimando.

Un estimador es **insesgado** cuando su media muestral coincide con el parámetro. Así, son insesgadas (y recomendables) estas estimaciones:

$$\hat{\mu} = \overline{X}$$

El estimador insesgado de la media de la población es la media de la muestra

$$\hat{P} = p$$

El estimador insesgado de la proporción es la proporción de la muestra

$$\hat{\sigma}^2 = S_{n-1}^2$$

El estimador insesgado de la varianza no es la varianza de la población, sino la cuasivarianza.

ESTIMACIÓN POR INTERVALOS

Al ser la estimación una operación arriesgada (¿cuándo aciertan totalmente las encuestas políticas?), en lugar de apostar por una estimación puntual, se rodea esta de un intervalo de seguridad, lo que la prensa llama "la horquilla", que técnicamente es el **Intervalo de confianza**.

Para construir un intervalo de confianza, además de la elección del estimador, debemos fijar el **nivel de confianza**, que para no correr riesgos, se suele tomar como una probabilidad grande: 95%, 96%, 99%…

A este nivel de confianza lo representaremos por **1-a**.

Su significado intuitivo es que si repitiéramos muchas veces un experimento con un nivel de confianza, pongamos el 95%, sólo corremos el riesgo de equivocarnos en la estimación un 5% de las veces, mientras acertaríamos un 95%. Así, el símbolo **a** representa el riesgo de que la estimación sea errónea.

Una vez elegido el nivel, sabiendo las distribuciones muestrales, se puede rodear al estimador de todo un intervalo en el que existe una probabilidad 1 - a de que se encuentre en su interior el parámetro estimado.

Los intervalos más populares son (para muestras con n>=30)

INTERVALO PARA LA MEDIA

$$\left(\bar{x} + z_{\alpha/2} \cdot \sigma/\sqrt{n}, \bar{x} + z_{1-\alpha/2} \cdot \sigma/\sqrt{n} \right)$$

Los valores de **z** son uno negativo y otro positivo, por lo que rodean la media. Corresponden a la distribución normal.

σ es la desviación típica de la población, supuesta conocida y **n** el número de elementos de la muestra.

Si no es conocida, recurriríamos a la **t de Student** o a la normal si la muestra es mayor que 120.

Estos casos los puedes consultar en los manuales.

INTERVALO PARA LA PROPORCIÓN

$$\left(p + z_{\alpha/2} \cdot \sqrt{\frac{pq}{n}}, p + z_{1-\alpha/2} \cdot \sqrt{\frac{pq}{n}} \right)$$

Los significados de **z, p, q** y **n** ya están explicados con anterioridad.

INTERVALO PARA LA VARIANZA

$$\left(\frac{nS_n^2}{\chi^2_{1-\alpha/2}}, \frac{nS_n^2}{\chi^2_{\alpha/2}} \right)$$

donde la **chi-cuadrado** se toma con n-1 grados de libertad

DISTRIBUCIONES EN LA REGRESIÓN Y CORRELACIÓN

En las estimaciones correspondientes a la Regresión lineal se admite como hipótesis el siguiente modelo teórico:

Se supone que en la población se han medido dos variables X e Y, que están relacionadas siguiendo estas hipótesis:

(1) - $\mathbf{Y_i = a + bX_i + e_i}$ **,** donde **a** y **b** son parámetros de la población (ordenada en el origen y pendiente) y $\mathbf{e_i}$ es el error de cada observación respecto al modelo lineal

(2) La media de los errores $\mathbf{e_i}$ es cero. La varianza de los errores $\mathbf{e_i}$ coincide con la de la población.

(3) Los errores de las observaciones son independientes entre sí.

Designaremos por **r** al coeficiente real de correlación entre X e Y que presenta la población estudiada.

Estimadores

Llamaremos A al estimador de **a** , B al de **b** , y R al del coeficiente de correlación **r**

Estimador B de la pendiente b

La fórmula del estimador B de la pendiente presenta es:

$$B = \frac{N\sum X_i Y_i - \sum X_i \sum Y_i}{N\sum X_i^2 - \sum X_i \sum X_i}$$

que en realidad es un desarrollo de la que se estudió en el Tema 5 y equivale al cociente entre la covarianza y la varianza de X

$$B = \frac{S_{xy}}{S_x^2}$$

Estimador de la ordenada en el origen a

La fórmula del estimador A de la ordenada en el origen es, como en el Tema 5:

$$A = \bar{Y} - B\bar{X}$$

Estimador de la varianza

La varianza se estima mediante

$$S^2 = \frac{\sum (Y - Y')^2}{N - 2}$$

N-2 son los grados de libertad y el numerador equivale a la suma de los cuadrados de las diferencias entre los valores de Y y sus pronósticos.

Estimador del coeficiente de correlación r

También nos vale la clásica fórmula de Pearson.

$$r = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

que equivale al cociente de la covarianza entre las dos desviaciones típicas (X e Y).

Estimador B

La varianza del estimador de la pendiente B viene dada por la expresión

$$VAR(B) = \frac{\sigma^2}{\sum (X_i - \overline{X})^2}$$

Si suponemos que la población es normal y su varianza conocida, el estimador B también seguirá una distribución normal. Si la varianza es desconocida, su distribución será la T de Student, y se deberá sustituir la varianza por su estimador S^2.

Estimador A

El estimador A posee una varianza algo más complicada de calcular

$$VAR(A) = \frac{\sigma^2 . \sum X^2}{N \sum (X_i - \overline{X})^2}$$

También A se distribuye normalmente o mediante la T de Student, según sea conocida o no la varianza de la población. En este último caso se deberá sustituir la varianza por su estimador S^2.

Estimador S^2

El cociente

$$\chi^2 = \frac{S^2(N-2)}{\sigma^2}$$

se distribuye según una χ2 con N-2 grados de libertad

Estimador r

El cociente $\chi^2 = \frac{S^2(N-2)}{\sigma^2}$

sigue una T de Student con N-2 grados de libertad. El valor de T puede dar una idea de si **r** es significativamente distinto de cero.

Si se aplica al coeficiente **r** la transformación de Fisher

$$Z = \frac{1}{2} \ln \frac{1+r}{1-r}$$

el estadístico resultante se distribuye de forma aproximadamente normal con una varianza igual a 1/(N-3)

Se puede usar esta transformación para construir un intervalo de confianza para el coeficiente de correlación.

ESTIMADORES EN LA REGRESIÓN Y CORRELACIÓN

Este tema se ha considerado siempre como de nivel superior, pero existen situaciones reales que demandan la estimación de parámetros en el modelo de la regresión. Por esto se incluye en estos temas de carácter práctico. Su justificación teórica es más compleja, pero se puede consular en manuales específicos.

Un ejemplo de la necesidad de estas estimaciones es el hecho de que muchos docentes calculan coeficientes de correlación, pero cuando obtienen valores de tipo

medio, como 0,55, 0,62, -0,4 no saben interpretar si estos valores indican que existe correlación significativa. Usaremos un ejemplo práctico para explicar algunas técnicas:

EJEMPLO

¿Es significativa la correlación que encuentro?

Supongamos que un orientador escolar está tratando el tema de la paz con un grupo de alumnos y alumnas. Les ha pasado una prueba en la que la puntuación depende más de los conocimientos previos, y otra, más inclinada a las actitudes. Llamémoslas A y B. La prueba A puntúa entre 0 y 20 y la B entre 0 y 10. Los resultados, ordenados según las puntuaciones en A han sido los siguientes:

Prueba A	0	2	2	4	5	5	5	7	8	8
Prueba B	2	1	1	6	3	7	2	4	6	6
Prueba A	8	9	9	10	10	11	11	11	11	12
Prueba B	8	5	2	4	9	3	6	3	7	5
Prueba A	12	13	14	14	14	15	15	17	17	18

Prueba B	0	9	7	7	6	8	7	9	6	8

¿Es significativamente distinta de cero la correlación entre los resultados de las dos pruebas?

En caso afirmativo, ¿se pueden ajustar estos datos a un modelo lineal **B = a+bA**?

Procedimiento práctico

En primer lugar, mediante una hoja de cálculo, por ejemplo regresion.ods

(http://www.hojamat.es/estadistica/tema5/open/regresion.ods),

calculamos la estimación de los parámetros a, b y r.

Copia los datos en dos columnas de la zona de cálculos.

Columna de la variable X	Columna para la Y	Estimación lineal
0	2	2,08
2	1	2,72
2	1	2,72
4	6	3,35
5	3	3,67
5	7	3,67
5	2	3,67
7	4	4,31
8	6	4,63
8	6	4,63
8	8	4,63
9	5	4,95

Los resultados son:

Coeficiente de correlación r = 0,575
Pendiente estimada B = 0,319
Ordenada en el origen A = 2,080

Media de X	9,900
Media de Y	5,233
Des.Tip. X	4,629
Des.Tip. Y	2,565
Regresión	
Pendiente	0,319
Ordenada	2,080

Coeficiente de correlación	0,575
Error de estimación	2,173
Pronóstico individual	
Escribe un valor de X	26,000
Su Y estimada será	10,361

¿Son significativos estos valores?

Coeficiente de correlación

El coeficiente obtenido 0,575 nos plantea dos cuestiones. La primera es si es significativamente distinto de cero. Para ello puedes usar la misma herramienta

Si has mantenido escritos los datos, pasa a la hoja ***Estimación***. Junto al valor 0,575 figura el valor de T obtenido de la fórmula contenida en la teoría

$$T = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}}$$

con un resultado de 3,716. Este valor es muy alto, y si lees la probabilidad de su aparición (p-valor=0,001) te darás cuenta de que es muy improbable una correlación así por puro azar, luego podemos aceptar que

El coeficiente de correlación 0,575 es significativamente distinto de cero, luego podemos concluir que verdaderamente existe una correlación entre las dos pruebas por el orientador.

