

Curso 2009-2010

Apuntes

de

BIOESTADISTICA

**Por Eduardo Buesa Ibáñez, Profesor de la asignatura en la Escuela
Universitaria de Enfermería Nª Sª del Sagrado Corazón. Castellón**

BIOESTADISTICA

OBJETIVOS: Realizar una introducción elemental en el campo de la Metodología Estadística para que el futuro Diplomado sea capaz de aplicar los procedimientos estadísticos fundamentales y valorar críticamente los informes y publicaciones que hagan uso de tales métodos.

CONTENIDOS: Temas de Estadística Descriptiva, de Estadística Inferencial y de algunas aplicaciones concretas de la Estadística en las Ciencias de la Salud. El alumno aprenderá a recoger datos procedentes de muestras, a ordenarlos y a presentarlos en forma de tablas, gráficos y números índice que los resumen (media, varianza, desviación estándar, etc). Además aprenderá a estimar parámetros y a realizar pruebas de conformidad, relación y contraste de variables.

METODOLOGIA: Exposición de los temas. Realización de más de 200 ejercicios prácticos. Manejo de programas estadísticos libres y gratuitos.

EXAMENES: Ante todo, resolución de uno o varios supuestos prácticos. Alguna pregunta sobre teoría "tipo test" o a contestar en una o dos líneas.

PROGRAMA:

Tema 1	Fundamentos y fines de la Bioestadística.
Tema 2	Operaciones matemáticas más usuales en Bioestadística.
Tema 3	Variables y su medida. Síntesis de datos estadísticos.
Tema 4	Tabulación de datos.
Tema 5	Representaciones gráficas.
Tema 6	Indices estadísticos de variables cuantitativas. Parámetros de tendencia central, dispersión, posición y forma.
Tema 7	Datos bivariados. Tabulación y representación gráfica. Correlación y regresión.
Tema 8	Series de tiempo.
Tema 9	Teoría de la probabilidad
Tema 10	Distribuciones fundamentales de probabilidad (normal, binomial, de Poisson). Otras distribuciones.
Tema 11	Planificación de estudios estadísticos. Clases de estudios.
Tema 12	Recogida de la información. Técnicas de muestro. Errores de los muestreos.
Tema 13	Intervalos de probabilidad y confianza. Hipótesis y decisiones estadísticas.
Tema 14	Estimación de parámetros. Pruebas de conformidad
Tema 15	Pruebas de contraste de variables.
Tema 16	Contraste de dos variables cualitativas. Odds ratios.
Tema 17	Contraste de una variable cualitativa y otra cuantitativa.
Tema 18	Contraste de dos variables cuantitativas.
Tema 19	Demografía sanitaria. Medida de la salud.
Tema 20	Errores de las medidas de laboratorio. Control de calidad. Valoración de pruebas diagnósticas
Tema 21	Programas para resolver problemas estadísticos.
Tema 22	La Estadística en Internet

Libros de consulta recomendados

- ESTADISTICA PARA LA INVESTIGACION BIOMEDICA. P Armitage, G Berry.
Edit. Doyma, Barcelona
- BIOMETRÍA. RR Sokal, FJ Rohlf. Ediciones Blume, Madrid.
- ESTADISTICA. Gilbert. Ed. Interamericana, Madrid
- ESTADISTICA PARA BIOLOGIA Y CIENCIAS DE LA SALUD. JS Milton.
Edit. McGraw-Hill, Madrid

Tema 1 : Fundamentos y fines de la Bioestadística

--Conceptos básicos

La **BIOESTADISTICA** es la Estadística aplicada a las ciencias biológicas.

La **ESTADISTICA** es muy difícil de definir. Esto hace que haya muchas definiciones y que incluso algunos libros la soslayan. Una definición aceptable es :**”La Estadística es el estudio científico de datos numéricos referidos a características variables”**.

Un estudio es científico si utiliza métodos rigurosos en su concepción y desarrollo, teniendo como normas básicas la objetividad, el espíritu crítico y la ética. Algunas afirmaciones aparentemente científicas no lo son al no cumplir alguna de estas normas básicas. Es frecuente cuando se tocan temas religiosos, políticos o económicos. Incluso los muy expertos en una materia no están libres de prejuicios y presiones crematísticas.

Los datos numéricos son números que expresan medidas (datos métricos) o recuentos de modalidades (datos categóricos).

Por característica se entiende una propiedad o condición claramente reconocible en diversos individuos. El individuo es la unidad estadística y puede ser una persona, un animal, una planta, un objeto o una acción. Las características pueden ser constantes o variables.

Las constantes no varían, siempre ocurren de la misma forma, como las constantes físicas o la certeza de la muerte en los seres vivos. Siguen el llamado modelo determinista de los fenómenos naturales. Tienen un resultado fijo, que se puede resumir por una fórmula matemática. Al lanzar una bola es posible saber con exactitud la velocidad y la aceleración que va a tener en un determinado momento.

Las variables presentan una gama de variaciones (al menos dos) en los diversos individuos, como el sexo o la talla de las personas. Siguen el modelo indeterminista (= probabilístico, casual o estocástico). No tienen un resultado fijo. Hay un conjunto de posibles resultados, conocidos de antemano, de los que sólo se producirá uno. Los factores que influyen en que se produzca ese resultado u otro son múltiples, complejos, incontrolables y en parte desconocidos, de forma que el resultado ocurre de forma aparentemente casual, al azar. El azar no es ciego, tiene sus modelos de comportamiento, predecibles con un margen de variación mediante fórmulas matemáticas, basadas en el cálculo de probabilidades. Son las llamadas distribuciones fundamentales de probabilidad (Distribución normal, de Poisson, binomial, hipergeométrica, etc.). Los fenómenos biológicos siguen uno u otro modelo, que una vez conocido nos permite calcular las probabilidades de que ocurra tal o cual resultado. ¡EL AZAR ES LA SUPREMA LEY DE LOS FENÓMENOS BIOLÓGICOS!.

En Estadística sólo interesan las características variables, que habitualmente son denominadas variables, sin más aditamentos.

--Etimología e Historia

Estadística proviene de Estado, ya que fueron los Estados los que iniciaron la recogida de datos para su mejor funcionamiento (impuestos, soldados...). Así, hay constancia histórica de censos de tierras y hombres en Egipto 3000 años A.C., en China 2200 años A.C. y en Israel (Moisés y David, 1500 y 1000 años A.C.). En los Evangelios se dice que Jesús nació cuando su familia se trasladaba para cumplimentar el censo ordenado por el César. Por este origen se han introducido términos “humanos” en el lenguaje estadístico, como individuo y población.

Esta Estadística era muy elemental, fundamentalmente recuentos. A partir del siglo XVII experimenta un gran impulso, que se intensifica en siglos posteriores. Se hace científica. En este desarrollo hay que destacar como motores importantes:

1. Los juegos de azar, sobre todo el de dados, que fascinaron a matemáticos insignes y de cuyo estudio nació la teoría de la probabilidad.
2. La Astronomía, con su interpretación de observaciones, cuantificación de posibles errores de medida y predicción de eventos.

3. La Agricultura, con sus estudios genéticos y de productividad.

4. Las compañías de Seguros norteamericanas, con sus estadísticas vitales y estudios de supervivencia y de los factores que más influyen en la misma (edad, tensión arterial, obesidad...)

Nombres como De Moivre, Bernouilli, Lagrange, Laplace, Gauss, Pascal, Quetelet, Galton, Spearman, Pearson y Fisher ocupan un lugar destacado en el progreso de la Estadística.

POBLACIONES Y MUESTRAS

Población: todos los individuos que poseen una determinada característica.

Por su tamaño las poblaciones pueden ser finitas o infinitas. En la práctica, y para facilitar los cálculos, una población se considera “infinita” a partir de un tamaño de 10.000 individuos.

La obtención de datos de una población se llama censo.

Teóricamente un individuo puede tener infinitas características y por tanto puede formar parte de infinitas poblaciones.

Muestra: es una parte de la población, un subconjunto de la misma. Cuando la muestra es representativa de la población, se pueden hacer extensivos a la población los resultados obtenidos en la muestra. En el tema 12 se estudian las muestras con detalle. Aquí se puede adelantar que la representatividad, el que la muestra reproduzca lo más fielmente posible a la población de la que procede, depende fundamentalmente de dos factores: un tamaño adecuado y la elección de los individuos al azar.

Un conjunto de individuos, según las circunstancias, puede ser población o muestra. Por ejemplo, los alumnos de esta Escuela serán “población” cuando tomemos a unos cuantos de ellos para estimar la talla de todo el alumnado de la Escuela. Y serán “muestra” si toda la Escuela ha sido seleccionada para participar en un estudio a nivel nacional.

Hay muchos sinónimos para los conceptos estadísticos:

Bioestadística: Biometría, Estadística biológica...

Población: universo, colectivo, conjunto...

Individuo: elemento, sujeto, efectivo, caso...

Dato: observación, registro, resultado...

CLASES DE ESTADISTICA

Hay que distinguir entre Estadística descriptiva y Estadística inferencial.

La E. descriptiva es la parte más antigua y la más conocida por los profanos. Comprende la obtención, clasificación y presentación de datos numéricos mediante tablas, gráficos, frecuencias, porcentajes, etc. . La vida diaria está invadida por estadísticas de este tipo: de consumo, producción, accidentes, desempleo, etc.

La E. inferencial (o deductiva) es la parte más moderna y científica. A partir de una muestra representativa permite sacar conclusiones razonablemente válidas para la población de origen (Problemas de estimación). Además permite contrastar variables (Problemas de contraste) y concluir si las diferencias o relaciones observadas son explicables o no por el azar.

La E. inferencial *clásica* proporciona un conjunto de “recetas” para realizar las inferencias. Modernamente se ha desarrollado con bastante éxito una variante, la E. *bayesiana*, que se basa en probabilidades condicionadas y que es la base del diagnóstico por computadora.

LA ESTADISTICA, ¿CIENCIA INEXACTA?

Aunque utiliza herramientas matemáticas, las conclusiones estadísticas no son dogmáticas. Incluyen un margen de variación (el llamado intervalo de confianza) y un grado de fiabilidad (nivel de aceptación o significación). Si se estudia por medio de una muestra la opinión de la población de Castellón sobre un determinado asunto y se encuentra que al 65% le parece bien, la Estadística dirá que el 65% está a favor , pero añadirá que este resultado tiene un margen de variación

del, digamos, 10% por encima y debajo de ese valor puntual obtenido y que esta afirmación se hace con una probabilidad de acierto del 95% (o probabilidad de error del 5%).

Es importante destacar que las conclusiones de los estudios estadísticos inferenciales son válidas a nivel de grupo. A nivel individual pueden no serlo por la existencia del llamado error muestral, que suele ser muy pequeño, pero nunca cero. Ejemplo: el medicamento A es eficaz en el 95% de los pacientes con la enfermedad X; el medicamento B sólo en el 5%. Un estudio estadístico permitirá sin duda concluir que el medicamento A es el de elección. La inmensa mayoría se curará sólo con el A. Pero habrá pacientes, pocos ciertamente, que se curen con el B y no con el A.

En la vida diaria se abusa mucho de expresiones como “estadísticamente demostrado” o “estadísticamente comprobado”. En realidad la Estadística no demuestra nada, sino que apoya con la fuerza de una probabilidad una determinada conclusión. Admite siempre una probabilidad de equivocarse, que aunque sea muy pequeña, ocurrirá de vez en cuando. Es una ayuda para la toma de decisiones razonables en caso de incertidumbre, aportando las probabilidades de éxito y fracaso de una decisión.

Por otra parte la existencia de una correlación entre dos cosas sólo permite establecer una relación de causalidad si se cumplen determinadas condiciones, ya que puede tratarse de correlaciones espurias, a veces difíciles de descubrir. Dos ejemplos: 1) si en una ciudad se comprueba que la venta de música clásica aumenta a la par que los espectadores que acuden al campo de fútbol, sería muy aventurado concluir que la visita de los campos estimula la afición musical clásica 2) Bernard Show destacó que los londinenses que usaban paraguas estaban mejor nutridos, gozaban de mejor salud y vivían más que los que no lo usaban. Sería peregrino pensar que eso se debía al paraguas. Más bien parecía deberse a que en aquellos tiempos los que usaban paraguas eran los ricos, que disfrutaban de una vida más saludable. En los medios de comunicación, en las argumentaciones de los políticos y grupos de presión e incluso en las publicaciones científicas se utilizan de forma mucho más sutil que en los ejemplos anteriores, de forma más o menos consciente, “conclusiones” estadísticas para hacer comulgar al lector u oyente con grandes ruedas de molino. La Estadística es siempre honesta. los que la utilizan a veces no.

DOS OPINIONES ILUSTRES SOBRE LA ESTADISTICA

1. Hay tres clase de mentiras: mentiras, mentiras viles y estadísticas (Disraeli)
2. El buen cristiano debe guardarse de los matemáticos y de los que practican la predicción... porque existe el peligro de que esta gente esté aliada con el diablo. (San Agustín)

...Y OTRA OPINION ALGO MENOS ILUSTRE...

Y todo esto...¿para qué sirve? (Un antiguo alumno de esta Escuela)



Fisher

Tema 2 : OPERACIONES MAS USUALES EN ESTADISTICA

---OPERACIONES

- 1) Las "4 reglas" clásicas : sumar, restar, multiplicar y dividir.
- 2) Potenciación : a^n , generalmente a^2 . Recordar que $a^0=1$ y $a^1=a$
- 3) Radicación : casi exclusivamente la raíz cuadrada
- 4) Resolución de ecuaciones : nosotros sólo veremos de primer grado
- 5) utilización del sistema de coordenadas rectangulares (x , y), a veces los 4 cuadrantes, pero habitualmente sólo el primer cuadrante.
- 6) logaritmos y antilogaritmos. Fáciles de obtener con una calculadora científica (log , ln , 10^x , e^x)
- 7) Factoriales : $n!$, que es igual a $n*(n-1)*(n-2)*(n-3)*.....*1$. Recordar que $1!=1$ y $0!=1$
- 8) Cálculo del número combinatorio o coeficiente binomial , n sobre r , que desarrolla los coeficientes del binomio de Newton

$$\binom{n}{r} = \frac{n!}{r!(n-r)!}, \text{ donde } r \text{ va tomando sucesivamente los valores } 0, 1, 2, 3, \dots, n$$

$$\binom{n}{0} = 1 ; \binom{n}{n} = 1$$

---ALGUNOS DE LOS SIMBOLOS EMPLEADOS

-operadores matemáticos

+ suma (a+b) ; - resta (a-b) ; * , . , **nada** : multiplicación (a*b , a.b , ab) ;
: , / , — división (a:b , a/b , $\frac{a}{b}$) ; ± más-menos (sumar y restar) ; = igual ;

≈ aproximadamente igual ; < menor ; > mayor ; ≤ igual o menor ;

≥ igual o mayor ; ≠ , <> (<>) no igual, distinto

|a| valor absoluto de a, siempre positivo ; $\sum X^2$ suma de todos los cuadrados de X ;

$(\sum X)^2$ el cuadrado de la suma de todas las X.

-otros

Δ incremento ; α letra griega alfa ; β letra griega beta ; λ letra griega lambda ; r coeficiente de correlación ; $E(a \div b)$ intervalo que va desde a hasta b ; Σ sumatorio abreviado,

que para simplificar es el único que utilizaremos. El símbolo normal es $\sum_{i=1}^{i=n} x_i$, que quiere decir sumar todos los valores de x, desde el primero hasta el que ocupa el lugar n . si la variable x vale 10 , 12 y 14 , $\sum X=36$

Clásicamente se utilizan letras griegas para simbolizar parámetros de poblaciones y letras latinas para las muestras. Aquí se utilizarán en aras de la sencillez siempre letras latinas tanto para poblaciones como para muestras, poniendo en caso de que pueda haber duda o confusión el subíndice p o m.

---LECTURA DE FORMULAS

consiste en traducirlas al lenguaje gramatical y lógico, separándolas en sus distintas partes, lo que nos permitirá resolverlas.

$F = \sqrt{\frac{\sum (x-5)^2}{2}}$ quiere decir: a cada valor de la variable x le restamos 5 y esta diferencia la

elevamos al cuadrado; luego sumamos todos los resultados obtenidos; esta suma se divide por 2 ; finalmente se extrae la raíz cuadrada del cociente. Así obtenemos el valor de F. No hay que asustarse de fórmulas muy complejas que se resuelven de forma similar, por partes. Como dice un proverbio indio: es posible comerse todo un elefante siempre que sea a trocitos...

---RESOLUCION DE LOS CALCULOS ESTADISTICOS

Muchos se pueden resolver manualmente, utilizando lápiz , papel y los conocimientos adecuados, facilitando el trabajo las calculadoras de bolsillo. Con una calculadora científica sencilla se pueden resolver todos los problemas de esta asignatura. Es absolutamente necesario estar familiarizado con el manejo del aparato para evitar errores. Existen programas estadísticos para ordenadores, algunos gratuitos, que se verán en los temas 21 y 22 . La hoja de cálculo **Excel** permite resolver muchos problemas. En todo caso, si no se sabe Estadística, el ordenador y los programas sirven de muy poco.

---REDONDEO DE NUMEROS

Redondear un número es expresarlo por otro más corto, con menos cifras; en general comporta una pequeña pérdida de exactitud. El redondeo puede hacerse voluntariamente para obtener números más manejables o más fácilmente comprensibles. En otros casos el redondeo es obligado, como en el caso de tener que expresar un número con la sensibilidad que le corresponde (cifras significativas). Cualquier número puede redondearse, pero sobre todo se aplica a números con muchas cifras, poco frecuentes en Estadística, o con decimales. En este último caso el redondeo se indica diciendo el nº de decimales deseado o bien el lugar del redondeo (décimas, centésimas, milésimas...).

Regla general del redondeo: se redondea al número más próximo. Siempre hay dos opciones, una por encima y otra por debajo del número original.

Ejemplos: 4,1 redondeado a enteros es 4 (hay que elegir entre 4 y 5; el 4 está más cerca).
25,8 redondeado a enteros es 26 , que es el número más próximo entre 25 y 26
3,1785 redondeado a 2 decimales es 3,18 (se elige entre 3,17 y 3,18)
3,141592 redondeado a todos los lugares posibles::

redondear a	elección entre		nº redondeado
unidades	3	4	3
1 decimal	3,1	3,2	3,1
2 decimales	3,14	3,15	3,14
3 decimales	3,141	3,142	3,142
4 decimales	3,1415	3,1416	3,1416
5 decimales	3,14159	3,14160	3,14159

Caso especial del 5 como última cifra para redondear al lugar anterior : se redondea al número par.

Ejemplos: 2,5 (¿2 ó 3?) → **2** ; 2,55 (¿2,5 ó 2,6?) → **2,6** ;
2,145 (¿2,14 ó 2,15?) → **2,14** ; 2,1235 (¿2,123 ó 2,124?) → **2,124**

Más ejemplos:

$$5! = 5 * 4 * 3 * 2 * 1 = 120$$

$$\binom{n}{r} = \frac{n!}{r!(n-r)!} ; \quad \binom{5}{3} = \frac{5!}{3!*2!} = 10$$

$$\sum x \quad \sum x^2 \quad (\sum x)^2 :$$

$$\text{si } x = (1, 2, 3, 5) :$$

$$\sum x = 11 \quad \sum x^2 = 39 \quad (\sum x)^2 = 121$$

redondear 6'28945 a todos los lugares posibles:

6 6'3 6'29 6'289 6'2894

Tema 3: Variables. Medidas. Síntesis de datos estadísticos.

--Variables. Como ya se vio en el tema 1, las variables son características que se distinguen por la variabilidad con que se manifiestan en los diversos individuos.

--Tipos de variables.

Hay variables: cualitativas (CL) y cuantitativas (CT)

nombre	datos	expresión	variantes	ejemplo	
CUALITATIVAS O ATRIBUTOS	Catagóricos	modalidades o categorías	2 modalidades más de 2 mod.	sexo caras dado	mujer-hombre 1, 2, 3, 4, 5, 6
CUANTITATIVAS	métricos	valores	-continuas -discretas	talla nº hijos	170 cm. 0, 1, 2, 3,

--Medida de las variables

Se hace según las llamadas escalas. Básicamente hay 4 escalas de medidas:

- nominales
- ordinales
- de intervalo
- de razón

Las variables ordinales son una variante de las nominales y las de razón de las de intervalo.

--Escalas nominales

Se utilizan para medir atributos, es decir, variables cualitativas. Se da un nombre a cada una de las modalidades, se asignan los individuos a ellas y se cuentan los individuos de cada modalidad (frecuencia). El orden en que se designan las modalidades es indiferente, p.e. alto y bajo o bajo y alto.

Ejemplo: la variable sexo tiene dos modalidades, hombre y mujer. Medimos este atributo en 100 personas y encontramos 52 hombres y 48 mujeres.

En vez de dar un nombre convencional a las modalidades se las puede designar con un número, lo que facilita sobre todo el tratamiento informático. Estos números son realmente un nombre y por tanto no pueden hacerse con ellos operaciones matemáticas. Así podríamos llamar a los hombres "1" y a las mujeres "2" (ó 7 y 8...)

--Escalas ordinales

Una escala ordinal es una escala nominal en la que las diversa modalidades guardan entre sí una relación de orden o jerarquía, que debe ser respetada, siendo indiferente que el orden sea de mayor a menor o viceversa. Ese orden viene marcado por el sentido común y también por la costumbre.

Un ejemplo clásico son las notas académicas tradicionales : sobresaliente-notable-aprobado-suspenso o suspenso-aprobado-notable-sobresaliente. En la variable "evolución de la enfermedad" podríamos distinguir las siguientes modalidades : muerto-peor-igual-mejor-curado , o bien, curado-mejor-igual-peor-muerto.

También pueden emplearse números como nombre de modalidades, pero respetando el orden. Podríamos hacer muerto=1, peor=2, igual=3, mejor=4, curado=5 . O bien, curado=1, mejor=2, igual=3, peor=4 , muerto=5 .

--Escalas de intervalo

Se utilizan para medir variables cuantitativas cuando no hay cero absoluto en la zona de medición, lo que permite valores negativos. El cero se asigna arbitrariamente así como la unidad de medida.. La escala ha sido diseñada de tal manera que sus números permiten valorar exactamente la diferencia que hay entre dos medidas (= intervalo). Ejemplo típico es la temperatura medida de la forma habitual, lo que puede hacerse de diversas maneras. En Europa se mide en grados

centígrados o Celsius (C). El “0” se asigna a la temperatura de congelación del agua destilada y el “100” a su temperatura de ebullición. Ese intervalo se divide en 100 partes y así se obtienen los grados centígrados. En USA se mide en grados Fahrenheit (F). 0° C equivalen a 32° F y 0° F equivalen a -17,78° C. Por tanto 32° C no representa el doble de calor que 16° C, simplemente el doble de grados C. Esas temperaturas medidas en grados Fahrenheit serían 0° F y -8,9° F. Un niño con un proceso febril en Castellón puede tener 40° C de fiebre; en USA tendría 104° F. Por la Física sabemos que hay un mínimo infranqueable de temperatura, el llamado “cero absoluto”, que en grados centígrados corresponde a -273,15°. Este cero no significa la ausencia de temperatura, sino el mínimo de temperatura posible. La escala de Kelvin asigna su 0 a esta temperatura.

--Escala de razón

Se utilizan para medir variables cuantitativas cuando hay un cero absoluto, siendo la unidad de medida lo único arbitrario. Una longitud puede ser medida en cm., Km., yardas, varas, etc. pero el cero es el mismo para todos. El tiempo de reacción a un estímulo siempre empieza en cero cualquiera que sea el sistema que utilicemos para medir el tiempo. Aquí sí puede decirse que una persona que pesa 50 Kg. pesa el doble que un niño que pesa 25. Y que la diferencia de peso entre una persona que pese 80 Kg. y otra que pese 50 Kg. es la misma que la existente entre dos piedras de 35 y 5 Kg., respectivamente. No hay valores negativos.

--Variables cualitativas

Las variables cualitativas (CL) o atributos se miden por escalas nominales u ordinales según corresponda. Cuando sólo tienen dos modalidades se llaman dicotómicas. Ejemplos: cara-cruz, varón-hembra, vivo-muerto. Todos los atributos, con independencia del número de modalidades que tengan, pueden ser siempre reducidos a dicotómicos si así se desea. Los 4 palos de la baraja española (oros, copas, espadas y bastos) pueden ser reducidos a oros-no oros, bastos-no bastos, etc. ; las marcas de coches a Seat-no Seat. ; el estado civil a casado-no casado...

--Variables cuantitativas

Las variables cuantitativas (CT) se miden por escalas de intervalo o de razón, según su naturaleza. Pueden ser continuas o discretas.

Una variable CT es continua cuando puede tomar cualquier valor en su zona de variabilidad. Son continuas la talla, el peso, la tensión arterial, el contenido de un frasco, la glucemia, etc.

Las variables CT discretas no pueden adoptar cualquier valor, sino solamente ciertos valores.

Una familia puede tener 0, 1, 2, 3, ... hijos, pero no 3,1416 hijos. El nº de pacientes que ingresa en un hospital, el nº de ataques que sufre un paciente en un mes, el nº de cápsulas de un envase medicamentoso... son discretas.

Una variable CT continua se mide a menudo, porque resulta más práctico, de forma “discretizada”. La edad suele expresarse en años enteros, o en meses en los niños pequeños, pero no por eso deja de ser continua.

--Transformación de variables

Las variables cuantitativas pueden ser transformadas en cualitativas, dicotómicas o no, con una pérdida en la calidad de la medida, que a veces se asume si mejora la información. La talla podemos medirla en alta-normal-baja. Los valores de colesterol en mayor de 200 mg/dl - igual o menor de 200 mg/dl. Como la variable CT proporciona más información que la CL debe ser usada siempre que no sea más conveniente hacerlo de forma cualitativa.

Las variables CL en cambio no pueden ser transformadas en CT.

Las variables CL son por su propia naturaleza discretas.
Por las limitaciones de los instrumentos de medida la mayoría de las CT continuas son discretizadas.

Dos ejemplos:

---variable “INGESTION DE ALCOHOL” .

He seleccionado 4 formas distintas en orden creciente de información:

- | | |
|--|--|
| 1) abstemio – bebedor | Variable CL con dos modalidades, nominal. |
| 2) abstemio – bebedor – alcohólico | Variable CL con tres modalidades, ordinal. |
| 3) nº de copas o vasos bebidos en una semana | Variable CT discreta |
| 4) gramos de alcohol tomados en una semana | Variable CT continua |

---“ESTUDIO DE 3 TRATAMIENTOS DE LA ISQUEMIA CORONARIA”.

Considerando las variables:

- | | |
|---------------------------------|-------------------------------|
| - sexo : hombre – mujer | CL con 2 modalidades, nominal |
| - medicamento: A – B – C | CL con 3 modalidades, nominal |
| - nº ataques del día anterior | CT discreta |
| - distancia caminada sin disnea | CT continua |

--Necesidad de una definición clara de las variables

Es esencial que todo el mundo sepa qué se está midiendo y cómo. Está claro lo que es medir el peso en Kg. o la talla en cm. Pero, ¿que es ser “fumador”? ¿El que fuma un pitillo, aunque sea una vez al año? ¿O el que fuma cada día o al menos cada tres?... Hay que concretar y decir por ejemplo: “en este estudio se considera fumador a quien fuma al menos un cigarrillo cada semana” o “se considera desnutridos a los niños que en los gráficos peso/talla de Tanner están por debajo del percentil 3”, etc., etc.

--Dominio de una variable

Es el conjunto de valores o modalidades que puede adoptar. El dominio de la variable CL “puntuación de la cara de un dado” es (1, 2, 3, 4, 5 y 6). El de la variable sexo: (hombre, mujer). El de la “longitud de las hojas de la planta P” cualquier valor entre 1 y 8 cm. o $\in (1 \div 8)$, etc.

--Variables aleatorias y controladas

Una variable es controlada o independiente cuando su valor o la modalidad elegida en cada uno de los individuos depende únicamente del investigador. En un estudio podemos seleccionar sólo individuos del sexo masculino. O fijar la dosis de medicamento que se da a los ratoncillos, etc. Una variable es aleatoria o dependiente cuando su valor en cada uno de los individuos no depende del investigador, sino de la naturaleza o reacción del propio individuo. Por ejemplo la talla de los alumnos de una clase, la tensión arterial de un grupo de pacientes, etc.

--Medida de una variable continua

Debido a la imperfección de los instrumentos de medida, aún de los más sofisticados, el valor exacto o real de una medida (**Xe**) es realmente desconocido y sólo podemos expresarlo de una forma aproximada mediante el valor medido (**X**). Supongamos que estamos midiendo una longitud con una regla graduada. Cuando la medida no se corresponde con un valor marcado en la regla, hay que aproximar (=redondear) a la marca más cercana. Si hay equidistancia se aproxima al valor par.

	5	6	7	8
	-x - -	- - -x -	- -x - -	
medida:	5	7	8	

La diferencia entre el valor exacto y el valor medido se llama ERROR ABSOLUTO. Toda medida tiene su error.

$$\boxed{E = |X_e - X| \text{ y por tanto } X_e = X \pm E}$$

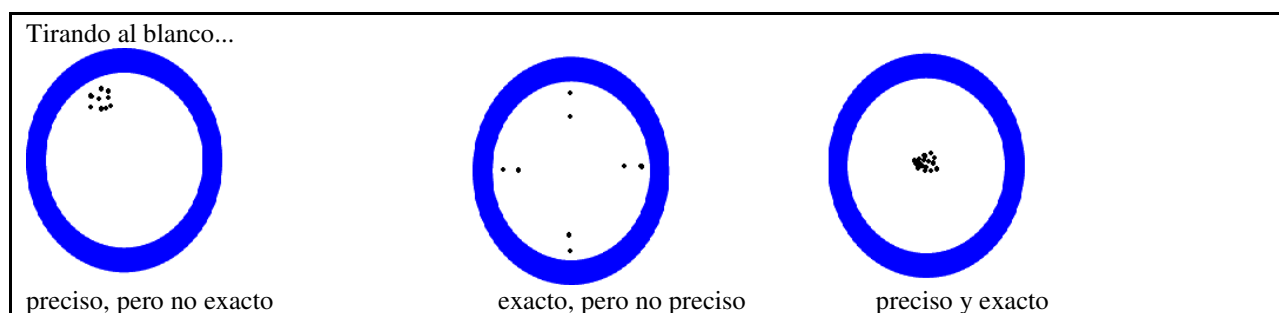
Este error, del que sólo podemos conocer su máximo (**E_{max}**), depende de la sensibilidad, precisión y exactitud de los instrumentos de medida.

La sensibilidad (se) es la unidad más pequeña que permite utilizar el instrumento de medida. En las reglas graduadas habituales $se = 1 \text{ mm}$. El **E_{max}** es igual a la mitad de la sensibilidad; **E_{max}** = $se/2$. Una regla milimetrada: tiene un **E_{max}** de $1/2 \text{ mm} = 0,5 \text{ mm}$.

Hay precisión cuando repetida la medida muchas veces da valores iguales o muy parecidos.

Hay exactitud si la media de repetidas medidas coincide con el valor exacto de la medida.

Así, si una longitud real de 9,0 cm. la medimos 4 veces y obtenemos 9,1; 9,0; 9,0 y 8,9 el instrumento es preciso y exacto. Si obtenemos 5,6 ; 5,5; 5,7; 5,6 será preciso, pero no exacto. Midiendo 9 ; 6 ; 12 ; 3 y 15 será exacto pero no preciso. La medida ideal es la que se obtiene con un máximo de sensibilidad, precisión y exactitud.



--¿Que sensibilidad se debe utilizar?

Una sensibilidad escasa proporciona datos de poca confianza, con mayor margen de error. Si es excesiva no es mala en sí, pero en general supone aparatos más caros y de manejo más difícil.

Hay que elegir la más adecuada para cada caso concreto, teniendo en cuenta la experiencia y el sentido común.

La sensibilidad es adecuada si la diferencia entre la medida más alta ,sin punto o coma decimal, y la medida más baja , también sin punto o coma decimal, está entre 30 y 300 .

Ejemplo:

1- medimos en mm. la longitud de las hojas de la planta XYZ. La medida mayor es 8 y la menor 4. Como $8-4=4$, que es menor de 30, la sensibilidad utilizada no es buena. En una medida de 5 mm. el error máximo es de 0,5 mm., o sea de un 10%. El instrumento de medida no es adecuado.
2- después utilizamos un aparato que mide en décimas de mm. Como valores extremos obtenemos 8,4 y 4,3 mm. $8,4-4,3=4,1$, que está entre 30 y 300. En una medida de 5,0 mm. el error máximo es de 0,05 mm., un 1%. Este instrumento sí es adecuado.

--Valor puntual y por intervalo de una medida

Al desconocer el valor exacto de una medida, **X_e** , hay que estimarlo. La medida se puede expresar de dos formas: puntual o por intervalo.

La medida puntual o valor puntual es el valor medido, X ; por tanto no es exacto..

El valor por intervalo o medida por intervalo es el intervalo en el que con seguridad (¡si se ha medido bien!) estará el valor exacto **X_e** de la medida. Se obtiene sumando y restando al valor puntual el error máximo, es decir, la mitad de la sensibilidad: $X \pm se/2$. De esta forma se obtienen los llamados *límites reales de la medida*, uno superior y otro inferior. Si medimos un lápiz con una regla milimetrada y obtenemos 151 mm., la medida puntual será 151 mm. Como la sensibilidad es de 1 mm., la medida por intervalo será $151 \pm 0,5 \text{ mm}$. o $\in (150,5 \div 151,5)$.

Si utilizamos una regla con nonius, que mide en décimas de mm. y obtenemos 151,1 mm. , el valor puntual será 151,1mm. Aquí la sensibilidad es de 0,1 y por tanto la medida por intervalo será $151,1 \pm 0,05$ ó $\in (151,05 \div 151,15)$.

Como es fácil equivocarse al realizar los cálculos, puede resultar útil el procedimiento siguiente:

- a) se toma el número, prescindiendo del posible punto o coma decimal y se añade un 0
- b) se le suma y resta 5
- c) si había decimales, se vuelve a poner la coma o punto decimal en su sitio. Así tenemos los dos límites del intervalo.

En el último ejemplo: $151,1 \rightarrow 15110 \rightarrow -5 = 15105$ y $+5 = 15115 \rightarrow 151,05$ y $151,15$

--Cifras significativas

Son las cifras del valor puntual de una medida, prescindiendo de los ceros a la izquierda de la primera cifra con valor distinto de cero. Son pues función de la sensibilidad.

medida	cifras significativas	medida	cifras significativas
65,5 m	3	4,53400 cm	6
0,0018 kg	2	1,00180 amp	6
1,0018 mm	5	0,10000 sec	5

En un número redondeado las cifras significativas llegan tan sólo hasta el lugar del redondeo. 18 millones como redondeo de 18 234 156 tiene 2 cifras significativas ; 3,14 como redondeo de 3,141592 tiene 3.

--Métodos de recuento (variables CL)

- a) observación, utilizando los órganos de los sentidos.
- b) gráficos: métodos de palotes, cuadrados...
- c) tarjetas de formas, contenidos o colores distintos
- d) lectura óptica, como en el escrutinio de quinielas y similares
- e) lectura magnética (de espacios marcados con lápiz de grafito)

--Síntesis de datos estadísticos

Una vez medida la variable en los diversos individuos se tiene una serie de datos, métricos o categóricos, los llamados DATOS ORIGINALES o DATOS AISLADOS, que sin más elaboración suelen ser poco útiles.

Es necesario ordenarlos y resumirlos para que proporcionen la máxima información de la forma más sencilla posible. Esto se hace de diversas formas:

- agrupando los datos según su frecuencia, con lo que se transforman en DATOS AGRUPADOS O DISTRIBUCION DE FRECUENCIAS, construyendo las correspondientes TABLAS y GRAFICOS ESTADISTICOS
- calculando los llamados INDICES o PARAMETROS ESTADISTICOS, como media aritmética, desviación estándar, porcentajes, etc.

Las Escuelas clásicas utilizan el término INDICE para las muestras y sus símbolos se representan con letras latinas, mientras que el término PARAMETRO se reserva para las poblaciones, con símbolos de letras griegas. Aquí utilizaremos ambos términos de forma indistinta, es decir, tanto para poblaciones como para muestras. Y salvo alguna rara excepción los símbolos serán de letras latinas.

Recordatorio : MEDIDA DE UNA VARIABLE CONTINUA

X_e	valor exacto, real, de la medida ; es desconocido
X	valor medido por el instrumento; es el valor puntual
E = X_e - X 	error de la medida ; por tanto $X_e = X \pm E$
E Máximo (E_{max})	se/2
Valor por intervalo de una medida	$X \pm E_{max}$ ó $\in (X - E_{max} \div X + E_{max})$ en ese intervalo está contenido el valor real X _e

Tema 4 : Tabulación de datos

La tabulación consiste en presentar los datos estadísticos en forma de tablas o cuadros.

--Partes de una tabla

- TITULO de la tabla, que debe ser preciso y conciso
- CONTENIDO, con
 - la *fila de encabezamiento o cabecera* (títulos de las columnas)
 - la *columna matriz*, con las modalidades o clases de la variable
 - *columnas de parámetros*
- NOTAS EXPLICATIVAS (opcional), como fuente de los datos, abreviaturas, etc.

--Forma de tabular

VARIABLES CUALITATIVAS

pueden representarse :

- la frecuencia absoluta (símbolo : **f** ó **n**), que es el n° de veces que aparece cada modalidad (resultado del recuento). La frecuencia total, de todas las modalidades juntas, se representa por **N**.
- la frecuencia relativa (**fr**) o proporción se obtiene dividiendo la frecuencia de cada modalidad entre el total de datos. $fr = f / N$. Los valores posibles oscilan entre 0 y 1. Suele expresarse con 3 decimales. La suma de todas las fr tiene que dar 1 ó un número muy cercano al 1, si ha habido redondeos.
- el porcentaje (**P** o **%**), que es la frecuencia relativa multiplicada por 100. $P = fr * 100$ ó $\% = (f*100)/N$. Suele expresarse con 3 dígitos. La suma de todos los porcentajes debe dar 100 o un número muy próximo, si ha habido redondeos.
- las frecuencia acumuladas (**Σf** ó **Σn**) que se obtienen sumando la frecuencia de cada modalidad a las frecuencias ya acumuladas anteriormente. En la primera modalidad no hay nada acumulado de antes y por tanto su frecuencia acumulada será su misma frecuencia. La última modalidad tiene que dar una frecuencia acumulada igual a N.
- las frecuencias relativas acumuladas y los porcentajes acumulados se obtienen de forma similar
- En las variables nominales las modalidades pueden ponerse en el orden que se quiera, pero en las ordinales hay que respetar el orden lógico.