La segunda cuestión es la de estimar ese mismo coeficiente en la población. El intervalo de confianza requiere para su cálculo la transformada de Fisher, explicada en la teoría. Aquí no profundizaremos más.

Coeficientes A y B

En los casos como el anterior, en los que existe correlación significativa, se pueden estimar mediante intervalos los coeficientes a y b mediante sus estimadores A y B, que en el ejemplo anterior presentaban los valores B=0,319 y A=2,080. En la teoría puedes consultar sus distribuciones muestrales.

En la herramienta regresion.ods se utilizan estas distribuciones muestrales para construir intervalos de confianza, resultando:

Para A: (0,16 , 4)

Para B: (0,14 , 0,49)

En estos temas no profundizaremos más en estas técnicas. Tan sólo se desea presentar sus posibilidades.

CUESTIÓN-EJEMPLO

Creo que vamos a peor…

Un director de un colegio tiene una especial preocupación por el alumnado de difícil comportamiento. Ha elaborado un criterio objetivo para calificar a ciertos alumnos o alumnas como conflictivos. Se basa en las faltas de asistencia, retrasos, calificaciones trimestrales partes de disciplina, etc. Lleva años calificando como conflictiva a una parte del alumnado, que supone, por término medio el 12% de la población estudiantil de la que procede su alumnado.

Últimamente está observando un incremento del porcentaje de este tipo de calificaciones. En efecto, en el presente curso, con una matrícula de 1385, el colegio presenta un número medio de 201 calificaciones de conflictividad. ¿Puede seguir manteniendo la hipótesis de que sólo supone un 12% del total?

Esta cuestión es un ejemplo claro de un contraste de hipótesis estadística. El director hace una afirmación o tiene una creencia: el grado de conflictividad es del 12% del alumnado. Los hechos, sin embargo, parecen hacerle sospechar que esto ya no es cierto. En efecto, el grado actual es del 201/1385 = **14,5%**

En estos casos surge siempre una duda: ***La diferencia que observo, ¿es debida al azar o a que en realidad la población estudiantil ha cambiado?***

Teóricamente, es imposible responder con seguridad a esta pregunta por lo que lo haremos en términos de probabilidad: Los 1385 alumnos y alumnas de este año constituyen una muestra de la población total. Si presentan un 14,5% de conflictividad puede ser debido a que en la actual promoción ha llegado al colegio, por puro azar, un alumnado de peor comportamiento que la media. Pero también puede ocurrir que haya cambiado toda la población.

Si calculáramos la probabilidad de que ocurra lo primero (por puro azar) y nos resultara muy pequeña, nos inclinaríamos más bien al caso contrario (que ha cambiado la población). Si la probabilidad fuera

razonable, por prudencia, mantendríamos la hipótesis del 12%.

¿Qué es una probabilidad pequeña o una probabilidad razonable? Según el tipo de trabajo estadístico que se emprenda, se suele tomar como límite 0,1, 0,05 ò 0,01. Si deseamos efectuar un contraste de hipótesis sobre la proporción, según la teoría, si np>5 se puede usar la distribución binomial, que desemboca en normal para muestras grandes.

En este caso np coincide con las 201 calificaciones de conflictividad, luego se cumple con creces. Además, conocemos P=0,12, Q=0,88 y n=1385, luego podemos pasar directamente al contraste.

En todo contraste de hipótesis se aconsejan un mínimo de pasos para concretar bien el problema:

(1) Planteamiento de las hipótesis nula y alternativa

En este caso la hipótesis previa es que el porcentaje era del 12%: **H_0 : P = 0,12**

La preocupación del director se justifica por la sospecha de que la proporción ha aumentado, luego: **H_1 : P> 0,12**

Así que planteamos una hipótesis de tipo unilateral por la derecha.

(2) Supuestos del contraste

Un contraste de proporción con una muestra tan grande se comporta como si la población fuera normal, por lo que podemos suponerla.

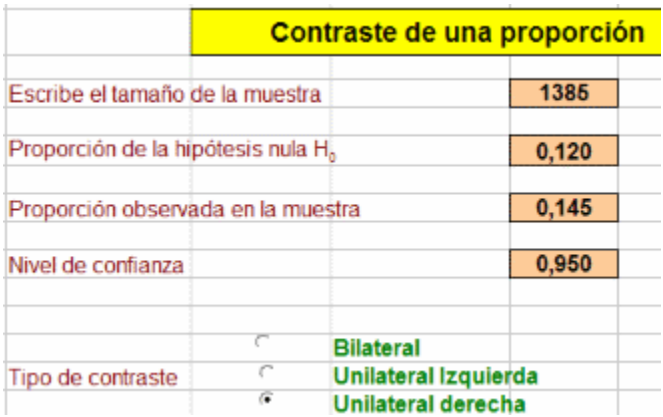
Suponemos muestra aleatoria simple procedente de una población normal.

(3) Estadístico de contraste

Para la elección del contraste debes consultar los manuales de Estadística o el apartado de teoría de este tema. En este caso usaremos

$$Z = \frac{p - P}{\sqrt{PQ/n}} \approx \frac{p - P}{\sqrt{1/(4n)}}$$

Es mucho más cómodo en nuestro caso usar la hoja de cálculo tproporcion.ods (http://www.hojamat.es/estadistica/tema8/open/tproporcion.ods), que contiene este contraste en su primera hoja "Una proporción", y tan sólo necesitamos rellenar los datos:



En la imagen vemos incorporado el dato del tamaño de la muestra, 1385, la proporción de la hipótesis nula, 0,12, y la alternativa de 0,145.

También se ha elegido ya el contraste unilateral por la derecha, porque el objetivo es contrastar si la proporción ha aumentado.

(4) Nivel de significación

Ya se explicó que los niveles más usados son los de 0,1, 0,05 y 0,01. Su complemento a 1 recibe el nombre de Nivel de confianza, que por tanto tendrá usualmente los valores de 0,90, 0,95 y 0,99

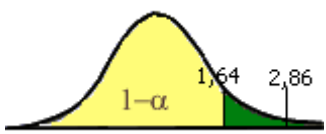
En el caso de Ciencias Humanas se suele elegir el 0,95. Así se ha hecho en nuestro caso.

(5) Toma de decisión

Si observas la parte inferior del esquema de contraste podrás entender cómo se toma la decisión.

Resultados		Valor crítico de Z	
Desviación muestral	0,01	Bilateral	
Estadístico de contraste	2,86	Unilateral izquierda	
P-valor	0,0021	Unilateral derecha	1,64
Decisión			
	Rechazamos la hipótesis		
Intervalo de confianza	(0,128 , 0,162)		
Error muestral	1,71%		

El estadístico de contraste presenta un valor de 2,86, que en la distribución normal está más a la derecha que el valor crítico de 0,95 que ves que es 1,64, luego el estadístico se sitúa en la zona de rechazo.



Otra forma de verlo es con el p-valor, que es la probabilidad, si la hipótesis nula fuera cierta, de que se produzca un resultado del 14,5%. Lee en el esquema su valor, que es de 0,0021, algo muy cercano a cero,

prácticamente imposible. Por tanto, nuestra decisión debe ser:

Se rechaza la hipótesis nula

La población de estudiantes ha cambiado.

(**6) Intervalo de confianza**

A veces, cuando se rechaza un hipótesis, es conveniente proponer una alternativa. Podemos conseguirlo estimando el verdadero valor que tiene la proporción ahora. Esto se consigue construyendo un intervalo de confianza (generalmente bilateral) para el nuevo dato de la población.

En nuestro caso sería el de **(0,128, 0,162)**, es decir, entre un 12,8% y un 16,2%, con un error de estimación del 1,71%. Resulta muy afinado porque la muestra es grande.

RESUMEN TEÓRICO

Tests de hipótesis

Concepto de test de hipótesis

Un test de hipótesis (o contraste) es un proceso, compuesto de varios pasos muy concretos, que nos permite aceptar o rechazar una hipótesis en términos estadísticos. Desarrollamos esto con más extensión:

Proceso de contraste de hipótesis

1) Planteamiento de las hipótesis

Un contraste se comienza con una afirmación. En este curso consistirá en afirmar un valor concreto de un parámetro o de la diferencia o cociente de parámetros. Llamaremos ***Hipótesis nula*** H_0. a la afirmación que hacemos sobre los parámetros de una población y cuya validez deseamos contrastar: La estatura media en Andalucía es de 1,74, este grupo tiene dos puntos más de media que este otro, la varianza de esta población siempre es 54, en todos los experimentos que se efectúen, etc.

Hipótesis nula H_0 es la afirmación cuya validez se desea contrastar.

Frente a esa afirmación podemos oponer otra, a la que llamamos ***hipótesis alternativa H_1***. Suele ser una desigualdad que se opone a la igualdad que afirmamos. Así, si llamamos **m** a la media de una población, podíamos plantear como hipótesis nula

H_0 : m=234

y su hipótesis alternativa podría ser **H_1 : m<>234**, es decir, que la media puede ser mayor o menor que 234. En este caso hablaremos de un *contraste bilateral* o *de dos colas*.

Si tomamos como hipótesis alternativa **H_1 : m>234** o bien **H_1 : m<234** expresaremos que estamos usando un contraste *unilateral* o *de una cola*.

Decidir un contraste u otro depende del contexto. Si experimentamos con un analgésico no nos planteamos que aumente el dolor, ni tampoco que una cantidad de agua disminuya la humedad.

2) Supuestos

Debemos tener en cuenta siempre qué supuestos estamos aceptando sobre la población, si es es simétrica, normal, continua… y sobre la muestra, si es aleatoria simple, es de tamaño mayor que 30…

En este curso, por su especial orientación, no daremos mucha importancia a los supuestos, aunque los nombraremos cuando sea oportuno.

3) Estadístico de contraste

Es la expresión matemática, calculada a partir de la muestra, que nos servirá para tomar la decisión. Debemos conocer la función de distribución que posee. Las tres más usadas son: la Normal, la T de Student y la Chi-cuadrado.