Ejemplo:

Residencia Sanitaria S. S. de Castellón
Ingresos en Pediatría. Marzo 1980

Sección	f	fr	%	Σf	Σfr	$\Sigma \%$
Neonatología	25	0,125	12,5	25	0,125	12,5
Lactantes	95	0,475	47,5	120	0,6	60
Preescolares	80	0,400	40	200	1	100
Total	200	1	100			

En la tabla definitiva no se presentan todos estos parámetros, sino los más adecuados en cada caso concreto. Casi siempre **f** y/o **%** . Sólo el porcentaje, sin que conste N, no es correcto. En este ejemplo bastaría con **f** y **%** .

VARIABLES CUANTITATIVAS

Los datos se agrupan según la frecuencia de los valores. Es lo que se denomina *Distribución de frecuencias*. La forma de tabular depende del nº de datos.

----Si son pocos (la mayoría de autores pone el tope en 30) , se hace una tabla simple de forma similar a lo visto para las variables CL. Cada dato equivale a una modalidad. Al final nos quedaremos con la f de cada número y si se prefiere también con el %. Los números se ordenan de menor a mayor o de mayor a menor. La tabla puede hacerse en sentido vertical u horizontal.

Ejemplo: Si $x = (4, 1, 7, 2, 2, 9, 7, 2, 2, 9, 7, 1, 4)$

x	1	2	4	7	9
f	2	4	2	3	2

o bien

x	f
1	2
2	4
4	2
7	3
9	2

----Si son muchos se agrupan en clases, que son intervalos sucesivos de valores. Los datos se asignan a la clase que les corresponde y se cuentan los datos de cada clase, que está representada por el punto medio o centro de clase (pm ó c).

Esta agrupación es **arbitraria** con dos condiciones esenciales: que las clases sean mutuamente excluyentes y que todos los datos puedan ser asignados a una clase. Ahora bien, la experiencia ha ido introduciendo una serie de normas, que permiten hacer esta agrupación de la forma más racional posible.

Yo recomendaría los siguientes pasos:

- 1) calcular el **RECORRIDO (R)** , (a veces mal llamado Rango)
= (límite real superior del dato mayor – límite real inferior del dato menor)
O si se prefiere: = (valor tabulado máximo – valor tabulado mínimo) + 1
- 2) calcular el **Nº DE CLASES (NC)** .

Es función de N (tamaño de la muestra) y no hay reglas fijas.

En general: “entre 4 y 20” .

Ayudas: $NC = 1 + 3,32 \cdot \log N$ ó $1 + 1,44 \cdot \ln N$

O la siguiente tabla: N 8 16 32 64 128 256 etc.

NC 4 5 6 7 8 9 etc.

De entrada nos quedamos con 2 ó 3 opciones

- 3) calcular la **AMPLITUD** de las clases ó INTERVALO (i) : $i = R / NC$
Si **i** no es número entero, se redondea al número entero superior para que $NC \cdot i \geq R$ y así queden englobados todos los datos
Como probamos con 2 ó 3 opciones, conviene elegir una i que sea impar, pues así el punto medio de la clase (pm ó c) tendrá una cifra menos.

En principio todas las clases deben tener la misma amplitud.

- 4) Ver si hay **SOBRAS**, que son la diferencia entre $NC \cdot i$ y R. Se reparten lo mejor posible entre ambos extremos de la distribución fijando así los límites definitivos de la tabla.

5) Construir el esquema de la tabla, poniendo **columnas** de

- CLASES ó LÍMITES TABULADOS
- LÍMITES REALES
- PUNTO MEDIO (pm ó c)
- FRECUENCIA (f ó n)
- FRECUENCIA RELATIVA (fr)
- PORCENTAJE (P ó %)
- FRECUENCIAS ACUMULADAS (Σf ó Σn)
- FRECUENCIAS RELATIVAS ACUMULADAS (Σfr)
- PORCENTAJES ACUMULADOS ($\Sigma \%$)

6) Hacer el **RECuento** de datos y rellenar las casillas correspondientes

7) Escribir la **TABLA DEFINITIVA**. Son obligadas las clases y la frecuencia absoluta, pudiendo añadir otros parámetros, si se considera que mejoran la información. Una tabla excesivamente prolija resulta más difícil de leer. Por tanto la norma es: poner todo lo necesario, pero no más de lo necesario.

Es recomendable probar con al menos 2 tablas y elegir la que quede mejor.

Algunos de éstos parámetros son los mismos que se han visto para las variables CL. Otros precisan una aclaración:

Los **límites de las clases** son los valores inferior y superior de cada clase. (Límite inferior y límite superior). Hay que distinguir entre los **límites tabulados (LT)** y los **límites reales (LR)**. Los límites tabulados son los datos originales que abren y cierran una clase. Los límites reales son el límite real inferior del primer valor (LRI) y el límite real superior del último (LRS).

El **punto medio o centro** de la clase (pm ó c) representa a la clase cuando se hacen operaciones matemáticas. Es la media de los límites. Da lo mismo tomar los límites reales que los tabulados, ya que ambos dan el mismo resultado.

En una distribución con todas las clases de la misma amplitud las diferencias entre los puntos medios, los límites inferiores y los límites superiores de dos clases consecutivas valen lo mismo y son igual a la amplitud de la clase (i). Esto facilita la construcción de la tabla.

Una **clase** es **abierta** cuando carece de un límite. Sólo pueden ser abiertas la primera clase (p.e. <10 ; no tiene límite inferior) y la última (p.e. >100 ; no tiene límite superior). No deben usarse, a no ser que no haya otro remedio.

EJEMPLO:

Tabular los 70 valores siguientes:

DATOS ORIGINALES (N = 70)

40 55 19 51 62 15 20 44 60 60 45 15 21 31 13 44 41 43 51 35 50 33 25 16 61
14 14 59 59 59 20 23 25 29 29 59 58 54 50 49 39 27 37 23 24 58 27 28 57 32
32 34 57 56 35 35 54 36 43 46 52 50 49 42 43 46 40 39 31 48

PASOS DE LA TABULACION

-dato mayor: 62, cuyo LRS es 62,5

-dato menor: 13, cuyo LRI es 12,5

-recorrido (R): $62,5 - 12,5 = 50$ ó $(62 - 13) + 1 = 50$

-nº de clases (NC): 7 u 8

-amplitud (i):

-si NC = 7 , $i = 50/7 = 7,1 \rightarrow 8$ (par)

-si NC = 8 , $i = 50/8 = 6,2 \rightarrow 7$ (impar)

-nos quedamos pues con NC = 8 de amplitud 7, que es impar

-sobras: $(8 \cdot 7) - 50 = 6$, que repartimos así: 3 abajo y 3 arriba

la 1ª clase empezará en 10 (13-3)

la última terminará con el 65 (62+3)

--ya se puede construir el esquema de la tabla (clases, LR y punto medio) y proceder al recuento de los datos que corresponden a cada clase, para completar las otras columnas

Clases (Límites tabulados)	Límites reales	punto medio c	f	fr	%	Σf	Σfr	Σ%
10 – 16	9,5 – 16,5	13	6	0,09	8,57	6	0,09	8,57
17 – 23	16,5 – 23,5	20	6	0,09	8,57	12	0,17	17,1
24 – 30	23,5 – 30,5	27	8	0,11	11,4	20	0,29	28,6
31 – 37	30,5 – 37,5	34	11	0,16	15,7	31	0,44	44,3
38 – 44	37,5 – 44,5	41	11	0,16	15,7	42	0,60	60,0
45 – 51	44,5 – 51,5	48	11	0,16	15,7	53	0,76	75,7
52 – 58	51,5 – 58,5	55	9	0,13	12,9	62	0,89	88,6
59 - 65	58,5 – 65,5	62	8	0,11	11,4	70	1,00	100
Suma			70	1,01	99,94			

***Esta no es la única tabla posible, aunque probablemente sea la mejor.

Podríamos hacerla con 7 clases de amplitud 8; sobras: 6 . Clases: 10 – 17 ; 18 – 25; ...; 58 - 65

O bien 6 clases de amplitud 9. Sobras: 4 . Clases: 11- 19 ; 20 – 28; ...; 56 - 64

o bien 10 clases de amplitud 5 . Sin sobras. Clases: 13 –22 ; 23 – 32 ; ; 53 - 62

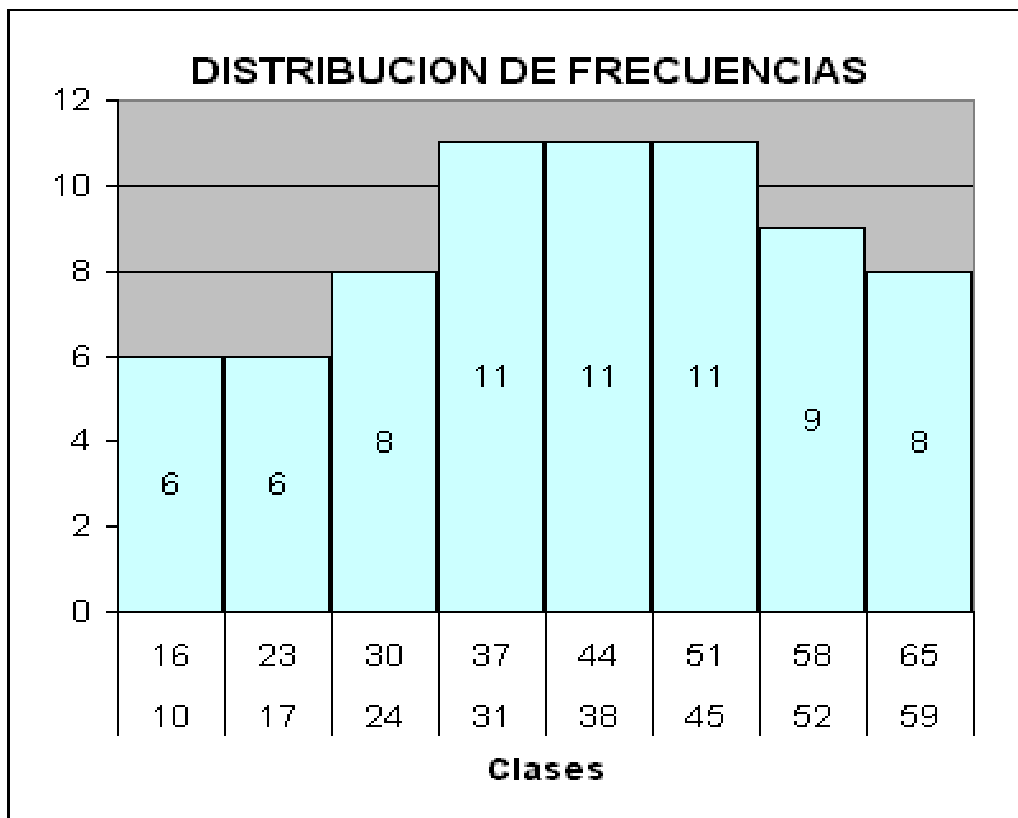
***En la tabla definitiva no suelen ponerse los LR. Las clases y la frecuencia están prácticamente siempre. Según la naturaleza de la variable puede ser conveniente añadir algún otro parámetro, que contribuya a una información mejor y más clara.

*** En la página siguiente puede verse la tabla y el gráfico que elabora automáticamente mi programa de Excel, Exceltabla.xls, a partir de los 70 datos del ejemplo anterior, introducidos en la columna A.

Tabla e histograma del ejemplo de la página 4-4 que hace “Exceltabla”

Lim.Tab.Inf.	LimTab.Sup.	pm	f	%	Σf	$\Sigma \%$
10	16	13	6	8,6	6	8,6
17	23	20	6	8,6	12	17,1
24	30	27	8	11,4	20	28,6
31	37	34	11	15,7	31	44,3
38	44	41	11	15,7	42	60,0
45	51	48	11	15,7	53	75,7
52	58	55	9	12,9	62	88,6
59	65	62	8	11,4	70	100,0
			0			

<i>Datos origin.:</i>	SESGO	-0,196	MODA	59,00
	CURTOSIS	-1,105	p3	14,07
	MEDIA GEO	36,53	p10	19,90
	MEDIANA	40,50	p25	28,25
	MEDIA	39,59	p75	51,00
	DS	14,41	p90	59,00
	VARIANZA	207,58	p97	60,00



Tema 5 : Representaciones gráficas

Los datos estadísticos pueden ser también representados por medio de **gráficos**. Un viejo proverbio chino dice que una imagen vale más que mil palabras (o que mil números, aplicado a la Estadística). Los gráficos son una simplificación y un complemento de una tabla estadística. Son más sencillos, más llamativos y a menudo más inteligibles, aunque se pierde información.

Componentes

Como en las tablas estadísticas se pueden distinguir:

- el título
- el gráfico en sí (casi siempre complementado con números)
- notas explicativas , si procede

Tipos de gráficos

- Diagramas
 - de barras
 - histogramas
 - polígonos de frecuencias
- Gráficos sectoriales
- Pictogramas
- Otros

Los **DIAGRAMAS** utilizan un sistema de coordenadas cartesianas. En el eje de abscisas (x) se representa la variable. En el de ordenadas (y) las frecuencias o porcentajes.

Si la variable es CL se marcan en el eje de abscisas las modalidades y sobre ellas se dibujan líneas o barras de altura proporcional al parámetro representado. Si la variable es CT se marcan los valores y clases correspondientes al recorrido de la variable.

La escala de y debe de empezar siempre en 0 para evitar manipulaciones y engaños ópticos.

Habitualmente se trata de una escala aritmética, pero cuando hay frecuencias o valores muy dispares el gráfico es apenas legible y es mejor utilizar escalas logarítmicas o semilogarítmicas. Una alternativa, algo chapucera, es quebrar claramente la escala y las barras. Todo antes que violar la norma del comienzo de y en 0.

En un buen diagrama la longitud de x debe de estar entre 1 y 2 veces la de y . Ambas escalas deben de estar claramente rotuladas, directamente o por medio de una nota explicativa. Son preferibles números cortos (redondeados) y hay que evitar dar excesivos datos, sobre todo en presentaciones, ya que el gráfico se muestra un corto espacio de tiempo. Otra cosa es un gráfico impreso al que el lector puede dedicarle el tiempo que quiera. Los ordenadores permiten fácilmente dibujar los gráficos en 3D. Las barras pasan a ser prismas o incluso cilindros o conos, a gusto del usuario.

-El diagrama de barras o columnas es propio de variables discretas (todas las CL y las CT discretas). Cada barra corresponde a una modalidad o valor de la variable.. La altura de la barra es proporcional a la frecuencia a representar. Todas las barras deben de tener la misma anchura y la distancia entre ellas debe de ser como máximo la anchura de las barras.

Se pueden distinguir tres tipos de diagramas de barras:

- a) simples (figuras 1 y 2)
- b) de barras adosadas o parcialmente superpuestas, cuando se presentan de forma paralela dos conceptos que interesa comparar, p.e. hombres y mujeres (figuras 3 y 4)
- c) de barras mixtas, apiladas, una variante del anterior (figura 5).

-El histograma es propio de variables CT continuas agrupadas en clases. Las barras están unas al lado de otras sin separación, a no ser que alguna clase tenga una frecuencia de 0. Cada barra

empieza en el límite real inferior de la clase que representa y termina en el límite superior, que a su vez es el comienzo de la clase siguiente. El punto medio de la clase coincide con el centro de la base. La superficie de cada barra es proporcional a la frecuencia de la clase. Si todas las clases tienen la misma amplitud, como en principio debe ser, la altura es la frecuencia de la clase. Si hay clases con distinta amplitud no puede ponerse la etiqueta de frecuencia (f ó n) en el eje vertical, ya que sería engañoso. Debe figurar la de “densidad de frecuencias” (fd). $fd = \frac{f}{i}$ (fig. 6)

Se pueden distinguir tres tipos de histogramas:

- 1) el H. simple, que es el que acabamos de ver (fig. 7)
- 2) el H. de frecuencias acumuladas, en el que cada barra representa las frecuencias acumuladas en cada clase. El gráfico tiene forma de escalera más o menos irregular. (fig 8)
- 3) el H. doble, cuyo paradigma es la **pirámide de población**. Este gráfico nos informa de la distribución por edades de un grupo poblacional, separando hombres y mujeres y rotando el gráfico de tal forma que las edades de las personas, agrupadas en clases, están en el eje vertical y la frecuencia de cada clase en el eje horizontal. (fig. 9).

Un **POLIGONO DE FRECUENCIAS** se obtiene uniendo los puntos medios de los techos de un hipotético histograma, que se corresponden, al ser la barra un rectángulo, con los puntos medios o centros de cada clase. La línea debe comenzar y terminar en el eje de abscisas, precisamente en el sitio que correspondería al punto medio de dos clases inexistentes, la que precedería a la primera y la que seguiría a la última. Si se superponen un histograma y el correspondiente polígono de frecuencias se ve que la superficie del histograma y el área que incluye el polígono es la misma. Por tanto ambos representan igualmente a la distribución. Los hay también simples y de frecuencias acumuladas. (fig. 10 y 11)

Cuando no se representa toda la distribución sino tan sólo una parte de la misma, no hay que bajar la línea hasta el eje de abscisas. Por delante y detrás de lo representado hay clases cuya frecuencia no es ofrecida al lector. Este gráfico se llama **diagrama lineal**.

Los GRAFICOS SECTORIALES o de TARTA equivalen a un diagrama de barras y por tanto sirven para representar variables discretas. Se utilizan círculos o semicírculos y a cada modalidad o valor se le adjudica un sector circular, cuya superficie sea proporcional a la frecuencia relativa o porcentaje. Para ello se calcula el ángulo que le corresponde mediante una simple regla de tres. A todo el círculo le corresponden 360° y si es un semicírculo 180° .

En el ejemplo de los ingresos en Pediatría:

al 100% (todos)	le corresponden 360°	
al 12,5% (Neonatos)	“ “	x° $x^\circ = 45^\circ$

y así para las otras Secciones se obtiene: Lactantes 171° y Preescolares 144°

Luego mediante un transportador se trazan en el círculo las líneas correspondientes.

Los sectores circulares se pueden desgajar del conjunto para que resalten más. (fig. 12 y 13)

Los **PICTOGRAMAS** utilizan figuras e imágenes de todo tipo, como pilas de monedas, balanzas, coches, muñequitos, mapas distorsionados, etc. Siempre deben respetar el espíritu del gráfico básico. (fig. 14)

La fantasía y la inspiración pueden sugerir **OTROS** tipos de gráficos. Pero lo esencial no es que sean bonitos, sino que informen bien. Pero si son buenos, bonitos y sencillos, mejor que mejor.

Los gráficos se prestan mucho a la manipulación (no respetando las normas básicas que se han citado) y pueden ofrecer por tanto una información falsa (fig. 15 y 16). En este caso se podría decir que una imagen puede mentir más que mil palabras.

Residencia Sanitaria de la S.S. de Castellón
Ingresos en Pediatría. Marzo 1980

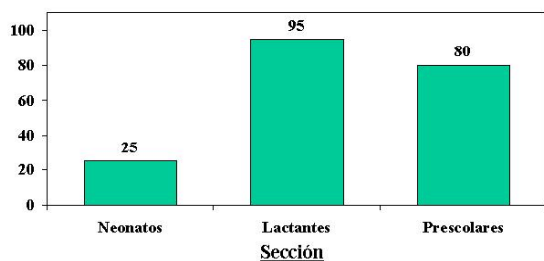


Figura 1
Diagrama de barras simple

Residencia Sanitaria de la S.S. de Castellón
Ingresos en Pediatría. Marzo 1980

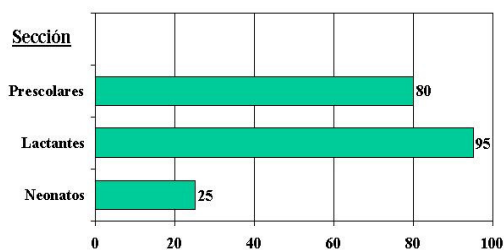


Figura 2
Diagrama de barras simple, rotado

Residencia Sanitaria de la S.S. Castellón
Ingresos en Pediatría. Marzo 1980

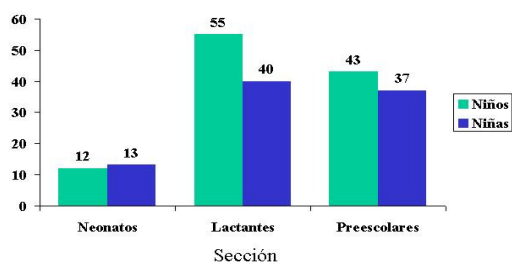


Figura 3
Diagrama de barras adosadas

Residencia Sanitaria de la S.S. Castellón
Ingresos en Pediatría. Marzo 1980

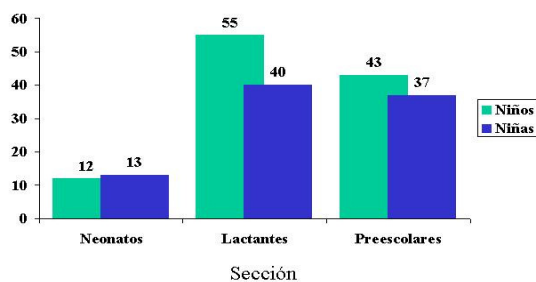


Figura 4
Diagrama de barras parcialmente superpuestas

Residencia Sanitaria de la S.S. Castellón
Ingresos en Pediatría. Marzo 1980

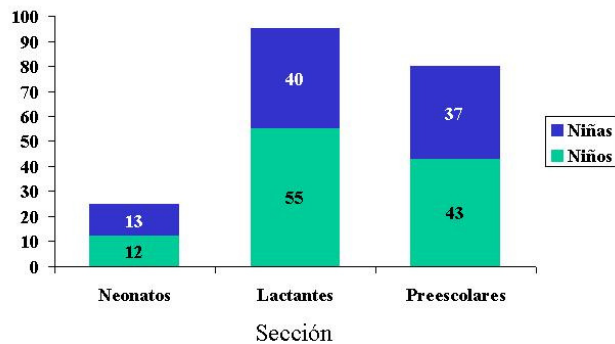


Figura 5
Diagrama de barras mixtas
o apiladas

	clases	f	i	fd=f/i
A	0-3	12	4	3
B	4-8	20	5	4
C	9-11	15	3	5

La amplitud de las clases de esta distribución varía.
La superficie de las columnas representa correctamente
a las clases; su altura depende no de la f sino de la df

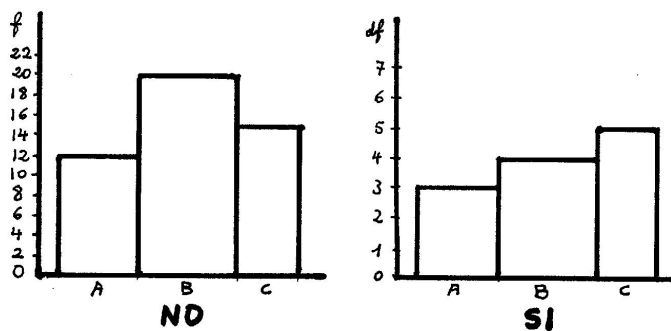


Figura 6
Si no son iguales todas las cla-
ses, hay una regla especial

"70 DATOS"
HISTOGRAMA

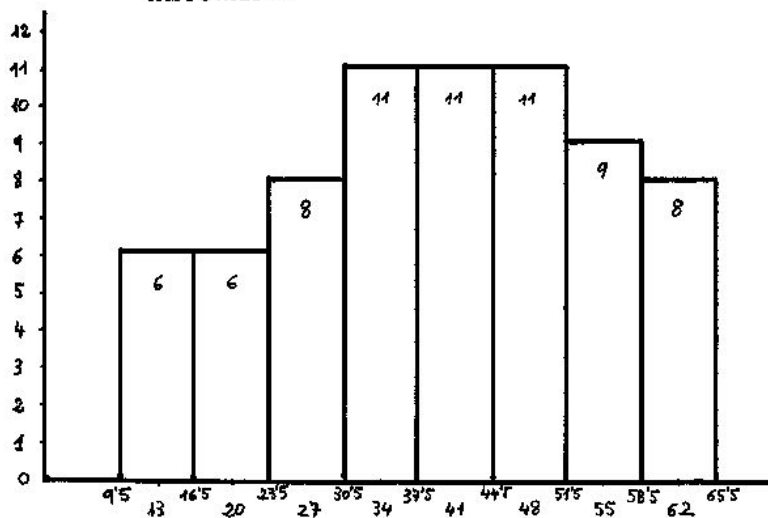


Figura 7
Histograma simple
"70 Datos" de la tabla
del tema anterior

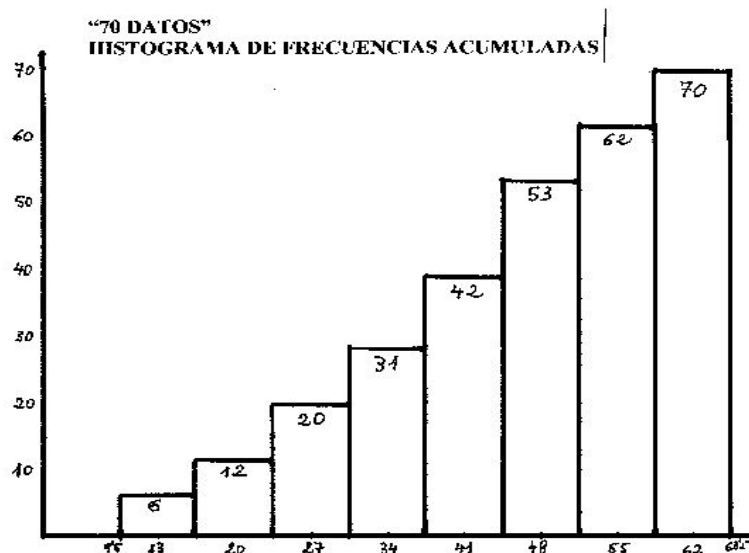


Figura 8
Histograma de frecuencias
acumuladas

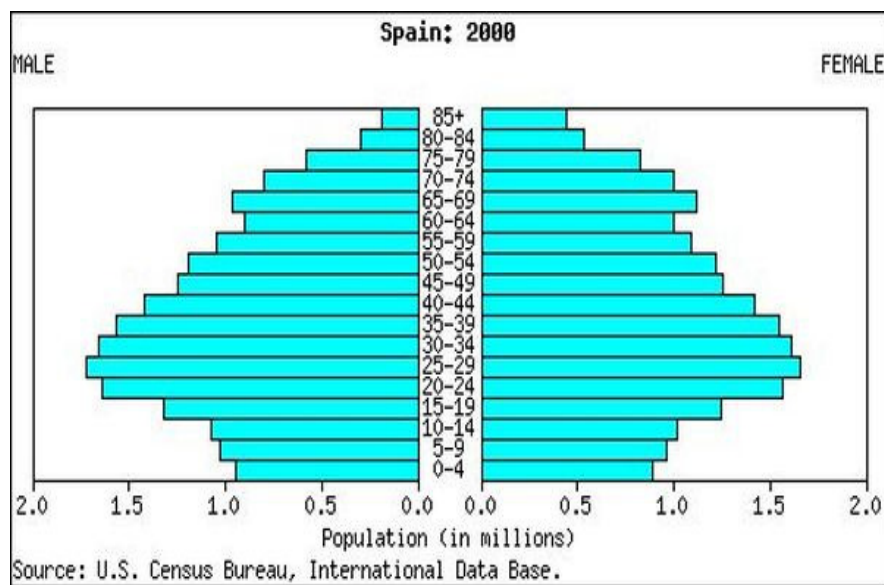


Figura 9
Pirámide de población
de España en 2002.

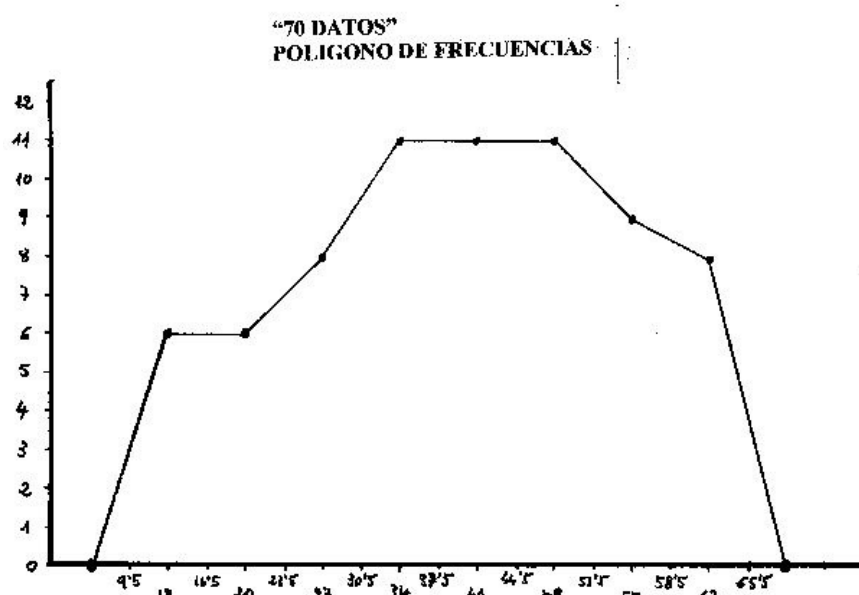


Figura 10
POLIGONO DE
FRECUENCIAS

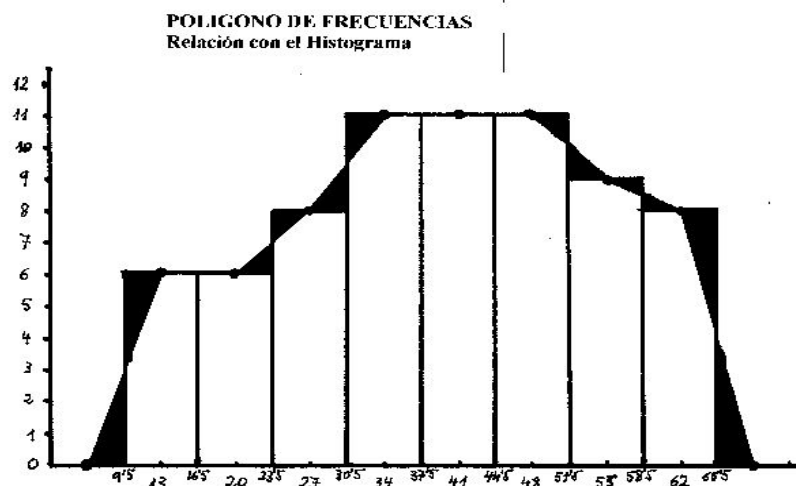


Figura 11
Relación entre el
histograma y el polígono

Residencia Sanitaria de la S.S. Castellón
Ingresos Pediatría. Marzo 1980

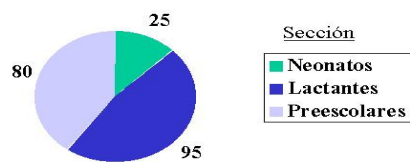


Figura 12
Diagrama circular
o de tarta

Residencia Sanitaria de la S.S. Castellón
Ingresos Pediatría. Marzo 1980

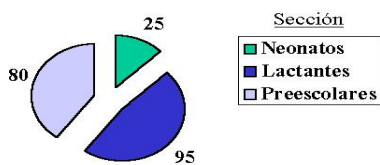
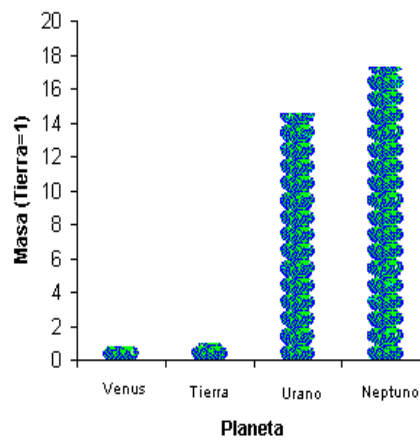
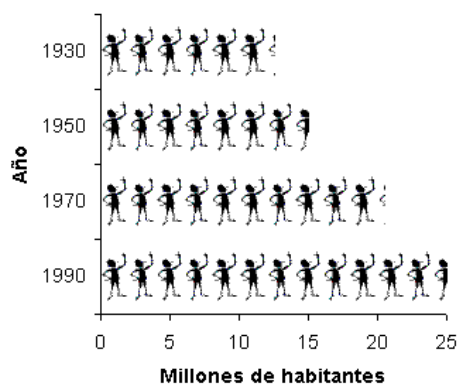
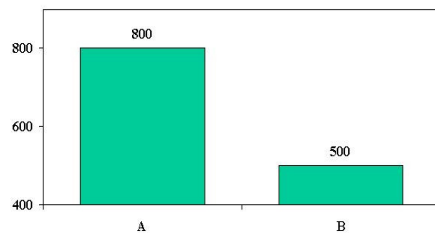


Figura 13
Diagrama circular,
cortado

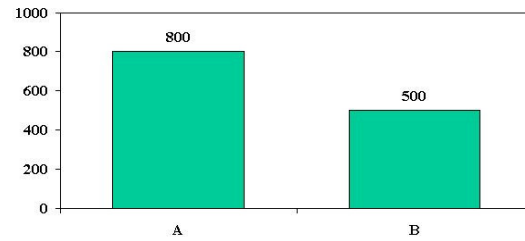


Figuras 14 v 15 Pictogramas

Estudio comparativo medicamentos A y B
Curaciones en 1000 pacientes



Estudio comparativo medicamentos A y B
Curaciones en 1000 pacientes



Figuras 16 y 17

El no empezar la escala en 0 , agranda las diferencias

El gráfico de la izquierda es incorrecto

Tema 6 . Índices estadísticos de variables cuantitativas. Parámetros de tendencia central, dispersión, posición y forma.

Los parámetros o índices (ya vimos en el tema 3 que consideramos ambos conceptos como equivalentes) son otra forma de presentar resumidos los datos estadísticos.

Hay que distinguir:

- parámetros de tendencia central, que informan del centro de la distribución
- parámetros de dispersión, que informan de la dispersión de los datos
- parámetros de posición, que sitúan a los datos en el conjunto de la distribución ordenada. Los más utilizados en Bioestadística son los percentiles. Algunos de ellos pueden ser considerados también como parámetros de tendencia central y otros como de dispersión.
- parámetros de forma, que precisan la forma de la distribución. Podría decirse que expresan numéricamente la forma del histograma.

Parámetros de tendencia central

Los más importantes son:

- la media aritmética, o simplemente la media
- la mediana
- la moda
- los percentiles “centrales” (p 25 a p75)

En la explicación de los parámetros se utilizarán tres grupos de datos en los ejemplos:

Supuesto A): 8 , 1 , 4 , 8 , 8 , 5 , 1

Supuesto B): los “70 DATOS” originales del tema 4

Supuesto C): la tabla que agrupa a esos 70 datos

--La **MEDIA** es la suma de todos los valores dividida por el número de ellos.

Símbolo: \bar{x}

Cálculo:

1) *datos aislados, originales:*

$$\bar{x} = \frac{\sum x}{N} \quad ; \text{ para el ejemplo A: } \bar{x} = \frac{8+1+4+8+8+5+1}{7} = 5$$

$$\text{para el ejemplo B: } \bar{x} = 39,6$$

2) *datos agrupados en clases:*

$$\bar{x} = \frac{\sum fc}{N} \quad ; \text{ en el ejemplo C:}$$

$$\bar{x} = \frac{(6*13)+(6*20)+(8*27)+(11*34)+(11*41)+(11*48)+(9*55)+8*62}{70} = 39,4$$

Propiedades de la media

- 1- si a cada valor de x le sumamos, restamos, multiplicamos o dividimos por una constante, la media queda sumada, restada, multiplicada o dividida por esa constante
- 2- la media es sensible a la variación de cada valor de x
- 3- la media se expresa en la misma unidad de medida que los datos originales
- 4- si la media tiene decimales es habitual expresarla con uno más que los datos originales

Media aritmética ponderada

Se usa cuando se quiere o se debe dar una fuerza distinta a determinados valores.

$$\bar{x}_{\text{pond}} = \frac{\sum xF}{\sum F}, \text{ siendo } x \text{ el valor original y } F \text{ el factor de ponderación}$$

Ejemplos:

- 1) Al introducirse los estudios de Diplomatura en esta Escuela, el Área de Ciencias de la Enfermería englobaba diversas asignaturas, de cuyas notas salía la nota del Área. Como eran de extensión e importancia dispares, se decidió que Microbiología (que para abreviar llamaremos A) participaría con el 33%, la Bioestadística (B) con el 28%, las Prácticas (C) con un 23% y el resto, la media de Salud Pública, Organización e Historia de la Profesión ((D1+D2+D3)/3) conjuntamente con un 16%.

Si las notas de las asignaturas fueron : 6 en A, 5 en B, 8 en C, 6 en D1, 8 en D2 y 10 en D3 , la nota del Área fué 6,5 y no la media aritmética 7,2

$$\bar{x}_{\text{pond}} = (6*33 + 5*28 + 8*23 + 8*16)/(33+28+23+16) = 6,5$$

- 2) la media de una distribución calculada a partir de una tabla es realmente una media ponderada en la que x es el punto medio de clase y f (frecuencia) el factor de ponderación F.

Otras medias

En circunstancias especiales (distribución con sesgo muy intenso) hay autores que prefieren otras medias como la media geométrica o la trimedia , en las que no vamos a entrar.

En los concursos varios jueces dan una nota al actuante. Para disminuir favoritismos e inquinas se utiliza la media recortada, que se obtiene prescindiendo del valor más alto y del más bajo. Este sistema se puede aplicar también para evitar errores, cuando se manejan grandes cantidades de datos y aparecen valores marginales “anómalos”. Así se puede decidir no tener en cuenta un pequeño porcentaje (no más allá de un 3%) de los valores más altos y más bajos.