4) Nivel de significación

En todo contraste hay que poner una barrera entre aceptar la hipótesis nula $\mathbf{H_0}$ o rechazarla, y en ese caso aceptar la alternativa $\mathbf{H_1}$. Construiremos dos zonas, la primera de las cuales, ***De rechazo***, compuesta por valores, que por estar alejados de la hipótesis nula, se suponen que no han aparecido al azar, y que por tanto

H_0 es la que ha cambiado. La probabilidad de que unos valores caigan en la región de rechazo, **a pesar de que H_0 sea verdadera**, se conoce con el nombre de ***nivel de significación a***, y se suele fijar con un valor pequeño: 5%, 1%, … Así los experimentos son conservadores, pues asignan muy poca probabilidad a la hipótesis alternativa **H_1**.

La zona complementaria, ***zona de aceptación***, tendrá como probabilidad **1 - a**, generalmente muy grande, del orden del 95%, 96%, o 99%.

5) Toma de decisión

Una vez realizado el experimento en una muestra y calculado el *estadístico de contraste*, según dónde caiga, ***aceptaremos o rechazaremos*** la hipótesis nula. A veces, cuando se rechaza, mediante las técnicas de estimación se propone un valor alternativo.

También se suele añadir a la decisión, no sólo el nivel de significación a, sino también la probabilidad de que en un experimento realizado en las mismas condiciones y con la hipótesis nula considerada verdadera, produzca valores del estadístico **menores** que el obtenido en nuestro caso. A este valor lo llamaremos ***p-***

valor. Si es mayor que 1-a en contraste unilaterales, o bien mayor que 1-a/2 en bilaterales, rechazaremos la hipótesis nula.

A continuación resumimos en dos tablas los contrastes más importantes existentes sobre una muestra y sobre dos. Para más detalles o ejemplos es preferible acudir a un texto de Estadística Inferencial.

Significado de los símbolos: **m** media de la población , **s** desviación típica de la población, **p** proporción en la muestra, **P** proporción en la población , **S** desviación típica en la muestra, **D** diferencia entre medias, $\mathbf{m_D}$ media de la diferencia en la población.

CONTRASTES SOBRE UNA MUESTRA

MEDIA

Estimador

$\overline{X}$

Supuestos

s^2 conocida

Población normal y/o n>30

Distribución

N(0,1)

Estadístico de contraste

$$Z = \frac{\overline{X} - \mu_0}{\sigma / \sqrt{n}}$$

Estimador

$$\overline{X}$$

Supuestos

s^2 desconocida

Población normal, al menos aproximada

Distribución

T_{n-1}

Estadístico de contraste

$$T = \frac{\overline{X} - \mu_0}{s/\sqrt{n-1}}$$

PROPORCIÓN

Estimador

P

Supuestos

Población binomial

np>5

Estadístico de contraste

$$Z = \frac{p - P}{\sqrt{PQ/n}} \approx \frac{p - P}{\sqrt{1/(4n)}}$$

VARIANZA

Estimador

S^2

Supiuestos

Población normal

Distribución

$$\chi^2_{n-1}$$

Estadístico de contraste

$$T = \frac{nS^2}{\sigma_0^2}$$

CONTRASTES SOBRE DOS MUESTRAS

IGUALDAD DE MEDIAS - s^2 CONOCIDA E IGUAL

Estimador

$\overline{X}_2 - \overline{X}_1$

Supuestos

s^2 conocida e igual en las dos muestras

Independencia

Población normal y/o n>30

Distribución

N(0,1)

Estadístico de contraste

$$\frac{X_2 - X_1 - (\mu_2 - \mu_1)}{\sqrt{\left(\frac{\sigma_2^2}{n_2} + \frac{\sigma_1^2}{n_1}\right)}}$$

IGUALDAD DE MEDIAS - s^2 DESCONOCIDA E IGUAL

Estimador

$\bar{X}_2 - \bar{X}_1$

Supuestos

s^2 desconocida e igual en las dos muestras

Independencia

Población normal

Distribución

$T_{n1+n2-2}$

Estadístico de contraste

$$\frac{X_2 - X_1 - (\mu_2 - \mu_1)}{\sqrt{\left(\frac{1}{n_2} + \frac{1}{n_1}\right)\left(\frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2}\right)}}$$

IGUALDAD DE MEDIAS - s^2 DESCONOCIDAS Y DISTINTAS

Estimador

$\overline{X}_2 - \overline{X}_1$

Supuestos

s^2 desconocidas y distintas en las dos muestras

Independencia

Población normal

Distribución

N(0,1)

Estadístico de contraste

$$\frac{X_2 - X_1 - (\mu_2 - \mu_1)}{\sqrt{\left(\frac{\sigma_2^2}{n_2} + \frac{\sigma_1^2}{n_1}\right)}}$$

Se supone que las dos desviaciones típicas de la población se sustituyen por sus estimadores insesgados, las cuasidesviaciones típicas

IGUALDAD DE MEDIAS – DATOS APAREADOS

Estimador

$\bar{X}_2 - \bar{X}_1$

Supuestos

s^2 desconocida e igual en las dos muestras

Datos apareados

Población normal

Distribución

T_{n-1}

Estadístico de contraste

$$\frac{\overline{D} - \mu_d}{\sqrt{\left(\frac{\sigma_d^2}{n}\right)}}$$

IGUALDAD DE PROPORCIONES

Estimador

$p_2 - p_1$

Supuestos

Población binomial

Igualdad de proporciones en muestras grandes e independientes

Se suponen proporciones iguales en la población

Distribución

N(0,1)

Estadístico de contraste

$$Z = \frac{p_1 - p_2}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Siendo

$$p = \frac{n_1 \cdot p_1 + n_2 \cdot p_2}{n_1 + n_2}$$

DIFERENCIA DE PROPORCIONES

Estimador

$p_2 - p_1 = d$

Supuestos

Población binomial

Diferencia de proporciones en muestras grandes e independientes, sin el supuesto de igualda de proporciones en la población

Distribución

N(0,1)

Estadístico de contraste

$$Z = \frac{p_2 - p_1 - (P_2 - P_1)}{\sqrt{P_1 Q_1 / n_1 + P_2 Q_2 / n_2}}$$

IGUALDAD DE VARIANZAS

Estimador

$$S_2^2 / S_1^2$$

Supuestos

Poblaciones normales

Distribución

Fn1-1,n2-1

Estadístico de contraste

$$F = \frac{S_2^2}{S_1^2}$$

AMPLIACIÓN

Contraste para la diferencia de dos proporciones

Unos grandes almacenes han instalado unas cajas de cobro automáticas. Durante los primeros meses, tan sólo las han usado un 8% de la clientela, por lo que deciden iniciar una campaña publicitaria a fin de incrementar ese uso en un 10%, y justificar así su instalación. Durante unos días, en horas elegidas aleatoriamente, han efectuado un recuento y han descubierto que de un conjunto de 2340 clientes, tan sólo han usado las cajas 208. Después de desarrollar la campaña, han repetido el estudio, y esta vez, de 1978 clientes, han pasado por las nuevas cajas 395. ¿Justifican estos resultados, al 95% de nivel de confianza, que se ha logrado el incremento deseado del 10%?

En este caso aplicaremos el estadístico de contraste

$$Z = \frac{p_2 - p_1 - (P_2 - P_1)}{\sqrt{P_1 Q_1 / n_1 + P_2 Q_2 / n_2}}$$

en el que las proporciones en la población son 8% y 18% respectivamente (si admitimos esto como hipótesis nula) y las de la muestra 208/2340=0,0889 y 395/1978=0,1887.

Abre la tercera hoja del libro **tproporcion.ods** (http://www.hojamat.es/estadistica/tema8/open/tproporcion.ods) y escribe en ella los datos. Como el error es pequeño, se toman aquí como parámetros de la población los mismos valores que en la muestra, y sólo hay que rellenar la diferencia de proporciones supuesta (aquí el 10%)

Suponemos contraste bilateral y fijamos el 95% de nivel de confianza:

Diferencia de proporciones			
Tamaño muestra 1	2340	Tamaño muestra 2	1978
Proporción 1	0,0889	Proporción 2	0,1997
Diferencia de proporciones supuesta		0,100	
Diferencia de proporciones observada		0,111	
	0,950		

El resultado del contraste será que se rechaza la hipótesis de un incremento del 10%. Si rellenas los datos observarás que ha subido un 11,1% de forma significativa.

Tipo de contraste	◉	Bilateral				
	○	Unilateral Izquierda				
	○	Unilateral derecha				
		Resultados		**Valor crítico de Z**		
	Desviación muestral		0,01	Bilateral	-1,96	1,96
	Estadístico de contraste		10,31	Unilateral Izquierda		
	P-valor		0,0000	Unilateral derecha		
	Decisión	Rechazamos la hipótesis. La diferencia no es la supuesta				
	Intervalo de confianza Para la diferencia		(0,090 , 0,132)			
	Error muestral		2,11%			

CASO PRÁCTICO

En una ONG se organizan encuentros trimestrales con todos los Delegados y Delegadas. Suelen asistir, salvo pequeñas variaciones y ausencias, las mismas personas. En cada encuentro se recoge una valoración posterior y se intentan mejorar los aspectos que se hayan puntuado menos. La Dirección está interesada en saber si las correcciones surten efecto, y por eso desearía averiguar si las medias de las encuestas de cada dos encuentros consecutivos son significativamente distintas entre sí. Los últimos encuentros produjeron estos resultados:

Encuentro	Octubre 07	Enero 08	Abril 08	Julio 08
Media	4,2	3,7	4,1	4,2
Desviación t.	0,8	1,2	1,4	1,2
Asistentes	47	54	45	49

¿Cómo tratar estadísticamente estos datos?

El interés de la Dirección está en comparar cada dos medias consecutivas, luego se está en el caso de Diferencia de Dos Medias Independientes. Como siempre asisten las mismas personas, salvo pequeños cambios, se puede suponer que la varianza de las poblaciones es desconocida, pero siempre la misma. Al ser el número de asistentes superior a 30, se puede suponer la normalidad de la población.