--La **MEDIANA** es el valor que ocupa el centro de la distribución una vez ordenados los datos. El símbolo es M

Cálculo:

1 – *datos aislados, originales* (¡que deben estar ordenados!)

a) N es impar: es el valor que ocupa el lugar (N+1)/2

b) N es par: es la media de los valores que ocupan los lugares N/2 y siguiente.

2 – *datos agrupados*

--de forma simplificada se toma como M el punto medio de la clase que contenga la mediana (el lugar se calcula como en los datos aislados) y se identifica la clase por la columna de frecuencias acumuladas.

--de forma un poco más exacta se utiliza la fórmula

$$M = L_i + i \left(\frac{N/2 - \sum f_M}{f_M} \right)$$

siendo L_i el límite inferior de la clase mediana, i su amplitud, N el nº total de datos, $\sum f_M$ las frecuencias acumuladas por debajo de la clase mediana y f_M la frecuencia de la clase mediana.

Ejemplos:

--supuesto A: se ordenan los 7 datos: 1 , 1 , 4 , 5 , 8 , 8 , 8 ; como N es impar la mediana será el valor que ocupe el lugar (7+1)/2 = 4 ; el 4º lugar es el 5

--supuesto B: se ordenan los 70 datos, número par. La mediana es la media de los valores que ocupen el lugar 70/2 = 35 y el siguiente, 36 . El 35º vale 40 y el 36º 41 , por tanto M = 40,5

--supuesto C: ***la clase mediana es la que contiene los valores 35° y 36°. En la columna de Σf se ve que pertenecen a la clase 38-44, que es la clase mediana. Por tanto $M = c = 41$

***aplicando la fórmula:
$$M = 37,5 + 7 \left[\frac{\frac{70 - 31}{2}}{11} \right] = 40$$

Propiedades de la mediana

Son las mismas que las de la media excepto la 2ª: la mediana sólo es sensible a la variación de los datos originales si se altera el orden en el centro de la distribución.

--La **MODA** es el valor más frecuente. Puede ocurrir que no haya moda o que haya más de una (empates en el máximo). El símbolo es M_o .

Cálculo:

-en *datos originales* se hace el recuento y se busca el valor más frecuente. Si hay empate, la moda es múltiple.

-en *datos agrupados* en tabla: la M_o será el punto medio de la clase modal, es decir, la más frecuente. En caso de empate se dan los puntos medios de las clases correspondientes.

Propiedades: como la mediana.

Ejemplos:

supuesto A: $M_o = 8$; supuesto B: $M_o = 59$; supuesto C: hay tres clases con frecuencia de 11; $M_o = 34, 41$ y 48

De estos tres parámetros de tendencia central el mejor es sin duda alguna la media, pero hay algunos casos concretos (clases abiertas, valores muy discordantes) en que la mediana o incluso la moda son mejores. Cuando $N \geq 30$ la media suele ser un buen parámetro. En todo caso si el CV (coeficiente de variación), que luego veremos, supera el 50% la media no es buen representante del centro de la distribución.

Parámetros de dispersión

Informan de la dispersión de los datos, de la amplitud del conjunto. Los más importantes son:

- El RECORRIDO, que ya vimos en el tema 4 , o simplemente citar el máximo y el mínimo.
- La VARIANZA, que se basa en las diferencias entre cada valor y la media de la distribución.
- La DESVIACION ESTANDAR, que es la raíz cuadrada de la varianza.
- El COEFICIENTE DE VARIACIÓN, que relaciona la desviación estándar y la media.

--Varianza

Símbolo : s^2 (σ^2 , en la nomenclatura con caracteres griegos)

Cálculo: hay fórmulas distintas según los datos pertenezcan a una población o a una muestra.

-- población

- **datos aislados:**
$$s^2 = \frac{N \sum x^2 - (\sum x)^2}{N^2}$$

- **datos agrupados:**
$$s^2 = \frac{N \sum (fc^2) - (\sum fc)^2}{N^2}$$

-- muestra

- **datos aislados:**
$$s^2 = \frac{N \sum x^2 - (\sum x)^2}{N(N-1)}$$

- **datos agrupados:**
$$s^2 = \frac{N \sum (fc^2) - (\sum fc)^2}{N(N-1)}$$

Propiedades de la varianza

- 1- si a cada valor de x le sumamos o restamos una constante k, la varianza queda igual
- 2- si cada valor de x lo multiplicamos o dividimos por una constante k, la varianza queda multiplicada o dividida por k^2
- 3- la varianza es sensible a la variación de cada valor de x
- 4- la varianza se expresa en el cuadrado de la unidad de medida utilizada en la variable.
- 5- si la varianza tiene decimales, es habitual expresarla con dos decimales más que los datos originales

Ejemplos:

Con datos originales es conveniente construirse una tabla auxiliar con dos columnas: x y x^2 .

--Así en el supuesto A (asumiendo que es una muestra):

x	x^2
8	64
1	1
4	16
8	64
8	64
5	25
1	1
-----	-----
35	235

$$s^2 = \frac{7 * 235 - 35^2}{7 * 6} = 10$$

--en el supuesto B : $s^2 = 207,58$

--en el supuesto C: la tabla auxiliar tendrá las columnas f , c , $f*c$, c^2 , fc^2 para que podamos tener los sumatorios necesarios para aplicar la fórmula.

$$s^2 = 218,96$$

--La **DESVIACION ESTANDAR** es la raíz cuadrada de la varianza y por tanto es un número más manejable y de utilización más frecuente.

Símbolo: s .También se usa mucho D.E. y la abreviatura inglesa S.D. Y la letra griega σ .

Fórmula: $s = \sqrt{s^2}$

Propiedades: como la media

Ejemplos:

-supuesto A: $s = 3,2$

-supuesto B : $s = 14,4$

-supuesto C: $s = 14,8$

--El **COEFICIENTE DE VARIACION** es un índice abstracto, que no tiene unidad de medida. Da igual que midamos la variable en cm , kg, sec., etc. , el coeficiente de variación se expresa siempre como %. (que puede ser mayor del 100%).

Símbolo: CV

Fórmula: $CV = \frac{100s}{\bar{X}}$

Aplicaciones:

- 1) comparar dispersiones de variables, incluso si están medidas en unidades distintas. La variable con el CV menor tiene la menor dispersión (y viceversa).
- 2) valorar la representatividad de una media. Es buena si no supera el 50%.

Ejemplos:

-supuesto A: 64%

-supuesto B: 36,4%

-supuesto C: 37,6%

-otro ejemplo: Los niños de 3 años de la ciudad C tienen una talla media de 93 cm con $s = 3,8$. Los niños de 15 años de esa ciudad miden en media 162 cm con $s = 6$. ¿A qué edad es la talla más variable?

Se calcula el CV: -a los 3 años: 4,09% -a los 15 años: 3,70%

Respuesta: La talla es más variable a los 3 años.

PARAMETROS DE FORMA

1) **SESGO** : es el grado de asimetría de una distribución, expresado por el coeficiente de sesgo o asimetría, cuyo valor ideal es 0 (entonces hay simetría). Cuando hay un Sesgo la parte más alta del histograma (o de la campana de Gauss) se desplaza hacia la derecha o la izquierda y la campana tiene una cola larga, donde estará la media, y otra más corta, en la que suelen estar la mediana y la moda. Si la media es menor que la M y/o la Mo, el sesgo es negativo y si es mayor, el sesgo es positivo.

Símbolo: Sg

Hay una fórmula, muy compleja, para calcular el coeficiente de sesgo, en la que no entramos.

Un cálculo aproximado es: $Sg = \frac{3(\bar{x} - M)}{s}$, aunque lo mejor es observar la campana o el

histograma. **Mirando la campana, si se desplaza a la derecha el sesgo es negativo; si lo hace a la izquierda, positivo.** Si nos ponemos en lugar de la campana, al revés.



Mirando el histograma de los “70 DATOS” (página 5.4) se ve que tiene un pequeño sesgo hacia la derecha, es decir, negativo. Con los datos originales el cálculo exacto da un sesgo de -0,196; la fórmula aproximada da -0,187. Con los parámetros calculados a partir de la tabla el sesgo vale según la fórmula aproximada -0,324.

2) **CURTOSIS**

es el grado de apuntamiento de una distribución, expresado por el coeficiente de curtosis, cuyo cálculo es complejo y no se ve aquí.

Símbolo: ct o k

Se toma como referencia a la campana de Gauss de la distribución normal, cuya k vale 0 y se dice que es mesocúrtica. Si la distribución es más alta y delgada, se dice que es leptocúrtica. y k es >0. Si es achatada y ancha se denomina platicúrtica y k es <0.

Los “70 DATOS” tienen una k = -1,105 y por tanto la distribución es algo platicúrtica.

PARAMETROS DE POSICION

1) **PERCENTILES**

Los percentiles (p) son parámetros de posición que nos indican la situación de cada valor en el conjunto de los datos ordenados, que se han dividido en 100 partes iguales. Se presentan como tabla o como gráfico.

Se expresan como pa siendo a el % de datos que queda por debajo del valor original al que corresponde ese percentil. Dicho de otra forma: a un valor le corresponde el percentil pa , cuando ordenados los datos el a% es menor que él y el (100-a)% es mayor.

Cálculo:

- 1- *en datos originales* : se ordenan los datos de menor a mayor y se calcula el lugar en el que estará el percentil (pa) buscado mediante la fórmula : lugar del pa = $N \cdot a / 100$. El valor que corresponda a es lugar o n° de orden será el pa
- 2- *en datos agrupados*: se utilizan la tabla o el gráfico de los porcentajes acumulados, interpolando, si es preciso. Hay una fórmula, parecida a la de la mediana, pero no suele ser necesaria.

Los percentiles se utilizan mucho en Pediatría en tablas y gráficos de crecimiento, pero en los últimos años su uso se ha extendido a muchos datos biológicos: colesterol, tensión arterial, densidad ósea... Han desplazado casi totalmente a otros parámetros de posición similares, como los deciles (el conjunto se divide en 10 partes iguales) y los cuartiles (el conjunto se divide en 4 partes).

Realmente hay 100 percentiles, que van del p1 al p100, pero en la práctica se utilizan para mayor claridad sólo algunos de ellos. En Europa en las tablas y gráficos de crecimiento se utilizan el p3 , p10 , p25, p50, p75, p90, y p97.

El p50 se corresponde con el centro de la distribución: el 50% de los valores es mayor y el 50% es menor. Por tanto coincide con la mediana: $p50 = M$

En las variables biológicas los valores normales se obtienen a partir de muchas determinaciones en individuos sanos. Si un valor está por debajo del p3 se considera anormalmente bajo; si está por encima del p97, anormalmente alto; entre el p10 y el p90, totalmente normal. Entre el p3 y el p10, así como entre el p90 y el p97, aunque son aún normales, se consideran como en “zona de riesgo” o “sospecha”, dada la proximidad de la zona anormal.

Los percentiles entre p25 y p75 pueden ser considerados también como parámetros de tendencia central y los mayores y menores como de dispersión.

Con los percentiles no pueden hacerse operaciones matemáticas, ya que son parámetros de posición . Así, pues, $p50 \neq (p25 + p75)/2$

Al final de este tema puede verse un ejemplo de gráficos percentilados del peso y talla de niños de 2 a 18 años. Un niño de 5 ½ años que pesa 23 kg y mide 106 cm tiene una talla en el p10, un peso <p90 y una relación peso/talla >p97.

2) La **PUNTUACION TIPIFICADA O NOTA TIPIFICADA** puede ser también considerada como un parámetro de posición. Se verá con detalle en el tema 10. Adelanto:

Símbolos: se utilizan varios según las escuelas: c, z, SDS , SDE...

Fórmula:
$$c = \frac{X - \bar{X}}{s}$$

Equivalencias aproximadas entre percentiles y puntuaciones tipificadas:

p	3	10	25	50	75	90	97
c	-2	-1,3	-0,7	0	0,7	1,3	2

Dos observaciones finales

1) una distribución queda perfectamente definida conociendo todos los parámetros que hemos visto. Como el sesgo y la curtosis son de cálculo más difícil, el mínimo son la media y la desviación estándar, que suelen anotarse así : $\bar{x} \pm s$ ó $\bar{x} \pm DE$.

Que la media sola no es suficiente lo aclara el clásico ejemplo del pollo:” si una persona se come dos pollos y otra no come ninguno, la Estadística dirá que se comen un pollo cada uno”. La media es ciertamente 1 . Pero si calculamos la desviación estándar la valoración puede ser distinta:

-uno come 2 pollos y el otro ninguno:

x	x ²
2	4
0	0
---	---
2	4

$$s = \sqrt{\frac{(2*4) - 2^2}{2*1}} = 1,4 \text{ y el CV} = 140\%$$

¡la media no es buena representante!

-cada uno come un pollo:

x	x ²
1	1
1	1
---	---
2	2

$$s = \sqrt{\frac{(2*2) - 2^2}{2*1}} = 0 \text{ y el CV será } 0\%$$

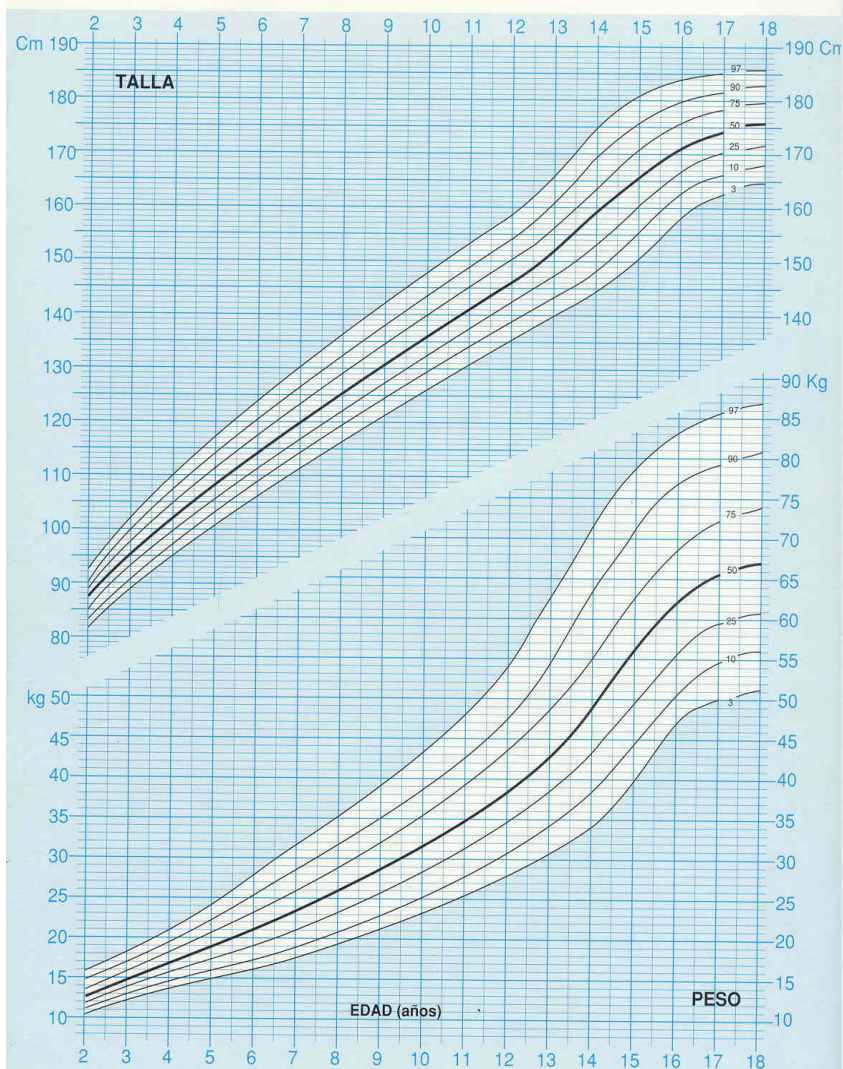
¡la media es buena representante!

2) siempre que sea posible, los índices se calcularán a partir de los datos originales, ya que los cálculos a partir de la tabla conllevan algo de error. Como puede verse en este resumen con parámetros de algunos ejemplos que se han ofrecido en este tema:

“70 DATOS”	Datos originales	Datos agrupados
Media	39,6	39,4
Desviación estándar	14,4	14,8
Mediana	40,5	40
Moda	59	34 , 41 , 48
Coefficiente de variación	36,4%	37,6%

NIÑOS: 2 a 18 años

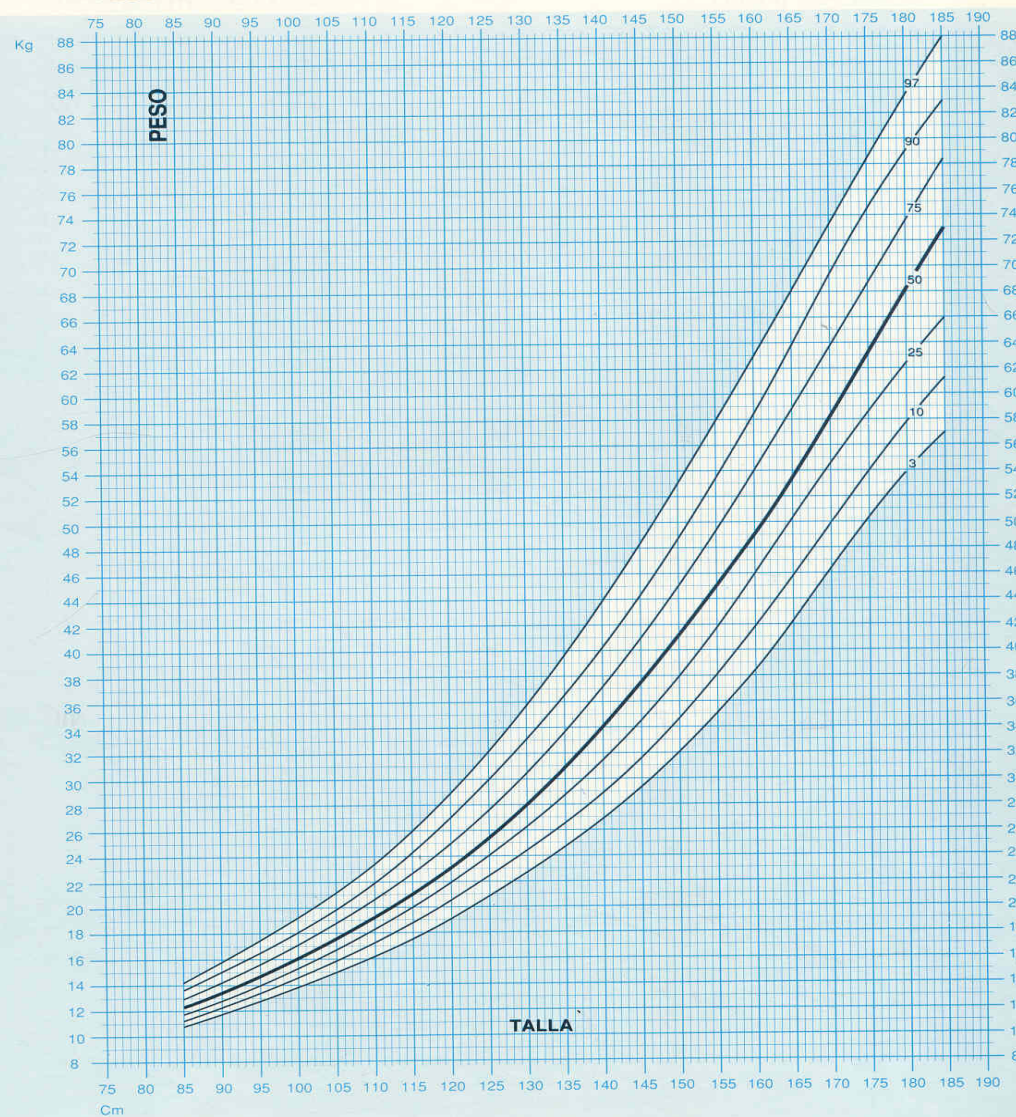
TALLA
PESO



ESTUDIO LONGITUDINAL DE CRECIMIENTO. CURVAS DE 0 A 18 AÑOS.
M. Hernández, J. Castellet, J. L. Narvaiza, J. M. Rincón, I. Ruiz,
E. Sánchez, B. Sobradillo y A. Zurimendi.

INSTITUTO DE INVESTIGACIÓN SOBRE CRECIMIENTO Y DESARROLLO.
FUNDACIÓN F. ORBEGOZO. María Díaz de Haro, 10 bis. 48013 BILBAO
Realización Gráfica: Garsi Editorial - Londres, 17 - 28028 MADRID

NIÑOS PESO-TALLA



ESTUDIO LONGITUDINAL DE CRECIMIENTO. CURVAS DE 0 A 18 AÑOS.
M. Hernández, J. Castellet, J. L. Narvaiza, J. M. Rincón, I. Ruiz,
E. Sánchez, B. Sobradillo y A. Zurimendi.

INSTITUTO DE INVESTIGACIÓN SOBRE CRECIMIENTO Y DESARROLLO.
FUNDACIÓN F. ORBEGOZO. María Díaz de Haro, 10 bis. 48013 BILBAO
Realización Gráfica: Garsi Editorial - Londres, 17 - 28028 MADRID

Tema 7 : DATOS BIVARIADOS. CORRELACION Y REGRESION.

Distribuciones uni- y pluridimensionales.

Hasta ahora se han estudiado los índices y representaciones de una sola variable por individuo. Son las distribuciones unidimensionales o univariadas .

En un individuo se pueden estudiar conjuntamente dos o más variables con objeto de ver si hay relación o dependencia entre ellas. Tenemos entonces distribuciones pluridimensionales, también llamadas plurivariadas. Cuando son dos se llaman bivariadas o bidimensionales. Son las únicas que veremos nosotros.

La simple medida de más de una variable en un individuo no tiene categoría de pluridimensional, sólo se tiene una serie de variables unidimensionales. ¡Hace faltar estudiarlas conjuntamente!

Estudio de variables bidimensionales

A una de las variables se la llama variable independiente y se representa por X. A la otra se la denomina variable dependiente y su símbolo es Y. (también se usan las minúsculas: x e y).

Los datos deben de ir siempre apareados. Para cada individuo se dan su X y su Y. (“Cada oveja con su pareja”). El nº de individuos se representa por N.

N es el nº de individuos, no el nº de datos, que siempre será el doble de N, pues cada individuo nos proporciona dos. ¡Es un error observado con frecuencia en los exámenes!

Ambas variables pueden ser cuantitativas (CT) o cualitativas (CL). En este tema veremos el caso de que ambas variables sean CT (que se completará en el tema 18) . En el tema 16 veremos la relación entre dos variables CL, expresada mediante la Odds ratio (OR). El caso de una variable CL y otra CT se trata en el tema 17.

--Ejemplos de variables bidimensionales

talla y peso, edad y tensión arterial, frecuencia cardíaca y frecuencia respiratoria, sexo y hábito de fumar, sexo y peso al nacer, velocidad de un vehículo y distancia de frenada...

Cuando ambas variables son CT, se pueden presentar:

- a) cada variable por separado (con sus tablas, gráficos e índices)
- b) conjuntamente (objeto de este tema) mediante:
 - a. la tabulación y representación gráfica de los datos
 - b. el cálculo de dos índices:
 - i. coeficiente de correlación
 - ii. ecuación de regresión

Tabulación

---de los datos originales

se hace una tabla, vertical u horizontal, con una columna (o fila) para X y otra para Y. Es opcional añadir otra para el número de orden del individuo. Los datos se ordenan en función del orden de los individuos o de los valores de X o de los valores de Y o no se ordenan en absoluto.

Ejemplo: Para X = (1 , 1 , 3 , 6 , 2 , 3 , 5 , 6) e Y = (1 , 1 , 4 , 4 , 2 , 5 , 1 , 5) :

Indiv.	X	Y	Ind.	1	2	3	4	5	6	7	8
1	1	1	X	1	1	3	6	2	3	5	6
2	1	1	Y	1	1	4	4	2	5	1	5
3	3	4									
4	6	4									
5	2	2									
6	3	5									
7	5	1									
8	6	5									

---de los datos agrupados en clases

Los valores de X e Y se agrupan en clases, siguiendo el método visto en el tema 4. La tabla es bidimensional: en la primera columna se representan las clases de X y en la primera fila las clases de Y. Al hacer el recuento los valores de cada individuo quedarán dentro de la casilla de la tabla que englobe a ambos.

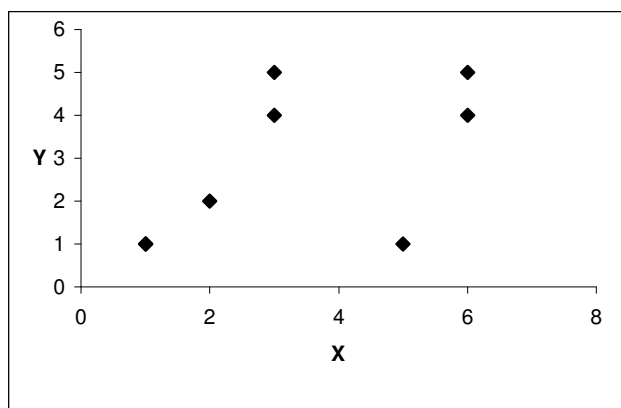
Ejemplo: Para los datos ya vistos la tabla podría ser así (presentada de forma simplificada y no del todo ortodoxa para mayor claridad):

X \ Y	1-2	3-4	5-6	TOTAL
1-2	3	0	0	3
3-4	0	1	1	2
5-6	1	1	1	3
TOTAL	4	2	2	8

Gráficos

--datos originales, aislados

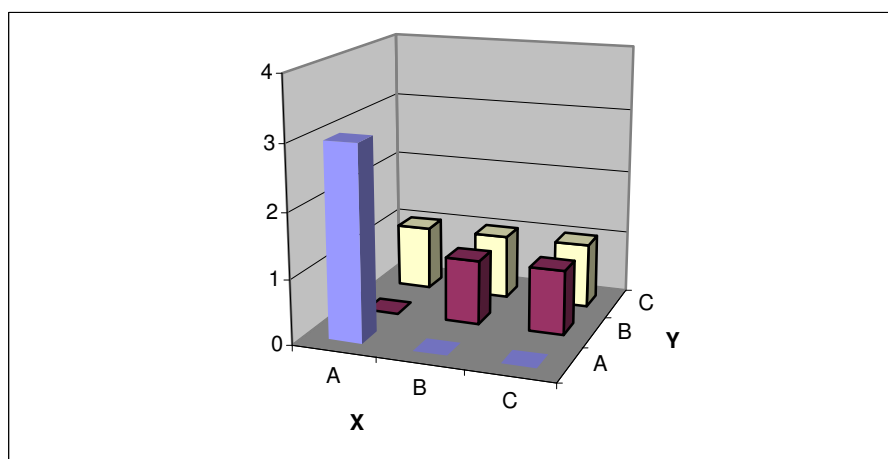
Es el diagrama de puntos, también llamado de dispersión o de nube de puntos. Los valores de cada individuo llevados aun eje de coordenadas originan un punto.



---datos agrupados en clases

El gráfico es el Estéreograma. Cada casilla de la tabla (que es la conjunción de dos clases, una de X y otra de Y) está representada por un prisma o cilindro (o incluso por una línea) cuya altura es proporcional a la frecuencia.

Para mayor claridad las clases en vez de como 1-2, 3-4 y 5-6 se representan como A, B y C



Índices estadísticos

Los típicos de estas distribuciones, aparte de los de cada variable por separado, son el coeficiente de correlación y la ecuación de regresión. Son los llamados índices o parámetros de asociación. Son distintos en función del tipo de variables (CL-CL, CL-CT, CT-CT). en este tema sólo nos ocuparemos del caso en que ambas variables son CT.

Correlación significa relación mutua y expresa el grado de asociación existente entre las variables, el CUANTO de la relación. Su parámetro es el coeficiente de correlación. Su símbolo es r , que puede acompañarse, si la claridad lo exige, de un subíndice con la notación de las variables (p.e. r_{xy}). Se puede calcular la correlación entre dos variables o más (correlación múltiple).

La **regresión** es la forma, el COMO de esa asociación. Expresa la relación entre las dos variables, X e Y, mediante la ecuación de regresión y su representación gráfica la línea de regresión. Mediante ella conocida una variable es posible predecir la otra. Por consenso X es la variable independiente e Y la dependiente. De esta forma $Y = f(X)$.

Coeficiente de correlación

Mide la intensidad de la asociación entre las variables. Es un número abstracto, independiente de la unidad de medida de las variables. Puede adoptar cualquier valor entre -1 y 1 . Dicho de otra forma: $r \in (-1 \div 1)$. Suele expresarse con 3 decimales, a no ser que valga -1 , 0 ó 1 . Aparte de su valor descriptivo sirve para ver la significación estadística de la relación (tema 18)

Aquí veremos sólo la correlación entre dos variables. Su coeficiente de correlación se llama de Pearson, aunque cuando se dice simplemente coeficiente de correlación, se sobreentiende que es éste. En el tema 18 se verá otro coeficiente, el de Spearman, que se usa cuando no puede utilizarse el de Pearson.

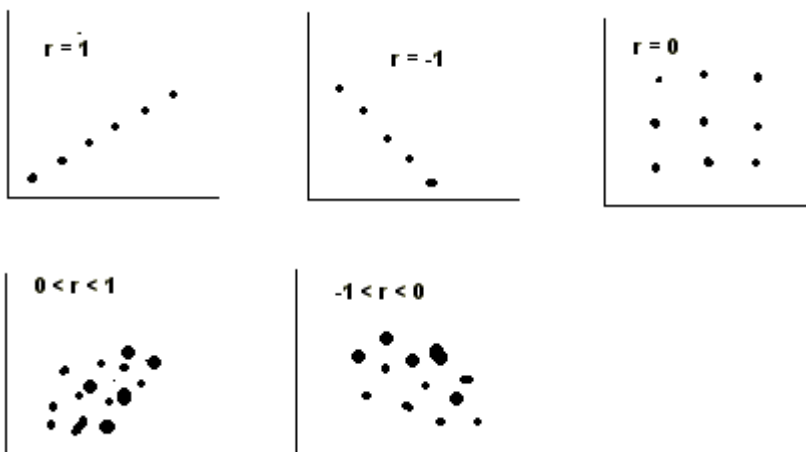
Si se observa una correlación aparentemente alta entre X e Y puede tratarse de dos situaciones:

--una variación de X provoca otra en Y. Por ejemplo, el aumento de la temperatura corporal produce un aumento de la frecuencia cardiaca.

--X e Y varían a la par por efecto de un a tercera o más variables. La correlación existente es pura coincidencia. Son las llamadas correlaciones espurias, ya citadas en el tema 1. Son las más frecuentes. De forma automática $\text{correlación} \neq \text{causalidad}$. Se requiere un estudio experimental con resultado significativo.

Si $r = 1$ hay una correlación total (perfecta) positiva.
Si $r = -1$ hay una correlación total (perfecta) negativa.
Si $r = 0$ no hay correlación.
Si está entre -1 y 0 , la correlación es parcial y negativa.
Si está entre 0 y 1 , la correlación es parcial y positiva.
Una r de 0 , -1 ó 1 apenas se encuentra en la práctica

Gráficamente esto se puede representar así:



Cálculo de coeficiente de correlación

Veremos únicamente el cálculo a partir de los datos originales, aislados.

$$r = \frac{N \sum XY - \sum X \sum Y}{\sqrt{\left[N \sum X^2 - (\sum X)^2 \right] \left[N \sum Y^2 - (\sum Y)^2 \right]}}$$

Para este cálculo y el de la ecuación de regresión es de gran ayuda construirse una tabla auxiliar como la que se utiliza en el siguiente ejemplo:

X = (2 , 1 , 3 , 2 , 5) ; Y = (3 , 5 , 4 , 2 , 6)

X	Y	X ²	Y ²	XY
2	3	4	9	6
1	5	1	25	5
3	4	9	16	12
2	2	4	4	4
5	6	25	36	30
13	20	43	90	57

$$\begin{aligned} r &= \frac{(5 * 57) - (13 * 20)}{\sqrt{[(5 * 43) - 13^2][(5 * 90) - 20^2]}} = \\ &= \frac{25}{\sqrt{46 * 50}} = 0,521 \end{aligned}$$

Este valor de r es el valor puntual. Cada día se utiliza más el valor por intervalo, cuyo cálculo veremos en el tema 13, en el que se estudian los intervalos de confianza (IC).

Regresión

Ya hemos visto el concepto de regresión. La fórmula matemática que la expresa puede ser una ecuación de primer grado (regresión lineal: $y = a + bx$) u otras ecuaciones más complejas (cuadrática: $y = ax^2 + bx + c$; exponencial: $y = ae^{bx}$; potencial: $y = ax^b$; hiperbólica: $y = a(b/x)$; logarítmica: $y = a + b \ln x$; etc...), que no trataremos, pues son muy complejas. Nos limitaremos a la regresión lineal, también llamada recta de regresión, pues su representación gráfica es una línea recta, que representa lo mejor posible a todos los puntos del diagrama de dispersión. Realmente se podrían trazar muchas rectas de regresión, pero sólo nos interesa la llamada “mejor línea de ajuste”, que es la que corresponde a la ecuación $y = a + bx$ (ó $y = bx + a$; el orden de los sumandos no altera la suma).

En esta fórmula **b** es el coeficiente de regresión, también llamado pendiente, pues de él depende la inclinación de la recta y nos indica en cuanto se modifica **y** en media cuando **x** varía en una unidad.

a es el valor de y cuando $x = 0$, por lo que también se la llama ordenada en el origen o intersección de y . Se ha comprobado que la mejor línea de ajuste es aquella en que la suma de los cuadrados de las diferencias entre cada punto original y la línea de regresión es la menor de todas las posibles. Por eso a este método se le llama “de los mínimos cuadrados”. Afortunadamente no hay que calcularlos, pues se ha desarrollado una fórmula mucho más manejable para encontrar la ecuación.

En principio se considera a y variable dependiente y a x variable independiente, por lo que la regresión se dice que es de y sobre x . En este sentido b es realmente b_{yx} y así se entiende cuando no hay subíndice. Matemáticamente también se puede calcular la regresión de x sobre y . Si interesara este cálculo, lo que no es habitual, escribiríamos b_{xy} para evitar confusiones.

Cálculo

Seguiremos el procedimiento que calcula primero **b** y a partir de él calcula **a**

$$b = \frac{N \sum XY - \sum X \sum Y}{N \sum X^2 - (\sum X)^2} \quad a = \bar{Y} - b \bar{X}$$

Ejemplo: Utilizando los datos empleados para calcular el coeficiente de correlación:

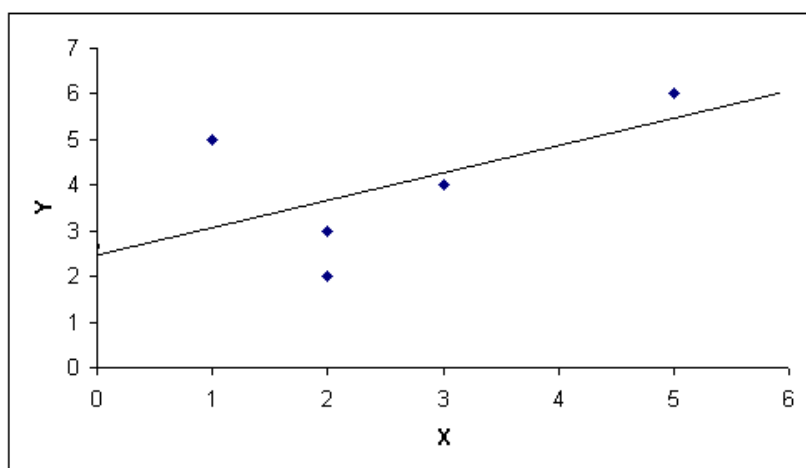
$$\bar{X} = \frac{13}{5} = 2,6 \quad \bar{Y} = \frac{20}{5} = 4 \quad b = \frac{(5 * 57) - (13 * 20)}{(5 * 43) - 13^2} = \frac{25}{46} = 0,54347$$

$$a = 4 - (0,54347 * 2,6) = 2,587$$

por tanto la ecuación es $y = 2,587 + 0,543x$

Representación gráfica

Para trazar una recta basta con dos puntos. En el diagrama de dispersión se busca el valor de y para $x = 0$. El otro punto se obtiene a partir de un valor cualquiera de x que nos de una y que no se salga del gráfico. En nuestro ejemplo: si $x = 0$, $y = 2,587$; para $x = 5$, $y = 5,302$



Se suele incluir en el gráfico la ecuación y el coeficiente de correlación y con menos frecuencia el IC (intervalo de confianza) de forma numérica y/o con dos rectas más que lo delimiten.

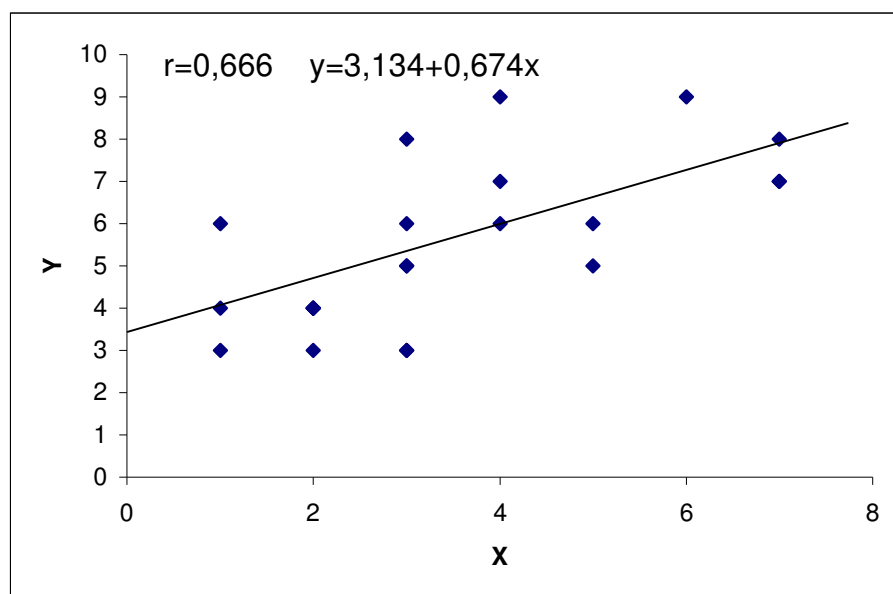
Coeficiente de determinación

Mide cuantitativamente la bondad o representatividad del ajuste de la recta a la nube de puntos. Es el cuadrado de r . Su símbolo es r^2 o R . En nuestro ejemplo $r^2 = 0,302$. Cuando se calculan diversas ecuaciones de regresión (lineal, exponencial, logarítmica, etc.) la que tenga el r^2 más alto será la mejor, la más representativa. r^2 unifica la fuerza de la asociación de positivos y negativos. (una $r = -0,400$ es más potente que una $r = 0,350$; sus r^2 son 0,160 y 0,122)

Ejercicio resuelto con Excel.