El análisis de esta situación se puede efectuar con la hoja de cálculo **tmedia.ods**

(http://www.hojamat.es/estadistica/tema8/open/tmedia.ods).

Ábrela y elige la hoja *Dos medias (independientes).* Fija antes de nada que el contraste sea Bilateral, porque no

tenemos motivos para inclinarnos por un sentido u otro. Activa también el caso Se suponen desconocidas, pero iguales.

Para cada par de encuentros consecutivos rellena los datos de media, desviación típica y tamaño de la muestra. Obtendrás estos resultados al 95% de Nivel de Confianza:

Primera y segunda: Estadístico de contraste 2,4, p-valor 0,0082. Son significativamente distintas.

Segunda y tercera: Estadístico de contraste 1,52, p-valor 0,0649. No hay razón para pensar que las medias son distintas.

Tercera y cuarta: Estadístico de contraste 0,37, p-valor 0,64. No hay razón para pensar que las medias son distintas.

ANÁLISIS DE LA VARIANZA (ANOVA)

CUESTIÓN-EJEMPLO

Se han aplicado cuatro métodos distintos para el aprendizaje del concepto de número primo a cuatro grupos de alumnos y alumnas elegidos aleatoriamente. Posteriormente se les ha pasado la misma prueba para valorar la adquisición del concepto, con los siguientes resultados:

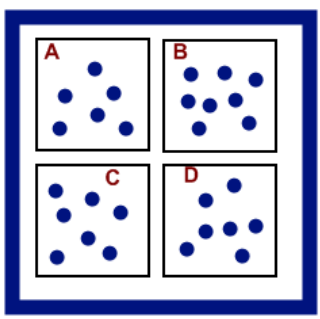
Método A	Método B	Método C	Método D
8	16	16	11
12	12	15	9
11	13	13	8
15	15	17	8
7	19	13	9
9	16	9	12
10	13	19	10
11	10	16	9
17	6	14	5
12	11	13	10

Se supone población normal y que las muestras son independientes entre sí. ¿Hay alguna evidencia, al 95% de Nivel de Confianza, de que exista un efecto en la aplicación de los distintos métodos?

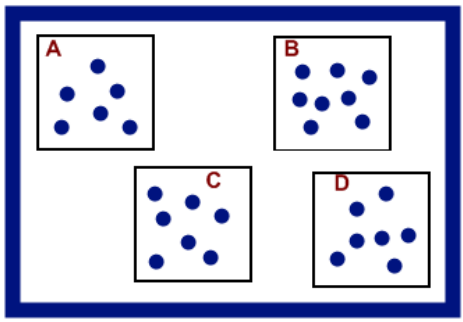
Aunque se haya expresado con otras palabras, lo que interesa en esta situación es averiguar si las medias de las cuatro poblaciones representadas por la aplicación de los métodos se pueden considerar iguales o no, es decir:

H_0: $m_1 = m_2 = m_3 = m_4$

Lo sorprendente de la técnica que vas a aprender es que para averiguar esto se acude a analizar la varianza. La razón es que si las medias son iguales, la varianza total disminuye, pero si son muy diferentes, aumenta. Es una idea intuitiva que podemos expresar con estas imágenes:



En esta situación, los cuatro grupos están muy cercanos. Su varianza total no será grande. Cada grupo tiene su propia varianza interna.



En esta otra, al separarse los grupos, **la varianza total aumentará**, porque hay más dispersión, pero la varianza interna de cada grupo es la misma. Lo que ha aumentado es la variabilidad **Intergrupos**

Observando las imágenes puedes entender que si la varianza total aumenta, esto puede deberse a dos causas, o a que haya aumentado la varianza interna de cada grupo, o, lo que es más probable, que se hayan separado las medias y eso ha aumentado la varianza total.

Cuando las medias de varios grupos relacionados se separan entre sí, aumenta la varianza total

El Análisis de la varianza (ANOVA) nos permite aceptar o rechazar la hipótesis nula $\mathbf{H_0: m_1 = m_2 = m_3 = m_4}$ descomponiendo la varianza total en dos sumandos: **Intragrupos e Intergrupos**. Según sean estas cantidades se tomará una decisión u otra.

En la práctica se forman tres sumas de cuadrados distintas y después se restan adecuadamente. Para entenderlo mejor, abre la hoja de cálculo

anova.ods (http://www.hojamat.es/estadistica/tema9/open/anova.ods)

y vuelca en ella los datos de la cuestión que estamos estudiando. Lo puedes conseguir con Copiar y Pegar.

S1: Consiste en sumar todos los cuadrados de los datos. En la hoja **anova.ods** figura a la derecha, y su valor es en este ejemplo **6207.**

S2: Se suman los cuadrados de las sumas de los distintos niveles dividido cada uno entre el número de datos. En el ejemplo su valor es de **5901,1**

S3: Se obtiene dividiendo el cuadrado de la suma total de todos los niveles dividido entre el número total de datos. En este caso vale **5736,03**

Una vez obtenidas estas sumas, se van restando y resultarán las sumas de cuadrados Intergrupos, Intragrupos y Total:

Suma de cuadrados INTRA: S1-S2 = 6207 - 5901,1 = 305,9

Es la suma de cuadrados que corresponde al interior de los niveles, sin tener en cuenta sus diferencias de medias. Sus grados de libertad se obtienen restando el número total (40) menos el número de niveles (4), es decir, 36. Su cociente es el mejor estimador de la varianza de la población, en este caso **8,5**

Suma de cuadrados TOTAL: S1-S3 = 6207 - 5736,03 = 470,98

Es la suma total de cuadrados. Sus grados de libertad son N-1, que en este caso son 39, con lo que la varianza total será 470,98/39 = **12,08**

Suma de cuadrados INTER: S2-S3 = 5901,1 - 5736,03 = 165,08

Esta suma refleja los desniveles en las medias. Si es alta, puede indicar que las diferencias entre medias son significativas. Sus grados de libertad equivalen al número de niveles menos 1, en el ejemplo 3. La varianza INTER será entonces igual a 55,03

Contraste

El punto importante del ANOVA es el contraste entre unas varianzas y otras, que se realiza, como vimos en el tema anterior, mediante la prueba F.

Observa en el archivo anova.ods cómo se contrasta la igualdad entre las varianzas INTER e INTRA mediante la prueba F. Al dividir nos resulta un valor de F=6,48, muy grande, con un p-valor de 0,001 que la convierte en significativa, luego las medias de los distintos niveles **no se pueden considerar iguales.**

Como resultado del ANOVA podremos afirmar que en nuestro ejemplo el método de enseñanza ha influido en los resultados.

CONCEPTOS GENERALES

La técnica del **Análisis de la Varianza** consiste en descomponer la variabilidad de una población (representada por su varianza) en diversos sumandos según los factores que intervengan en la creación de esa variabilidad. Por ejemplo, si estudiamos la varianza que presenta una colección de calificaciones que provienen de tres asignaturas en cuatro cursos distintos a lo largo de los últimos años, la varianza total se puede descomponer en cuatro sumandos:

- Parte proveniente del factor asignatura
- Componente aportado por los distintos cursos
- Influencia de la evolución temporal en los últimos años
- Varianza propia (interna) de la población.

Este ejemplo sería bastante complejo, porque depende de **tres factores**: asignatura, curso y año. Son mucho más frecuentes los ejemplos de **un solo factor** (por ejemplo, tres métodos distintos aplicados simultáneamente a alumnado del mismo nivel) o de **dos**

factores (lo sería la influencia del sexo y la edad en un rendimiento)

Lo original del Análisis de la Varianza es que su verdadero objetivo no es la variabilidad, sino otros contrastes, como la igualdad de medias o el ajuste en un problema de Regresión. Lo veremos en los casos que vamos a estudiar.

ANÁLISIS DE VARIANZA DE UN FACTOR

MODELO Y SUPUESTOS

Supongamos la existencia de varias muestras distintas que corresponden a los resultados obtenidos en una población bajo la influencia de distintos niveles de un factor. La palabra niveles no se debe interpretar en sentido ordinal. Pueden ser niveles distintos métodos de enseñanza, lugares de nacimiento o sexo. Se consideran igualmente válidos niveles cualitativos o cuantitativos (fijos).

Para fijar ideas, supongamos un experimento consistente en medir los minutos transcurridos en la

desaparición de un dolor después de la administración de tres tipos de analgésicos a una muestra de pacientes con migraña

Analgésico	A	B	C
	12	11	14
	18	12	18
	14	13	17
	10	12	21
	21	8	17
	15	15	16
		18	19
		11	21

En este caso el factor es el tipo de analgésico, que actúa a través de tres niveles distintos A. B y C.

En la tabla se observa que dentro de cada nivel existe bastante variabilidad (cada paciente tendrá su forma de reaccionar), y que parece que también existen diferencias entre unos niveles y otros. Si calculáramos las medias nos resultaría

m_1=15; m_2=12,5; m_3=17.875

Si las medias fueran iguales, negaríamos que existan diferencias en el efecto de los distintos analgésicos, pero como no lo son, deberemos plantearnos un contraste de hipótesis para la igualdad de medias.

De hecho, el verdadero contraste que se propone el Análisis de la Varianza es el de **igualdad de medias**. Plantearemos la hipótesis nula:

H_0: $m_1=m_2=m_3$

Pero en realidad la contrastaremos descomponiendo la varianza. Para ello supondremos que cada medida de minutos se puede descomponer en tres sumandos:

$a_{ij} = m + a_i + e_{ij}$

a_{ij}: Es la medida real que se observa en los sujetos (12, 18, 13, 10...) y se considera descompuesta en tres factores aditivos

m: Es la media general de todo el experimento. En el ejemplo equivaldría a 15,14.

a_i: Mide la influencia del factor, mediante la diferencia entre la media de cada columna y la media general. En

el ejemplo se darían estas diferencias: a_1=15-15,14=-0,14 a_2=12,5-15,14=-2,64 a_3=17,875.15,14=2,735

e_{ij}: Mide la variación propia de cada individuo.

Para entenderlo mejor descompondremos dos datos:

La medida 8 de la segunda columna equivale a 8=15,14-2,64-4,5. En esta suma 15,14 es la media general del experimento, -2,64 la influencia del medicamento B y -4,5 la diferencia aportada por el individuo, que ha reaccionado muy rápido.

La medida 19 de la tercera columna equivale a 19=15,14+2,735+1,125. El factor medicamento aporta 2,735, porque es el más lento en actuar, y el individuo 1,125, que no es tan rápido como el anterior.

MODELO

El conjunto de supuestos más aceptado en este caso, porque permite inferencias muy simples, es el siguiente:

Se trabaja sobre una variable aleatoria Y_{ij}, a la que se le supone descompuesta de la siguiente forma:

$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$ donde μ es la media de la población, α_i la influencia del factor, que equivale a la diferencia entre la media general y la del grupo. Finalmente, ε_{ij} se corresponde con la diferencia propia de cada individuo.

Se supone que todas las Y son **normales e independientes**.

Llamamos n_i al número de sujetos por grupo, y N al número total, con lo que $n_1+n_2+n_3\ldots=N$

ESTIMADORES

La media general μ se estima mediante

$$m = \frac{\sum_i \sum_j Y_{ij}}{N}$$

Las medias de cada grupo o nivel de forma similar:

$$m_i = \frac{\sum_j Y_{ij}}{n_i}$$

La influencia del factor (α_i) se estima mediante la diferencia $\alpha_i = m_i - m$

Para la estimación de la varianza deberemos antes abordar la operación fundamental del Análisis de la Varianza, que consiste en descomponer en sumandos la suma de cuadrados de los datos corregida con la media. Se distinguen tres sumas distintas:

Suma de cuadrados total (SCT)

Viene dada por la fórmula

$$SCT = \sum_i \sum_j (Y_{ij} - \bar{Y})^2$$

Coincide con el numerador de la varianza total. Esta fórmula se puede simplificar mediante esta otra:

$$SCT = \sum_I \sum_J Y_{IJ}^2 - N\bar{Y}^2$$

En el ejemplo de arriba el valor sería

SCT = 5339 – 229,109504*22 = 298,59

Si esta suma la dividimos entre los grados de libertad, que son N-1, nos resultará la Media cuadrática Total. En este caso: MCT = 298,59/(22-1)= 14,22

Suma de cuadrados Intra o de error (SCE)

Representa la suma de cuadrados corregidos que se da dentro de los grupos, es decir, las diferencias de los datos entre la media de cada grupo.

$$SCE = \sum_i \sum_j (Y_{ij} - \overline{Y}_i)^2$$

Y su expresión reducida

$$SCE = \sum_i \left(\sum_j Y_{ij}^2 - n_i \overline{Y}_i^2 \right)$$

En el ejemplo su valor sería

SCE=(1430-15^2*6)+(1312-12,5^2*8)+(2597-17,88^2*8) = 182,88

Si lo dividimos esta suma entre los grados de libertad N-n nos resultará la media cuadrática de error, que es el mejor estimador de la varianza de la población.

MCE = 182,88/(22-3)=9,63

Suma de cuadrados Inter (entre grupos)

Se define mediante la fórmula

$$SCI = \sum_i n_i(\bar{Y}_i - \bar{Y})^2$$

Aunque también se puede calcular restando, ya que se demuestra que

SCT = SCI + SCE

En el ejemplo valdría

SCI = 298,59 – 182,88 = 115,72

También se puede hallar la media cuadrática Inter dividiendo entre los grados de libertad i-1

MCI=115,72/2 =57,86

En la práctica se forman tres sumas de cuadrados:

$$S1 = \sum_i \sum_j Y_{ij}^2$$

que consiste en sumar todos los datos por separado elevados al cuadrado. En el ejemplo tendría un valor de 5339.

$$S2 = \sum_i \frac{\left(\sum_j Y_{ij}\right)^2}{n_i}$$

que equivale a sumar los datos de cada nivel, elevar al cuadrado y dividir entre el número de datos. En el ejemplo:

$S2=90^2/6+100^2/8+143^2/8= 5156,13$

Y por último, la suma S3 equivale al cuadrado de la suma total de datos dividida entre el número total de los mismos.

$S3 = 333^2/22 = 5040,41.$

De esta forma, la suma de cuadrados total es la diferencia entre S1 y S3 (se puede demostrar)

SCTotal = S1 – S3 = 5339 – 5040,41 = **298,59**

De igual forma, la suma de cuadrados Intra es la diferencia entre S1 y S2

SCIntra = S1 – S2 = 5339 – 5156,13 = **182,88**

Y la otra diferencia será la suma Inter:

SCInter= S2-S3 = 5156,13 - 5040,41 = **115,72**

ANÁLISIS DE VARIANZA DE DOS FACTORES

MODELO Y SUPUESTOS

Supongamos la existencia de varias muestras distintas que corresponden a los resultados obtenidos en una población bajo la influencia de distintos niveles de dos factores.

Por ejemplo, imaginemos que las medidas de la tabla siguiente se han obtenido en tres barrios distintos A,B y C y en tres niveles de edad: 10-30, 31-50, 51-70. Podemos imaginar las medidas como una valoración que se ha recogido en una encuesta:

	Barrio A	Barrio B	Barrio C
10-30	3,4,4,5,4 2,4,5,3,1	6,2,3,4,5,4 4,5,6,2,7	2,4,5,6,6 3,4,3
31-50	6,8,4,6,9 7,3,4,8,7	8,9,6,7,7 10,6,9,8,7	5,7,5,6,6 3,5,4,6
51-70	4,2,2,4,5 1,3,2,4,5	6,6,4,5,3,8 4,0,1,4	5,6,4,5,3 5,2,1,1,1,0

Al igual que en el caso de un factor, podemos descomponer las medidas en varios sumandos:

$a_{ijk} = m + a_i + b_j + ab_{ij} + e_{ijk}$

a_{ijk} es una medida cualquiera, individual, que la consideramos descompuesta en cuatro sumandos:
m: Es la media total de toda la tabla.

a_i : Mide el efecto del factor A. En el ejemplo podría ser el barrio, que influyera en la valoración efectuada por los sujetos.

b_j : Mide el efecto del otro factor B, en nuestro caso el nivel de edad.

ab_{ij} : Puede que los efectos de A y B no sean aditivos sin más, sino que exista interacción entre ellos. Este sumando mide dicha influencia mutua. Si se supone que A y B son independientes, valdrá 0, y consideraremos un modelo **sin interacción.**

e_{ijk} : Contiene las diferencias individuales. Se supone que su distribución es Normal de media 0.

La hipótesis nula en este caso es la de que todas las medias de los subgrupos son iguales.

Como en el caso anterior, el análisis se basa en sumas de cuadrados y en grados de libertad, para después dividirlos, obtener estimadores de la varianza y compararlos mediante un contraste F.

Si se ha entendido el modelo de un factor, para abordar éste hay que considerar que existen cuatro fuentes de variación en este problema.

Explicaremos cada fuente mediante la resolución que del ejemplo propuesto nos brinda la hoja de cálculo. En unos temas prácticos como estos, no llenaremos la teoría de sumatorios, remitiendo a manuales específicos el estudio detallado de los mismos.

Fuente variación	SC	G.L.	CM	F
Factor A	29,26	2	14,63	5,05 P=0,165
Factor B	149,04	2	74,52	25,73 P=0,005
Interacción AB	11,65	4	2,91	1,01 P=0,410

Error	231,68	80	2,9	
TOTAL	421,62	88		

Fuente de variación Barrio: SCA=29,26. Esta suma representa la variabilidad de los tres grupos formados por los barrios. Se consigue de forma similar a la de un factor. Sus grados de libertad son 2, equivalentes al número de barrios menos 1.

Fuente de variación Edad: SCB= 149,04. Representa la variabilidad entre edades. Como existen tres niveles, sus grados de libertad también son cuatro.

Interacción: SCAB=11,65. En algunos modelos no se considera que haya influencia entre los dos factores. Esta decisión se debe tomar teniendo en cuenta conocimientos anteriores, y no como consecuencia de los datos obtenidos en el ANOVA. En estos temas usaremos siempre modelos con interacción.

Sus G.L. se calculan multiplicando los de los dos factores.

Error: SCE=231,68. Las sumas correspondientes al error y sus grados de libertad se suelen calcular restando los totales de los otros tres. Así se consigue más rapidez. Este sumando representa la variabilidad interna de los datos, independientemente de la influencia de los factores. Es el verdadero estimador de la varianza, y hay quien plante el ANOVA sólo para conseguir este estimador.

En el ejemplo la mejor estimación de la varianza sería 2,9.

Total: SCT=421,62. Es la suma de los factores, la interacción y el error. Su utilidad reside en facilitar los cálculos y comprobar que las sumas cuadran bien.

Todos los cuadrados medios estiman la varianza de la población, aunque el mejor estimador sea 2,9. Si aplicamos el contraste F a la comparación de estimadores, los sesgos significativos que encontremos se deberán a influencias de los factores.

En el ejemplo ha resultado significativo el factor edad, al tener un p-valor inferior al 5%.

ANÁLISIS DE LA REGRESIÓN

MODELO Y SUPUESTOS

Las técnicas de descomposición en sumas de cuadrados propias del ANOVA también se pueden aplicar a la regresión entre dos variables. El modelo teórico es el de suponer que entre dos variables X e Y existe una relación lineal de la forma:

$$\mathbf{Y_{ij} = \alpha + \beta X_i + e_{ij}}$$

En esta fórmula supondremos lo siguiente:

X es cuantitativa y presenta valores fijos, como los niveles en el modelo ANOVA. Estos valores dividen a los de Y en distintos subconjuntos. Se supone que los valores de Y en ellos son independientes entre sí (covarianza cero)

Y presenta valores aleatorios dependientes de X según la relación lineal $\alpha + \beta X$ a cuyo valor se añade un sumando aleatorio $\mathbf{e_{ij}}$. Se supone que $\mathbf{e_{ij}}$ se distribuye normalmente y que las medias de Y en los distintos

conjuntos dependen de las medias de X según la misma relación lineal.