Ejercítese en el cálculo de la media, desviación estándar, CV, coeficiente de correlación y ecuación de regresión.

X	Y				X	Y
2	4	Error est. Y	1,472	media	3,478	5,478
3	3	r	0,666	s	1,904	1,928
5	5	Ecuación: b	0,674	CV	54,7	35,2
2	4	Ecuación: a	3,134	p50 ó M	3,000	5,000
1	3					
2	4	N = 23				
7	7					
3	6					
2	3					
4	6					
1	6					
3	3					
1	4					
3	5					
2	4					
6	9					
4	9					
3	8					
4	7					
3	5					
5	6					
7	7					
7	8					



Notas adicionales. 1) Con los datos del ejercicio anterior se han calculado otras ecuaciones de regresión con sus respectivos r y r^2 . Se dan aquí a título puramente informativo para que se vea que la mejor ecuación que relaciona a X e Y es la cuadrática, ya que tiene la r^2 más alta.

ECUACION	a	b	c	r	r^2
Cuadrática	-0,034	0,950	2,703	0,668	0,447
Lineal	3,134	0,674		0,666	0,443
Exponencial	3,334	0,125		0,659	0,434
Logarítmica	3,262	2,034		0,630	0,397
Potencial	3,412	0,378		0,625	0,390

2) aunque no es lo correcto, en la práctica se calcula en ocasiones r cuando se contrastan 2 Vbles. CT procedentes de individuos distintos, siempre que estén emparejados. Aquí N es el nº de parejas de datos, no el de individuos.

Tema 8 : Series de tiempo

Concepto

Una serie de tiempo representa las variaciones o evolución de un fenómeno a través del tiempo. Se concreta en una serie de observaciones de una variable, hechas en determinados intervalos de tiempo, generalmente iguales. Son datos bivariados en los que la variable independiente es el tiempo, que se simboliza por t en vez de por x .

Son muy utilizadas en la vida diaria: evolución en un determinado periodo de tiempo de la producción de coches, exportaciones, turistas que nos visitan, paro, etc. La clásica curva de la fiebre y pulso de un paciente es una serie de tiempo. Los modernos monitores de las llamadas constantes vitales y los barógrafos, termógrafos y aparatos similares hacen un registro continuo de una o más variables.

Representación

- a) de forma numérica o tabular. La columna base es el tiempo.
- b) de forma gráfica. La más usada es el diagrama lineal, la variante del polígono de frecuencias que no baja al eje de abscisas ya que no se abarca toda la distribución sino sólo una parte de la misma. Si abarca toda la distribución se usará el polígono de frecuencias. En el eje de abscisas se representa el tiempo y en el de ordenadas la frecuencia correspondiente.

La tabla puede acompañarse de una columna con números índice, que en general parten de considerar como 100 ó 100% al valor de Y en el primer periodo de tiempo. Para los demás periodos se hace el cálculo por una simple regla de tres. También puede ponerse una columna que represente una tasa.

Ejemplo:

HOSPITAL H			
Ingresos del Servicio S			
año	ingresos	Nº índice	tasa/100.000 hab.
2000	800	100	200
2001	915	114	229
2002	980	122	245
2003	1040	130	260
2004	1000	125	250
2005	980	122	240

Otros cálculos

Los más utilizados son el coeficiente de correlación y la ecuación de regresión.

Lo esencial de las series de tiempo

Su estudio ha permitido comprobar que están sometidas a **variaciones típicas**, siendo las más importantes las tres siguientes:

--variaciones a largo plazo o tendencia secular. Representan la variación general de la serie, suavizada por la absorción de otras variaciones menores en intervalos de tiempo largos. Podría decirse que los datos utilizados son medias de otros muchos datos. Un ejemplo típico es la talla media de los chicos españoles cuando se incorporaban al servicio militar obligatorio, registrada durante casi un siglo.

--variaciones a medio plazo o fluctuaciones periódicas, obtenidas en intervalos de tiempo menores. Pueden ser estacionales y cíclicas. Son estacionales cuando el plazo es menor de un año.

Ejemplo típico son las ventas de unos grandes almacenes en Navidad-Reyes, San Valentín, Día de la Madre, etc. Las cíclicas ocurren a intervalos mayores de un año, como los ciclos de la economía. Suelen ser más suaves.

--variaciones irregulares o accidentales. No son previsibles, como el aumento de las ventas de determinados alimentos cuando se rumorea que van a subir mucho de precio o la disminución de la producción de una fábrica durante una huelga. Estas variaciones pueden originar nuevos ciclos o tendencias, como la crisis pesquera de los años 70, que elevó mucho los precios, sin vuelta atrás. O el aumento imparable del precio del petróleo tras la primera invasión de Irak.

Análisis de las series de tiempo

Es una especialidad de la Estadística. No podemos entrar en sus procedimientos, pues son muy complejos y desbordan las posibilidades de tiempo de esta asignatura. Únicamente veremos sus aplicaciones. Las principales son:

--descripción y estudio de un fenómeno a lo largo del tiempo con todas sus variaciones.

--predicción de la tendencia para el futuro. Se basa en la ecuación de regresión, mejor con su intervalo de confianza, lo que da una horquilla de posibles situaciones. Aquí hace falta una buena dosis de experiencia y sentido común. Utilizando la ecuación de regresión de la mortalidad de una enfermedad en los primeros años tras introducir una vacuna eficaz, se puede llegar fácilmente a una mortalidad negativa, es decir, a la resurrección de los muertos...

Precauciones

Las series de tiempo se prestan mucho a la manipulación. Por ejemplo utilizando variaciones cíclicas, o incluso accidentales, como si fueran tendencias a más largo plazo. O tomando como punto de partida de la serie un “momento conveniente” para lo que interesa. Valorarlas siempre con espíritu crítico.

Otro ejemplo:

Hotel del Golfo
Estancias agosto últimos 5 años

Año	Estancias	Nº índice
2001	2980	100.0
2002	3050	102.3
2003	3130	105.0
2004	3020	101.3
2005	3260	109.4

$r = 0,757$

$$Y = 48,2 * X - 93420,2$$

o sea, **Estancias = $48,2 * \text{año} - 93420,2$**

Predicciones:

año 2006: Estancias = $48,2 * 2006 - 93420,2 = 3269$

año 2007: Estancias = $48,2 * 2007 - 93420,2 = 3317$

Tema 9 : Teoría de la probabilidad

Definición

Veremos dos:

---La **definición clásica de Laplace** dice que la probabilidad, (p), de ocurrencia de un fenómeno A (o evento, suceso, modalidad de una variable...) en un experimento aleatorio de resultados equiprobables es igual al n° de casos favorables, también llamados éxitos, (símbolo: f ó r) dividido por el n° de casos posibles (N).

$$p_A = f/N$$

Como f puede estar entre 0 y N, los valores posibles de p van de 0 a 1. Suelen expresarse, salvo el 0 y el 1, con 3 ó 4 decimales. También se puede expresar como porcentaje, entre 0% y 100%. A veces es conveniente, por ser más manejable, expresarlo como fracción.

Tres aclaraciones a esta definición

1-Un experimento aleatorio

- no tiene resultado fijo, sino un conjunto de posibles resultados (2 ó más)
- el resultado no se conoce de antemano, ocurre de forma aparentemente casual.
- se puede repetir indefinidamente bajo las mismas condiciones.

2- *Equiprobable* quiere decir que todos los resultados tienen la misma probabilidad de ocurrir. Ejemplo: la probabilidad de que al tirar un lado salga un 3 es 1/6. (1/6 es preferible a 0,1667). El modelo de Laplace es un modelo teórico, intuitivo, en el que por simple reflexión se pueden saber las probabilidades.

3- *Éxito* se utiliza cuando ocurre el evento. El término es un clásico y se introdujo estudiando tiradas de dados, aplicándose aunque el evento sea algo negativo. Si se estudia la mortalidad, un fallecimiento será un “éxito”...

---La **definición de Richard von Mises** es más amplia y universal, basada en un modelo experimental, práctico: “La mejor estimación de la probabilidad de la ocurrencia de un fenómeno en un experimento aleatorio es su frecuencia relativa”.

Ejemplo: Teóricamente al lanzar una moneda bien hecha la p de cara es de 0,5. Hacemos un experimento tirando la moneda repetidamente. Vamos anotando como éxito las caras que van saliendo y después de cada tirada se calcula la f.r. de éxitos. Tras variaciones de cierta amplitud al principio pronto la f.r. se mueve cada vez más cerca de 0,5, con el que coincidirá exactamente en el infinito.

De esta forma calculando la f.r. podemos hallar la probabilidad de sucesos en los que no podemos utilizar la intuición. Por ejemplo, tirando varios cientos de chinchetas del modelo X al suelo, la f.r. de las que queden con la punta hacia arriba nos dará la p de tal resultado en ese modelo.

No tiene valor estadístico la llamada probabilidad subjetiva, que es una mezcla del conocimiento de los factores que pueden influir en un resultado con factores emocionales. Como la p de que nuestro equipo favorito gane el próximo partido o de aprobar una asignatura a la primera..

Sucesos elementales y complejos

Suceso elemental es el suceso básico, como p.e. nacer chica, cuya p es de 0,5

El suceso complejo comprende varios elementales, como p.e. tirar dos dados o el n° de chicas en una familia de 5 hijos. En algunos casos es fácil calcular sus probabilidades de ocurrencia con las reglas que se ven a continuación, pero en la mayoría hay que recurrir a las distribuciones fundamentales de probabilidad, que se verán en el tema 10

Algunos conceptos básicos de la probabilidad

- 1- $0 \leq p \leq 1$ ó $0\% \leq p \leq 100\%$
- 2- $\sum p(A_x) = 1$, siendo A_x el dominio de la variable, o sea todas sus modalidades o valores

- 3- Si A es el suceso elemental con probabilidad p_A , la probabilidad de que no ocurra A, es decir, de que ocurra el suceso contrario o complementario (\bar{A}) es $1-p$ ó q .
 Por tanto $p_{\bar{A}} = 1-p = q$; $q_A = 1 - q$
 Un suceso elemental y su complementario son mutuamente excluyentes, incompatibles, no pueden ocurrir simultáneamente. Un suceso complementario puede ser simple o múltiple. Simple o sencillo, cuando sólo tiene una modalidad (caso de una moneda). Múltiple o compuesto, cuando engloba varias modalidades (caso de un dado).
- 4- $p + q = 1$ ó $p + q = 100\%$
- 5- Son sucesos independientes aquellos cuya ocurrencia no depende de otro u otros sucesos. Por ejemplo, que al tirar dos dados en una salga 4 y en el otro 2.
 Son sucesos dependientes aquellos cuya ocurrencia depende de otro u otros sucesos. Si sacamos dos cartas de una baraja española, la p de que la segunda seaoros depende del palo de la primera carta. se formula así: $p(A_2/A_1)$, “p de A_2 dado A_1 ”.
- 6- Ley multiplicativa. Rige la p de que ocurran a la vez dos o más sucesos (que por fuerza tienen que ser compatibles).
 a. si son independientes: $p(A_1 \text{ y } A_2) = p_{A_1} * p_{A_2}$
 b. si son dependientes: $p(A_1 \text{ y } A_2) = p_{A_1} * p(A_2/A_1)$
- 7- Ley aditiva. Rige la p de que ocurra un suceso u otro.
 a. si son incompatibles. $p(A_1 \text{ o } A_2) = p_{A_1} + p_{A_2}$
 b. si son compatibles: $p(A_1 \text{ o } A_2) = p_{A_1} + p_{A_2} - p_{A_1} * p_{A_2}$
 ya que hay que restar la compatibilidad.

Ejemplos

- a) p de que al tirar un dado dos veces salgan en ambas un 6.
 “seis en la 1ª tirada y 6 en la 2ª”
 $p(2 \text{ veces } 6) = 1/6 * 1/6 = 1/36$ (mejor que 0,0278)
- b) p de que al tirar dos dados salga en ambos un 6
 “seis en el primer dado y seis en el segundo”
 es el mismo caso que a)
- c) La p de ser rubio es de 0,3 y la de llevar gafas es de 0,2 . Calcular la p de que una persona cualquiera sea rubia y lleve gafas (se asume que son independientes)
 $p(\text{rubio y gafas}) = 0,3 * 0,2 = 0,06$ (ó 6%)
- d) en una caja hay 3 bolas blancas y 2 negras. Calcular la p de que sacando dos bolas, las dos sean negras.
 Nos piden la p de que sea negra la primera y negra la segunda.
 la p de ser negra de la 1ª bola es $2/5$; una vez sacada quedan 4 bolas (una, negra)
 la p de ser negra de la 2ª bola es de $1/4$
 $p(2 \text{ bolas negras}) = 2/5 * 1/4 = 2/20 = 1/10$ (ó 0,1 ó 10%)
- e) p de que al sacar una carta de una baraja española de 40 cartas seaoros o copas.
 $p(\text{oros o copas}) = 10/40 + 10/40 = 20/40 = 1/2$ (ó 0,5 ó 50%)
- f) p de que al sacar una carta de esa baraja sea as o espadas.
 hay 4 ases , 10 espadas y 1 as de espadas (que cuenta como as y como espada, 1 entre 40, que debe ser compensada)
 $p(\text{As o Espada}) = 4/40 + 10/40 - 1/40 = 13/40 = 0,325$
- g) p de acertar 6 en la Primitiva
 Hay 49 bolas. Como no hay reemplazo, cada vez que sale una bola, queda una menos en el bombo. Para acertar los 6 resultados hay que acertar el primer número y el segundo y el tercero...y el sexto.
 $p(6 \text{ aciertos}) = 6/49 * 5/48 * 4/47 * 3/46 * 2/45 * 1/44 = 1 / 13.983.816$
- h) p de que tirando un dado 4 veces, la primera vez que salga un 5 sea en la 4ª tirada.
 $p(5 \text{ sólo en la } 4^a) = p(\text{no } 5 \text{ en la } 1^a) * p(\text{no } 5 \text{ en la } 2^a) * p(\text{no } 5 \text{ en la } 3^a) * p(5 \text{ en la } 4^a)$
 $= 5/6 * 5/6 * 5/6 * 1/6 = 125/1296 = 0,096$

a) p de al menos un éxito (es decir, uno o más, uno como mínimo) en n intentos

se resuelve así : $p(r \geq 1) = 1 - p(r=0)^n$ ¡Ojo! no es $1 - p(r=0) \cdot n$

Ejemplo: Un problema importante en la prevención del tétanos cuando no había vacunas o gammaglobulinas y había que administrar suero antitetánico eran las reacciones, a veces muy graves, que ocurrían en un 10% de los inyectados.

En una persona que hubiera recibido 10 inyecciones ¿cual es la p de que al menos tuviera una reacción?

Si la p de tener una reacción es de 0,1, la de no tenerla es de 0,9. Por tanto $p(r \geq 1) = 1 - 0,9^{10} = 0,651$. Si, falsamente, se hubiera calculado $1 - 0,9 \cdot 10$ se obtendría un resultado imposible: $p = -8$

Distribución de probabilidad

es el conjunto de las p de todas los valores o modalidades que puede adoptar una variable X.

Veamos el caso más sencillo, el de una variable cualitativa:

- se establece el dominio de la variable (todas las modalidades)
- se calcula la p de cada modalidad
- se tabula y se representa gráficamente

ejemplo: X = suma de puntos al tirar dos dados

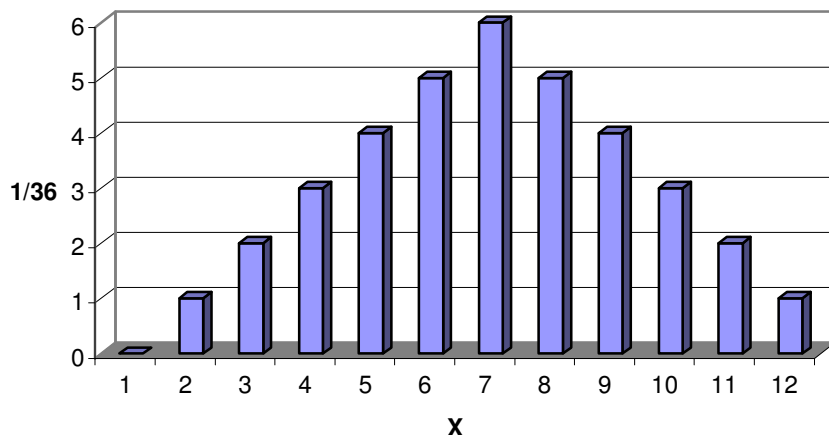
dominio: hay 36 combinaciones posibles (zona sombreada)

		dado 1					
		1	2	3	4	5	6
dado 2	1	2	3	4	5	6	7
	2	3	4	5	6	7	8
	3	4	5	6	7	8	9
	4	5	6	7	8	9	10
	5	6	7	8	9	10	11
	6	7	8	9	10	11	12

probabilidad:

x	1	2	3	4	5	6	7	8	9	10	11	12
px	0	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

gráfico :



Método de Bayes

El modelo estadístico bayesiano se basa en probabilidades condicionadas y ha permitido el desarrollo, aún bastante imperfecto, del “diagnóstico por ordenador”. A partir de las frecuencias de determinados síntomas en diversas enfermedades calcula la p de padecer una u otra enfermedad. Es un compleja especialidad dentro de la Estadística, cuyos detalles escapan a la intención de esta asignatura. Veremos su fórmula general y un ejemplo.

Fórmula de Bayes

$$p(A_x / E) = \frac{pA_x * p(E / A_x)}{\sum_{i=1}^n [pA_i * p(E / A_i)]}$$

pudiendo valer x entre 1 y n

Ejemplo

Se sabe que la presencia de determinados síntomas se da en el 60% de pacientes con la enfermedad A1, en el 30% de los que padecen la enfermedad A2 y en el 10% de los que tienen la enfermedad A3.

Al análisis E sale positivo en el 30% de los casos de A1, en el 70% de los casos de A2 y en el 70% de los de A3.

Si un paciente tiene esos síntomas y el análisis sale positivo, ¿qué probabilidades hay de que tenga una u otra enfermedad?

Enferm. (Ax)	pAx	p(E+/Ai)	pAx*p(E/Ai)	pAx/ E+
A1	0,6	0,3	0,18	0,18/0,46 = 0,391 = 39,1%
A2	0,3	0,7	0,21	0,21/0,46 = 0,456 = 45,6%
A3	0,1	0,7	0,07	0,07/0,46 = 0,152 = 15,2%
Suma .			0,46	

La enfermedad más probable es la A2, seguida de cerca por la A1 y más lejos por la A3.