Los valores de la varianza en los distintos subconjuntos son iguales (homocedasticidad)

Lo anterior es un breve resumen de los supuestos. En manuales de Inferencia Estadística puedes estudiarlos con más amplitud.

En los temas 5 y 7 estudiamos los estimadores de α y β y los valores pronosticados $Y' = \alpha + \beta X$. Aquí nos interesarán más bien las descomposiciones en sumas de cuadrados y las técnicas de ANOVA.

Hipótesis nula: $\beta=0$

Hipótesis alternativa: $\beta\neq0$ (o $\beta<0$ o $\beta>0$)

La anulación de β equivale a que todas las medias de subgrupos sean iguales, porque la recta de regresión sería horizontal, luego esta hipótesis nula coincide con la del ANOVA de igualdad de medias. Por eso nos vale esta técnica también para la regresión. Explicaremos cómo:

Suma de cuadrados total:

$$SCT = \sum_i \sum_j (Y_{ij} - \overline{Y})^2$$

Tiene la misma expresión que en el ANOVA, y sus grados de libertad serán N-1, porque se ha estimado un valor, que es la media de Y.

Suma de cuadrados explicada:

$$SCT = \sum_i \sum_j (Y'_{ij} - \overline{Y})^2$$

Si representamos los pronósticos del modelo de regresión como Y', se dará entonces la identidad **Y'=**α **+** β**X**. La diferencias de Y' respecto a la media general representarán a la variabilidad explicada por el modelo. Sólo tiene un grado de libertad, pues todo depende del valor de β.

Suma de cuadrados no explicada (o de error):

$$SCT = \sum_i \sum_j (Y_{ij} - Y'_{ij})^2$$

En ella se suman las diferencias entre los valores reales de Y y los pronosticados Y'. Es decir, se suman los cuadrados de $\mathbf{e_{ij}}$. Representa, pues, la suma de errores, y de ahí su nombre. Le quedarán N-2 grados de libertad, por lo que el estimador de la varianza de la población será el cociente de esa suma entre N-2.

Estas tres sumas se pueden estructurar de forma similar al caso de ANOVA con un factor. Lo veremos con un ejemplo:

Se ha sometido a unos sujetos a unas horas de entrenamiento para una prueba en la que el número de aciertos depende en gran parte del manejo de un mando de juegos para ordenador de nuevo diseño. En la siguiente tabla se recogen los distintos niveles de tiempo de entrenamiento y las puntuaciones obtenidas en un determinado juego.

Tiempo en minutos	Resultados en puntos de 0 a 10
10	3 3 4 5 4 6 4
15	4 5 4 6 5 7 8
20	4 6 6 5 7 8 7 5 6
25	5 9 9 8 6 7 10 4 7
30	4 8 8 9 6 8 9 10 10

¿Se puede considerar que estos datos siguen un modelo de tipo lineal? ¿Cuál sería su ecuación? ¿Qué varianza presenta la población?

Aplicamos el ANOVA y nos queda:

Fuente variación	SC	G.L.	CM	F
Regresión	**69,39**	**1**	**69,39**	**28,34 P=0,00**
Error	**95,49**	**39**	**2,45**	
TOTAL	**164,88**	**40**	**4,12**	

La F=28,34 es claramente significativa, luego existe influencia de tipo lineal.

La estimación de la varianza de la población es el cuadrado medio de error, es decir, 4,12

La ecuación de la recta de regresión la obtendríamos por los métodos tradicionales y resultaría ser Y' = 2,43+0,187X

PRUEBA DEL ANÁLISIS DE REGRESIÓN

Podemos combinar el análisis de regresión con el de varianza para probar simultáneamente la anulación de la pendiente y la hipótesis de linealidad. El esquema de cálculo sería el mismo pero añadiendo también las sumas de cuadrados INTER e INTRA.

Sólo daremos el esquema al que daría lugar el ejemplo, pues se explica por sí solo:

Análisis de la regresión comparado con el ANOVA

Fuente variación	SC	G.L.	CM	F
INTER	**70,75**	**4**	**17,69**	**6,76**
Regresión	**69,39**	**1**	**69,39**	**26,54**
Desviación regresión	**1,36**	**3**	**0,45**	**0,17**
INTRA	**94,13**	**36**	**2,61**	
TOTAL	**164,88**	**40**	**4,12**	

En él aparece somo significativa la F-INTER, luego podemos afirmar que hay efecto de los niveles. También es significativa la F-REGR. y no lo es la desviación, por lo que nos reafirmamos en que el efecto de los niveles es de tipo lineal con pendiente no nula.

EJEMPLO DE REGRESIÓN

ANÁLISIS DE LA REGRESIÓN

Las técnicas del Análisis de la varianza se pueden aplicar también al estudio de la regresión lineal entre dos variables. Basta considerar los valores de X como niveles de un factor y sustituir las sumas INTER e INTRA por sus equivalentes REGRESIÓN y ERROR.

En tiempos de crisis se ha efectuado un estudio sobre el nivel de gasto de unas familias. Se han comparado cuatro niveles de ingreso familiar con el gasto mensual para intentar descubrir una relación lineal entre ambos. Los resultados, en miles de euros, han sido los siguientes:

Nivel en miles de euros	Gastos
1,5	1 1,2 0,9 1,4 1,5 1,3 1,2 1,1 1,4 1,3
2	1,5 1,5 1,9 2 1,8 1,7 1,5 1,3
2,5	2,4 2,5 2 1,7 2 1,8 1,9 1,8 2 2,4 2,5

3	2,4 2,3 2,6 3 2,8 2,7 2,8 2,6 3

¿Es significativa la relación lineal entre ambos? Expresado de otra forma, ¿Es la pendiente significativamente distinta de 0?

Volcamos estos datos en la tercera hoja (Regresión) del archivo anova.ods.

En las celdas L18 y L19 podemos leer los coeficientes de la ecuación de regresión Y' = 0,95789X - 0,24211. Podemos interpretar que cada incremento de un euro en el ingreso se traduce en un incremento de 0,95 en el gasto. Como de hecho se ha ahorrado más, querrá decir que hay una base fija del mismo (representada por -0,24), y que aumentos de ingreso no se traducen en incrementos proporcionales en el gasto, sino que hay una base fija que se dedica al ahorro.

Pero, ¿es significativo?

En el análisis de ANOVA vemos lo siguiente:

Fuente variación	SC	G.L.	CM	F
Regresión	**10,9**	**1**	**10,9**	**179,33**
Error	**2,19**	**36**	**0,06**	
TOTAL	**13,08**	**37**	**0,35**	

P-valor de F 0,000

Fcrítica 90%	**al**	**2,85**	**Significativa**
Fcrítica 95%	**al**	**4,11**	**Significativa**
Fcrítica 99%	**al**	**7,4**	**Significativa**

Casi toda la suma de cuadrados (13,08) es explicada por la regresión (10,9), por lo que F es claramente significativa a todos los niveles usuales. El error estimado es muy pequeño (0,06), lo que indica que la población es bastante homogénea.

Podemos, pues, afirmar que existe una relación lineal con pendiente significativamente distinta de cero, lo que traducido a lenguaje llano significa que sí existe influencia lineal de los ingresos en los gastos.

PEQUEÑO DICCIONARIO DE ESTADÍSTICA

A

Agrupación de datos

Si la variable que se estudia es continua, o discreta con muchos valores distintos, se organizarán sus datos en forma de intervalos. Para ello se fija un valor mínimo y otro máximo, de forma que todos los datos estén comprendidos entre ellos (a veces esto no se garantiza y quedan intervalos abiertos). La diferencia entre ambos se denomina rango de los datos y posteriormente se divide en un número de intervalos mediante valores intermedios.

Aleatorio

Experimento aleatorio

Un experimento se llama aleatorio cuando repetido indefinidamente presenta siempre resultados totalmente impredecibles.

Variable aleatoria

Llamaremos **Variable aleatoria simple** (discreta) a un conjunto de valores X_1, X_2, X_3, ...,X_n (llamados también sucesos) a los que les corresponden unos números (llamados probabilidades) , p_1, p_2, p_3, ..., p_n que cumplen:

a) Todas las probabilidades son positivas o nulas.

b) La suma de todas ellas es igual a la unidad

Amplitud

Se llama amplitud de un intervalo de datos agrupados a la diferencia entre los valores de sus extremos.

Aplastamiento

Sinónimo de curtosis.

Asimetría

Asimetría de una distribución de frecuencias es la característica por la que los datos pierden su simetría respecto a la media. Expresado de otra forma, es el mayor o menor grado de desviación que existe entre la media (reparto equitativo) y la mediana (punto medio de la distribución).

B

Bernouilli

Una distribución de Bernouilli se compone de dos sucesos contrarios A y B, a los que se les suele llamar éxito y fracaso, con probabilidades p y q respectivamente

Binomial

Distribución binomial

Esta importante distribución se aplica a pruebas repetidas de la ley de Bernouilli, con las siguientes condiciones:

a) Se realizan experimentos repetidos del tipo Bernouilli, n en total.

b) La probabilidad p permanece constante en todos ellos

c) Cada experimento es independiente del resultado anterior.

C

Campana de Gauss
Nombre asignado coloquialmente a la representación gráfica de la distribución normal.

Característica
Es cualquier propiedad de objetos o personas que deseamos estudiar en Estadística

Censo
Es el estudio y recuento de todos los elementos de una población.

Constante
Llamaremos constante a una característica que sólo admite una modalidad, por ejemplo la constante de gravitación universal

Continua
Una variable se llama continua si entre cada dos valores suyos pueden existir infinitos otros, como el peso, la estatura, etc.

Contraste
Contraste de hipótesis
Sinónimo de Test de hipótesis

Correlación

Coeficiente
Es el cociente de dividir la covarianza de una distribución bidimensional entre las desviaciones típicas de X e Y respectivamente.

Covarianza
Es la varianza conjunta en una distribución bidimensional X-Y. Se calcula como el cociente de los productos de las diferencias de X y de Y respecto a sus medias, entre el número de pares de la distribución.

Cualitativo/a
Se aplica a la variable (o dato, o medida) que sólo admite una medida nominal

Cuantil
Diremos que un número es el cuantil de orden p en una distribución de frecuencias si el porcentaje de datos inferiores a él es igual a p (y los superiores 100-p). Por ejemplo, el cuantil C85 será un punto que cumple que el 85% de los datos es inferior a él.

Cuantitativo/a

Se aplica a la variable que admite medidas de intervalo o de razón

Cuartil
Los cuantiles que dividen a la distribución en cuatro partes iguales, es decir, C25, C50 y C75, reciben el nombre de cuartiles, y se representan por Q1 o primer cuartil es el número que deja inferiores a él un 25% de los datos. Q2 o segundo cuartil o mediana es el número que deja inferiores a él un 50% de los datos. Q3 o

tercer cuartil es el número que deja inferiores a él un 75% de los datos.

Cuasivarianza

Cuasivarianza o varianza insesgada es similar a la varianza, pero dividiendo las sumas de cuadrados entre n-1.

Curtosis

Independientemente de su asimetría, una distribución puede presentar los datos con un reparto más uniforme, en el que las frecuencias sean muy parecidas. El gráfico aparecerá como aplastado y diremos que la distribución es **platicúrtica** o de poca curtosis. En el otro extremo, si las frecuencias cercanas al centro son mayores (con diferencia) que las alejadas, diremos que es **leptocúrtica** o con gran curtosis. Al caso intermedio lo denominaremos como distribución **mesocúrtica**

CH

Chi-cuadrado
Es la distribución teórica que representa la distribución muestral de la suma de cuadrados de los datos dividida entre la varianza de la población.

D

Dato
Es el valor cuantitativo o cualitativo que representa un atributo o medida en la población.

Decil
Se suelen definir 9 deciles D1, D2, ... D9, que son los puntos que dividen al intervalo en diez partes iguales, correspondientes a los cuantiles de porcentajes 10%, 20%, ...90% respectivamente.

Desviación

Desviación media

Es una medida de la dispersión consistente en la media aritmética de las desviaciones individuales respecto a la media, tomadas en valor absoluto. También se usan desviaciones respecto a la mediana.

Desviación típica

Es la raíz cuadrada de la varianza.

Determinación

Coeficiente

El coeficiente de determinación es el cociente entre la varianza explicada y la total en un ajuste a la recta de regresión.

Dicotómico/a

Adjetivo que se aplica a toda medida o proceso que sólo puede presentar dos valores, como SÍ/NO, Hombre/Mujer, Encendido/Apagado.

Discreta
Si una variable solo admite un número finito de valores entre cada dos, recibirá el nombre de discreta (edades medidas en años, número de hermanos, etc.).

Distribución

De frecuencias
El conjunto formado por los valores de la variable y sus frecuencias constituye la distribución de frecuencias de la población o muestra, y se representa en las tablas de frecuencias.

Bidimensional
Si en un experimento todas las medidas que se obtienen son dobles, pertenecientes a dos variables distintas, a las que llamaremos X e Y respectivamente, se denominará distribución bidimensional a la formada por los pares X-Y de valores relacionados en ambas variables.

Muestral
Distribución muestral es la resultante de considerar, de forma teórica, todas las posibles muestras que se puedan elegir. Es una distribución teórica, construida

sobre variables aleatorias, y sus elementos se obtienen mediante técnicas matemáticas.

Distribución teórica

Llamaremos distribución teórica a la correspondiente distribución de probabilidades en una variable aleatoria. Las principales distribuciones teóricas son:

Uniforme

Una distribución se llama uniforme cuando todas las probabilidades son iguales. Como todas suman 1 (caso discreto), cada una será igual a 1/n.

E

Error

De predicción

Es la diferencia entre un valor de Y y su estimación Y' en una recta de regresión (o en una curva de regresión general)

Típico de estimación

Es la raíz cuadrada de la varianza residual en una operación de estimación.

Escala

Escala de medida

Es un conjunto básico de modalidades y números (considerados como sus medidas) a partir del cual se construye un procedimiento para medir las restantes modalidades. Así, la escala centígrada de temperaturas se basa en asignar 0º a la temperatura de fusión del agua y 100º a la de ebullición

Escala nominal

Una escala se llama ***nominal*** si la única relación que tiene en cuenta es la de igualdad (y su contraria la desigualdad). Suele estar formada por nombres, códigos o números considerados como etiquetas (como el DNI). Así, son nominales los apellidos, la Comunidad Autónoma, el distrito postal, etc.

Escala ordinal

La escala ***ordinal*** añade a la nominal la posibilidad de ordenar los datos, es decir, considera las relaciones de

mayor y menor, aunque no se plantea una distancia entre unas medidas y otras. La escala de Insuficiente, Suficiente, Bien, Notable y Sobresaliente es ordinal.

Escala de intervalos

Se introduce una medida tipo (o patrón) llamada unidad y se tiene en cuenta cuantas unidades están comprendidas entre dos medidas distintas. Tienen sentido, además de la igualdad y el orden, las diferencias entre dos medidas. Podemos sumar y restar medidas, pero no tienen sentido sus cocientes. Son de intervalo la gran mayoría de las escala de las ciencias experimentales: temperatura, peso, velocidad, intensidad de la corriente eléctrica, etc.

Escala de razón

En esta escala se le da también un sentido a las razones entre dos medidas, es decir, las veces que una medida contiene a la otra. Fue la medida por excelencia de la Geometría griega y se ha trasladado a todas las Ciencias Sociales y de la Naturaleza. Se distingue también por la existencia de un cero verdadero, no convencional. Así, la escala centígrada de temperatura es sólo de intervalo y la Kelvin es de razón.

Esperanza
La esperanza matemática de una variable aleatoria discreta es la suma de los productos de sus valores por sus probabilidades. Equivale a la media en una distribución de frecuencias.

Estadístico
Se llama estadístico a todo valor numérico extraído mediante cálculos de los datos de una muestra. Normalmente se usa para estimar un parámetro de la población.

Estadístico de contraste
Es la expresión matemática, calculada a partir de la muestra, que nos servirá para tomar la decisión en un contraste de hipótesis.

Estimación
Es la operación por la que se asigna a un parámetro de la población el mismo valor que a un estadístico calculado a partir de una muestra.

Estimación Puntual
La estimación se llama puntual cuando identificamos, sin más, el parámetro con el estadístico. En ese caso

añadiremos un acento circunflejo al parámetro para representar que estamos estimando.

Estimación por intervalos

Al ser la estimación una operación arriesgada, en lugar de apostar por una estimación puntual, se rodea esta de un intervalo de seguridad, que es el Intervalo de confianza.

Estimador

Es un estadístico calculado en una muestra que estima un parámetro de la población. Los más importantes son los que estiman la media y la varianza.

Extremo inferior

Es el valor mínimo que puede tener un valor incluido en un intervalo de datos agrupados.

Extremo superior

Es el valor máximo posible en un intervalo de datos agrupados. Se considera no alcanzable. Así si un

intervalo comprende desde 5 hasta 10, incluiremos en el mismo los valores comprendidos entre estos dos, incluyendo el 5 y sin incluir el 10.

F

Frecuencia

Definición

El número de veces que se repite un valor concreto en una recogida de datos se llama frecuencia absoluta o simplemente frecuencia.

Frecuencia absoluta

Es sinónimo de frecuencia. Se representa por la letra n o por la f, según los distintos textos.

Frecuencia relativa o proporción

Es el cociente de dividir cada frecuencia absoluta entre el total de valores N. Se representa por f o por h.

Frecuencia acumulada

Es el número de datos del conjunto que son menores o iguales a u valor dado. Por tanto, se calculará sumando todas las frecuencias de datos menores o iguales al mismo. Podemos acumular las frecuencias absolutas y también las relativas y los porcentajes.

Frecuencias conjuntas

Son los pares de frecuencias formados en una distribución bidimensional

Frecuencia marginal

Llamaremos frecuencia marginal de un valor de X en una distribución bidimensional X-Y a la que le corresponde a ese valor si no tenemos en cuenta la existencia de Y. En la práctica coincide con la suma de todas las frecuencias contenidas en la fila correspondiente a ese valor.

Frecuencias condicionadas

Son las frecuencias que posee una variable si sólo consideramos un valor (o varios) de la otra variable en una distribución bidimensional X-Y. En la práctica se traduce a considerar sólo una fila o sólo una columna, según el valor elegido.

Función

Función de distribución

Llamaremos **función de distribución F(x)** de una variable aleatoria, a la formada por las probabilidades acumuladas, es decir: F(m) = Prob(x≤m) (El símbolo Prob designa a la probabilidad de que sea cierta la comparación del paréntesis)

G

Gauss

Distribución de Gauss

Sinónimo de distribución normal.

H

Hipótesis

Hipótesis nula

Llamaremos Hipótesis nula H_0. a la afirmación que hacemos sobre los parámetros de una población y cuya validez deseamos contrastar.

Hipótesis alternativa

Frente a la hipótesis nula podemos oponer otra, a la que llamamos hipótesis alternativa H_1. Suele ser una desigualdad que se opone a la igualdad que afirmamos.

Histograma

Representación gráfica de una distribución de datos agrupados en intervalos. Es similar al diagrama de barras, pero con los rectángulos adosados y de áreas proporcionales a las frecuencias de los intervalos.

I

Índice

Índice simple de base fija

Un término de la serie se identifica (convencionalmente) con el número 1, o el 100%. Diremos que este valor y_0 posee el índice 1. Para el resto de valores se define el índice como el cociente entre su propio valor y_i y el valor y_0 identificado como de índice 1.

Índice simple de base variable (o en cadena)

Tiene la misma definición que el anterior, pero en lugar de elegir un valor **y0** como base, en el cociente se toma el término anterior yi-1.

Índice compuesto

Cuando se desea comparar la evolución de varios conjuntos a la vez, se definen índices compuestos, obtenidos combinando los índices simples. Una técnica sencilla es sustituir los múltiples valores de cada término por su media ponderada.

Inferencia

Inferencia estadística
Es la ciencia que estudia las operaciones de estimación

Insesgado
Un estimador es insesgado cuando su media muestral coincide con el parámetro

Intervalo

Intervalos en distribuciones de frecuencias
Si la variable que se estudia es continua, o discreta con muchos valores distintos, se organizarán sus datos en forma de intervalos, que son conjuntos formados por los números reales comprendidos entre un máximo y un mínimo.

Intervalo de confianza
Es el intervalo del que se rodea una estimación puntual acompañada de una probabilidad de que el parámetro estimado pertenezca a dicho intervalo.

L

Leptocúrtica
Distribución de frecuencias con gran curtosis.

Ley
Ley de los grandes números
"Las frecuencias observadas tienen como límite las probabilidades cuando n tiende al infinito"

M

Marca de clase
Promedio entre los dos extremos (o punto medio de un intervalo de datos agrupados), que se elige como representante de todos los valores comprendidos.

Media

Media aritmética

Llamaremos media aritmética o simplemente media al valor resultante de sumar todos los datos y después dividir el resultado entre el número de ellos.

Media geométrica

Es la raíz enésima del producto de los datos. Se usa cuando el producto es más representativo que la suma, como ocurre cuando se promedian cocientes o razones.

Media armónica

Es la media diseñada para promediar cantidades inversamente proporcionales y equivale al inverso de la media de los inversos de x

Media cuadrática

Es muy usada en la teoría de errores y en estudios sobre ajustes de datos. Es la raíz cuadrada de la media de los cuadrados de los datos.

Media ponderada

En esta media se multiplica cada dato por un peso (valor numérico), se suman todos los productos se divide el resultado entre la suma e los pesos.

Mediana

Llamaremos mediana de un conjunto de datos de tipo ordinal (o de intervalo o razón) al dato que ocupa el punto medio de la distribución ordenada de datos. Es decir, es el punto que divide a la distribución en dos partes iguales: el total de frecuencias de los datos inferiores a la mediana es igual al de las frecuencias de los datos mayores.

Medida

Es la operación de asignar un número a cada una de las modalidades de una característica, convirtiendo algunas relaciones entre modalidades en sus correspondientes relaciones entre los números que representan su medida.

Medida directa

Llamaremos medida directa en cualquier estudio o experimento, a aquella que se ha obtenido directamente

sobre los objetos, individuos o entidades con los instrumentos usuales de medida.

Medida diferencial
Dada una medida directa X, llamaremos **medida diferencial** x a su diferencia con la media del grupo.

Medida típica Z
Si se divide una medida diferencial entre la desviación típica del grupo, se obtiene la medida o **puntuación típica Z**.

Mesocúrtica
Distribución de frecuencias con curtosis media.

Moda
Llamaremos **Moda** al valor de la distribución de datos que presente una frecuencia mayor.

Modalidad
Las distintas formas de presentarse una característica se llaman *modalidades*. Por ejemplo, 1,82 y 1,65 cm. son dos modalidades de la característica *altura*, y varón y mujer dos modalidades de la característica *sexo*

Muestra

Definición
Es un subconjunto de la población que es más fácil de estudiar que la población.

Muestreo
Definición
Es un conjunto de operaciones o técnicas dirigidos a la elección de la muestra adecuada.

N

Nivel de confianza
Es la probabilidad de que un valor estimado pertenezca al intervalo de confianza que rodea a la estimación. Los más usados son 90%, 95% y 99%

Nivel de significación
La probabilidad de que unos valores caigan en la región de rechazo n un contraste de hipótesis, a pesar de que

H_0 sea verdadera, se conoce con el nombre de nivel de significación a,

Normal

Distribución normal

La distribución Normal o ley de Gauss es la más usada de las distribuciones teóricas continuas. La popularizaron Gauss, en el estudio de los errores de las medidas, y también Laplace, pero ya la había usado Moivre como límite de la binomial.

Por su característica forma, se la conoce también como campana de Gauss. Aquí sólo nos interesa su definición y uso dentro de la Estadística. La expresión de su función de densidad con media 0 y desviación típica 1 es

$$f(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z^2\right)$$

O

Ordenada en el origen
Su significado más usual es el del término independiente de la ecuación de la recta de regresión. Se puede representar como el corte de esa recta con el eje Y.

P

Parámetro
Un número que caracterice o describa una población recibe el nombre de parámetro. La estatura media de los alumnos y alumnas de 16 años es un parámetro de esa población, o la Renta per cápita de la población española

Pendiente
Su significado más usual es el del coeficiente de la variable X en la recta de regresión lineal.

Percentil
Similares a los deciles, P1, P2, P3, ...P99, son 99 números que dividen la distribución en 100 partes iguales.

Poisson
Esta distribución, llamada de los sucesos raros, es el caso límite de la binomial, con las siguientes condiciones:
a) El número de intentos n debe tender a infinito.

b) La propiedad p debe ser muy pequeña (de ahí el nombre de suceso raro)

c) El producto de n.p ha de ser constante, y al que llamaremos m.

Platicúrtica
Distribución de frecuencias con poca curtosis.

Población
Llamaremos población a un conjunto bien definido por ciertas características que deseamos estudiar: La población de una Comunidad Autónoma, los aprobados de 2º de Bachillerato en mi Centro, los profesores de E.S.O. en la Delegación Norte, etc.

Porcentaje
Equivale a la frecuencia relativa expresada como tanto por ciento o porcentaje.

Predicción
Llamaremos pronóstico o predicción para un valor de X a su imagen Y´en la recta de regresión. Esta definición se extiende a cualquier otra curva de ajuste de datos.

Proporción
Es sinónimo de frecuencia relativa

P-valor
El p-valor de un resultado en un experimento es la probabilidad de obtener ese valor u otros menores (o

mayores, según sea el experimento) si se satisface la hipótesis nula.

R

Rango

Si se fija un valor mínimo y otro máximo, de forma que todos los datos de un recuento estén comprendidos entre ellos (a veces esto no se garantiza y quedan intervalos abiertos), la diferencia entre ambos se denomina **rango** de los datos.

Rango percentil

Es la medida inversa del percentil. Dada una medida concreta, como puede ser la calificación de una alumna en Música, su rango percentil equivale al percentil más cercano a esa calificación. Un alumno que tenga rango percentil de 78 es aquel en el que el 78% de sus compañeros tiene una puntuación inferior a él.

Regresión

Recta de regresión

La recta de regresión de Y sobre X es aquella que minimiza la suma de cuadrados de las diferencias entre los valores de Y y los correspondientes Y' medidos en dicha recta.

S

Sesgo

Sinónimo de asimetría

Sumas de cuadrados

En ANOVA

Total

Es la suma de las diferencias al cuadrado entre los datos experimentales y su media.

Intragrupos

Representa la suma de cuadrados corregidos que se da dentro de los grupos, es decir, las diferencias de los datos entre la media de cada grupo.

Intergrupos
Es la suma ponderada de las diferencias al cuadrado entre as medias de los grupos y la media total.

Interacción
En un modelo con varios factores representa la influencia mutua entre ellos.

Supuesto
Es una afirmación que se hace de una población en la Estadística Inferencial: si es es simétrica, normal, continua… y sobre la muestra, si es aleatoria simple, es de tamaño mayor que 30…

T

T de Student
Distribución que sigue la estimación de la desviación típica.

Teorema

Teorema central del límite

Si las variables x_1, x2, x3,... xn, tienen todas la misma distribución, con los mismos valores m para la media y s para la desviación típica, la variable

$$\frac{x_1 + x_2 + \ldots x_n - nm}{s\sqrt{n}}$$

sigue asintóticamente la distribución normal N(0,1).

Test

Test de hipótesis

Un test de hipótesis (o contraste) es un proceso, compuesto de varios pasos muy concretos, que nos permite aceptar o rechazar una hipótesis en términos estadísticos.

Tipificación

Es la operación de convertir una medida en típica restándole la media y dividiendo entre la desviación típica.

V

Variable

Variable aleatoria

Llamaremos ***Variable aleatoria simple*** (discreta) a un conjunto de valores $\mathbf{X_1, X_2, X_3, ...,X_n}$ (llamados también *sucesos*) a los que les corresponden unos números (llamados *probabilidades*) , $\mathbf{p_1, p_2, p_3, ..., p_n}$ que cumplen:

a) Todas las probabilidades son positivas o nulas.
b) La suma de todas ellas es igual a la unidad

Variación

Coeficiente de variación

Es el cociente de dividir la desviación típica entre la media.

Varianza

Definición

Es el cociente de dividir la suma de los cuadrados de las desviaciones de los datos respecto a la media entre el número total de datos. Su raíz cuadrada es la desviación típica.

Explicada

Es la parte de una varianza que se considera producida por un factor determinado que influya en un experimento. En la regresión lineal es la varianza de las predicciones.

Total

Es la varianza total observada en un experimento, independientemente de las variables que puedan influir en los resultados.

Residual

Es la diferencia entre la varianza total y la explicada.

Análisis de Varianza

La técnica del Análisis de la Varianza consiste en descomponer la variabilidad de una población (representada por su varianza) en diversos sumandos según los factores que intervengan en la creación de esa variabilidad.