Tema 10 : Distribuciones fundamentales de probabilidad

Ya hemos visto que los fenómenos naturales siguen el modelo indeterminista, es decir las leyes del azar, entendido como la combinación de múltiples factores, en gran parte desconocidos e incontrolables, que conducen a resultados no previsibles de antemano, aunque sí conocidos, que se caracterizan por su variabilidad en los diferentes individuos. A cada uno de los posibles resultados se asocia una probabilidad, que en sucesos sencillos o poco complejos es fácil de calcular por las leyes básicas o fundamentales de la probabilidad, pero al aumentar la complejidad el cálculo se hace muy difícil o imposible. Entonces hay que recurrir a una serie de modelos teóricos, las llamadas distribuciones o leyes fundamentales de la probabilidad, que nos permiten hacer el cálculo con relativa facilidad. Al aumentar el nº de individuos todas las distribuciones se van aproximando y acaban confluyendo y haciéndose una en el infinito.

Clasificación

- a) para variables discretas
 - D. binomial
 - D. polinomial
 - D. de Poisson
 - D. hipergeométrica
- b) para variables continuas
 - D. normal
 - D. de la t de Student
 - D. de la χ^2 de Pearson
 - D. de la F de Snedecor-Fisher

Para todas valen los principios que ya conocemos:

$$0 \leq p \leq 1$$

$$p + q = 1$$

$$\sum p(x) = 1$$

En este tema nos ocuparemos de las distribuciones binomial, de Poisson, normal y hipergeométrica. En el Anexo se verán la t de Student, la χ^2 y la F. No veremos la polinomial.

DISTRIBUCION BINOMIAL

Concepto

es el modelo básico de distribución de las variables discretas (o discretizadas), que como ya sabemos pueden ser reducidas en última instancia a dicotómicas.

Experimentos binomiales

Pueden ser elementales y complejos

Los elementales tienen dos resultados posibles: Éxito (cuando aparece el resultado que se pretende) y fracaso, que puede ser único o múltiple. Sus probabilidades respectivas son p y q

En los complejos

- el experimento elemental se repite n veces
- obteniendo r éxitos (de 0 a n) : $0 \leq r \leq n$
- cada modalidad de la variable va asociada a una r . Como r empieza en 0 siempre hay n+1 modalidades: la de r=0 y las de r entre uno y n.
- un experimento binomial complejo puede repetirse N veces. Cada modalidad aparecerá **Nr** veces.

Notación

La distribución suele designarse como DB, pero cuando se dan los parámetros típicos, la n y la p del suceso elemental, se utiliza sólo B . Así: **B(n , p)**

Algunos ejemplos:

Experimento	Éxito	p	n	r	notación
elemental: lanzar 1 moneda	salir cara	0,5	1	0, 1	B(1, 0,5)
complejo: lanzar 4 monedas	salir cara	0,5	4	0, 1, 2, 3, 4	B(4, 0,5)
elemental: lanzar un dado	salir 1	1/6	1	0, 1	B(1, 1/6)
complejo: lanzar 5 dados	salir 1	1/6	5	0, 1, 2, 3, 4, 5	B(5, 1/6)
elemental: familia con 1 hijo	ser chica	0,5	1	0, 1	B(1, 0,5)
complejo: familia con 4 hijos	ser chica	0,5	1	0, 1, 2, 3, 4	B(4, 0,5)

El lanzamiento de las 4 monedas se puede repetir N veces.

O podemos estudiar N familias de 5 hijos.

Cálculo de las p de r

1) fórmula
$$p(r) = \binom{n}{r} p^r q^{n-r} = \frac{n!}{r! * (n-r)!} p^r q^{n-r}$$

$\binom{n}{r}$ da los coeficientes del desarrollo del binomio de Newton

2) tablas (en la pagina 16 hay una para $n \leq 8$ y ciertos valores de p)

3) Método intuitivo (la clásica “cuenta de la vieja”) posible en algunos casos.

Gráfico : diagrama de barras

Otros parámetros

Media o esperanza matemática: $\bar{X} = np$

la media representa el n° esperado de éxitos en el experimento

Varianza: $s^2 = npq$

y por tanto, **desviación estándar:** $s = \sqrt{npq}$

n, p, N y Nr

conviene insistir en estos símbolos que son básicos en la DB.

n : veces que se repite el suceso elemental en un experimento binomial. Si n=1 es un experimento simple; si >1, es complejo

p : probabilidad del suceso elemental

N : veces que se repite el experimento complejo. Si no se dice nada, N=1

N_r : frecuencia de cada modalidad tras N repeticiones. $\sum N_r = N$

----Si tiramos una moneda 1 vez, es una B(1, 0,5) . Podemos obtener 0 ó 1 cara (r). N=1

Si este experimento lo repetimos 3000 veces (N) seguirá siendo una B(1, 0,5) pero con N=3000. r sigue valiendo 0 y 1. Nos pueden salir p.e. 1450 caras. Entonces $N_0 = 1550$ y $N_1 = 1450$

----Si tiramos de una vez 3000 monedas pueden salir entre 0 y 3000 caras (r). Es una B(3000, 0,5) ; n=3000 ; N=1 Si obtenemos 1450 caras (c), habrá habido 1550 cruces (k). Como sólo se hace una vez, se suele asimilar al caso anterior y se dice que $N_0 = 1550$; $N_1 = 1450$, aunque realmente no es correcto. Mejor sería N_c y N_k

-----Si tiramos tres monedas 1000 veces y obtenemos 0 caras en 115 ocasiones, una cara en 380, dos caras en 370 y tres caras en 130: es una B(3 ; 0,5) , n=3 , N=1000 , $N_0=115$, $N_1=380$, $N_2=370$ y $N_3=130$

Problemas asociados a la DB

1) **calcular $p(r)$** : nos pueden pedir el cálculo de una r en concreto o de todas ellas. Como ejemplo vemos la p de 2 caras lanzando 3 monedas. Es $B(3, 0,5)$

1- aplicando la fórmula (de las dos que se han visto la más fácil es la segunda)

$$p(r=2) = \frac{3!}{2! \cdot 1!} 0,5^2 0,5^1 = 0,3750$$

2- consultando la tabla (ver página 16) ya que en este caso se puede utilizar. Es una tabla de doble entrada con valores de n y r en la primera columna y ciertos valores de p en la primera fila.

En una $B(3, 0,5)$ $p(r=2) = 0,3750$

3- método intuitivo (“cuenta de la vieja”). Válido para una p elemental de 0,5. Veremos no sólo la $p(r=3)$ sino todas las $p(r)$. Hay que considerar todas las combinaciones posibles de cara (c) y cruz (k)

r	modalidades	$\binom{n}{r}$	$p(r)$
0	k k k	1	1/8
1	c k k k c k k k c	3	3/8
2	c c k c k c k c c	3	3/8
3	c c c	1	1/8
Σ		8	1

$$3/8 = 0,3750$$

2) **calcular N_r** : es decir, la frecuencia de cada modalidad al repetir el experimento binomial N veces $N_r = N p(r)$

Si el lanzamiento de las 3 monedas se repite 200 veces, teóricamente se obtendrán lo siguiente:

$$0 \text{ caras : } N_0 = 200 * 1/8 = 25$$

$$1 \text{ cara : } N_1 = 200 * 3/8 = 75$$

$$2 \text{ caras : } N_2 = 200 * 3/8 = 75$$

$$3 \text{ caras : } N_3 = 200 * 1/8 = 25$$

3) calcular la media, varianza, desviación estándar

$$\bar{x} = np ; s^2 = npq ; s = \sqrt{npq}$$

En el ejemplo de las monedas:

$$\bar{x} = 3 * 0,5 = 1,5$$

$$s^2 = 3 * 0,5 * 0,5 = 0,75$$

$$s = \sqrt{3 * 0,5 * 0,5} = 0,866$$

4) **calcular los parámetros de una DB**, n y p , a partir de las frecuencias de las modalidades, es decir, a partir de N_r

n lo conocemos por los datos que nos dan.

$$p \text{ se calcula a partir de } \bar{x} = np \text{ y } \bar{x} = \frac{\sum(rN_r)}{N}$$

Ejemplo:

Lanzadas 4 monedas 10000 veces se han obtenido los resultados que se muestran en la tabla:
0 caras en 4096 ocasiones, 1 cara en 4096, 2 caras en 1536, 3 caras en 256 y 4 caras en 16.

r	Nr	r*Nr
0	4096	0
1	4096	4096
2	1536	3072
3	256	768
4	16	64
Σ	10000	8000

$$\bar{x} = \frac{8000}{10000} = 0'8 \quad 0'8 = 4p \quad p = 0'2$$

por tanto es una $B(4, 0'2)$

6) al crecer n la DB se llega a hacer inmanejable y la solución es **aproximarla** a otra Distribución fundamental transformando los parámetros originales en los propios de la distribución a la que se aproxima. Siempre que se cumplan ciertas condiciones.

- **a la DN**, si p y $q \geq 0,1$ (ó 10% si es %) y np y $nq \geq 5$ (ó 10 y 500 si es un %) se verá al tratar la DN
- **a la DP**, si p o $q \leq 0,1$ (ó 10% si es %) y np o $nq \leq 5$ (ó 10 y 500 si es %), aunque algunos admiten np o nq hasta 10 (ó 1000 si es %). Como veremos enseguida la DP es una variante de la DB y su parámetro λ es igual a $n \cdot p$, por lo que la aproximación es muy fácil.

7) **comprobar el ajuste** de unos datos (una distribución real u observada) a una DB ideal

Para ello hay que calcular una distribución binomial teórica, que tenga los mismos parámetros que la real. Como partiremos de las frecuencias de cada modalidad, hay que utilizar el procedimiento visto en 5). Luego se contrastan las frecuencias teóricas con las observadas por medio de una prueba de contraste de frecuencias, cuyo resultado se valora por χ^2 . Si no se encuentran diferencias significativas, el ajuste es bueno, En caso contrario es malo.

Ejemplo: En un lote de 800 piezas cada una de las cuales tiene tres soldaduras se han observado las siguientes frecuencias de defectos de soldadura: 0 defectos en 97 ; 1 defecto en 305 ; 2 defectos en 297 y 3 defectos en 101. Comprobar el ajuste a una DB.

$$a) \bar{x} = \frac{(0 \cdot 97) + (1 \cdot 305) + (2 \cdot 297) + (3 \cdot 101)}{800} = 1,5 \quad p = \frac{1,5}{3} = 0,5$$

b) cálculo de una $B(3 ; 0,5)$ con $N=800$

r	p(r)	N _r
0	0,125	100
1	0,375	300
2	0,375	300
3	0,125	100
Σ		800

Las p (r) se pueden leer directamente en la tabla de la DB

recordar que $N_r = N \cdot p(r)$

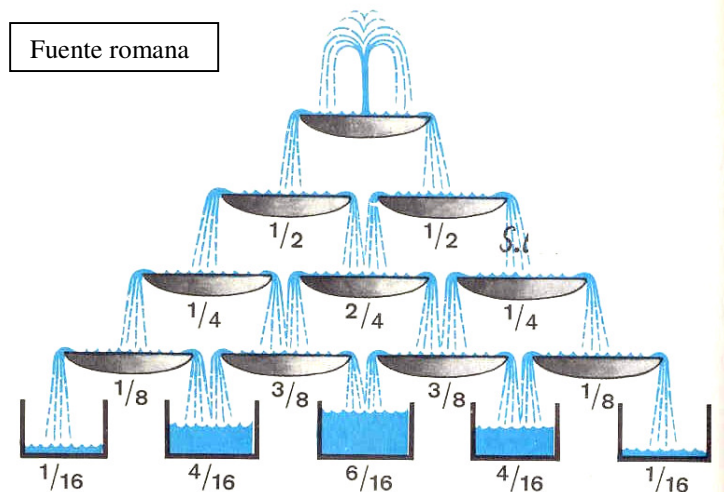
c) Ahora se contrastan las frecuencias observadas y las teóricas:

f observadas	97	305	297	101
f teóricas	100	300	300	100

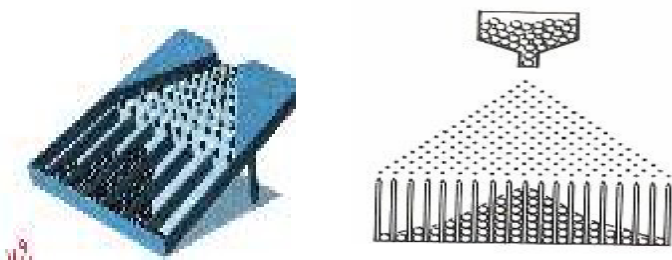
A simple vista se ve que el ajuste es muy bueno. Aplicando la prueba de contraste que veremos en el tema 16 la $z=0,213$ que no es significativa y por tanto el ajuste es bueno.

Modelos clásicos de la distribución binomial

Los más importantes son las fuentes romanas, el aparato de Galton y el triángulo de Pascal.



La mitad del agua que sale por la fuente de arriba cae por cada lado. Y lo mismo ocurre con las demás fuentes. Al final unos recipientes recogen el agua. Siguiendo el camino del agua, se ve que el volumen recogido aumenta hacia en el centro. Una fuente perfecta sigue exactamente la DB. El primer recipiente corresponde a $r=0$, el 2º a $r=1$, el 3º a $r=2$, etc. El nº de recipientes por tanto es igual a $n+1$.



El aparato de Galton sigue el mismo principio. Es una especie de embudo inclinado con filas de clavos, situados como las fuentes. Al final hay unos cajones receptores. Se lanza una bola que cada vez que choca con un clavo tiene la misma probabilidad de ir a la derecha que a la izquierda.

1
1 1
1 2 1
1 3 3 1
1 4 6 4 1
1 5 10 10 5 1
1 6 15 20 15 6 1
1 7 21 35 35 21 7 1
1 8 28 56 70 56 28 8 1
1 9 36 84 126 126 84 36 9 1

El triángulo de Pascal empieza por el 1 de la primera fila. Los números de las otras filas se obtienen sumando los dos que están por encima de él a derecha e izquierda. Como en los lados siempre se suma el 1 con nada, todos son 1. Se pueden construir el nº de filas que uno quiera. En cada fila los números corresponden a los coeficientes $\binom{n}{r}$ para cada valor de r , de 0 a n . Por tanto n es igual al nº de coeficientes menos 1. La suma de los coeficientes de cada fila es igual a 2^n .

DISTRIBUCIÓN DE POISSON

también llamada de los sucesos raros o de las probabilidades pequeñas.

Es una variante de la DB cuando p o q son muy pequeñas y n no es muy grande. En esta situación la DB se hace inexacta. La frontera se fija como se ha visto al tratar la aproximación de la DB a una DP en p ó $q \leq 0,1$ (ó el 10%, si se expresa en %; algunos admiten hasta 0,2 ó 20%) y np ó $nq \leq 5$ (ó 500 si se expresa como %), aunque últimamente se acepta hasta 10 (ó 1000). Como en origen es una DB, es válido lo que hemos visto sobre n , r , N_r y N .

Aunque un suceso sea raro, ocurre de vez en cuando. Incluso con cierta frecuencia, si aumenta el nº de ocasiones para que ocurra. Ya vimos que la p de acertar 6 en la Primitiva es bajísima, pero como se hacen millones de apuestas, hay muchas semanas con uno o más acertantes. En un determinado cruce puede ser que la probabilidad de que un coche tenga un accidente sea muy baja, pero si el tráfico es muy intenso, puede haber accidentes incluso todos los días.

Al contrario, un hecho frecuente, como las llamadas que se reciben en la centralita telefónica de un hospital, se puede convertir en raro si consideramos las llamadas en una unidad de tiempo muy pequeña, p.e. segundos. En 24 horas quizá en la mayor parte de los segundos no haya ninguna llamada.

¡Fijarse también en q , no sólo en p ! . Una $B(5, 0'98)$ tiene la $q=0,02$ y debe ser aproximada a una $P(4,9)$

Notación

$P(\lambda)$, siendo $\lambda = np$ (λ es la letra griega lambda)

Cálculo de $p(r)$

$$p(r) = \frac{\lambda^r}{r!} e^{-\lambda}$$

el valor de $e^{-\lambda}$ (e es la base de los logaritmos neperianos) se puede hallar con una calculadora científica o leer en una tabla (página 15). La tabla tiene dos partes: una va de λ entre 0,00 y 0,99 . La otra parte da $e^{-\lambda}$ para valores enteros de λ entre 1 y 10. Para valores con decimales en este intervalo se descompone λ en dos partes: una entera y la otra decimal . Por ejemplo: $\lambda = 3,48$ se descompone en 3 y 0,48. Los valores de $e^{-\lambda}$ se pueden leer en la tabla y hay que multiplicarlos, ya que este procedimiento se basa en que el producto de dos potencias de la misma base es otra potencia con la misma base y cuyo exponente es la suma de los exponentes.

Ejemplos: Calcular $p(r=3)$ para una $P(0,25)$ y para una $P(3,48)$

$$1) p(r=3) = \frac{0,25^3}{3!} e^{-0,25} = 0,0020$$

$$2) p(r=3) = \frac{3,48^3}{3!} e^{-3,48} = 7,024 * (0,04979 * 0,6188) = 0,2164$$

Media, varianza y desviación estándar

$$\bar{X} = \lambda = np \quad \bar{X} = \frac{\sum(rN_r)}{N} \quad s^2 = \lambda \quad s = \sqrt{\lambda}$$

Gráfico : es también el diagrama de barras

Problemas asociados a la DP

son similares a los vistos en la DB, ya que es una variante de la misma.

1) **calcular $p(r)$** : utilizando la fórmula

2) **calcular N_r** : es decir, la frecuencia de cada modalidad al repetir el experimento N veces

$$N_r = N * p(r)$$

3) **calcular el parámetro λ** a partir de las frecuencias de las modalidades, es decir, a partir de

N_r , **utilizando las fórmulas ya conocidas de la DB**: $\bar{X} = np$, $\bar{X} = \frac{\sum(rN_r)}{N}$ y $\lambda = np$

4) **calcular la media, varianza, desviación estándar**: $\bar{X} = \lambda = np$; $s = \sqrt{\lambda}$; $s^2 = \lambda$

5) **comprobar el ajuste** de unos datos a una DP

Veremos un ejemplo para comprobar el ajuste de una distribución real a una DP teórica.

Sabemos que a partir de los datos que nos den hay que calcular el parámetro λ . Luego se calculan las p teóricas asociadas a cada una de las modalidades deseadas y se multiplican por N, obteniendo de esta forma las N_r teóricas, que hay que contrastar con las observadas mediante la prueba estadística correspondiente.

--El veterinario militar alemán Borotkiewitz estudió las defunciones por coces de caballo en 20 regimientos prusianos durante 10 años ("Ley de los pequeños números", 1898). Encontró que seguían la distribución de los sucesos raros de Poisson y que por tanto eran fruto del azar y no eran imputables en principio a fallos de organización.

De los 200 regimientos-año (20*10) hubo 109 que no registraron muertes, 65 con un fallecimiento, 22 con dos, 3 con tres y 1 con cuatro.

Como λ es igual a la media, se utiliza la fórmula ya conocida $\bar{X} = \frac{\sum(rN_r)}{N}$

r	N_r
0	109
1	65
2	22
3	3
4	1
Σ	200

$$\bar{X} = \frac{(0*109)+(1*65)+(2*22)+(3*3)+(4*1)}{200} = 0,61$$

Hay que desarrollar una P(0,61) con N=200

r	p(r)	N_r
0	0,543	109
1	0,331	66
2	0,101	20
3	0,021	4
4	0,003	1
Σ		200

Los valores de N_r se presentan redondeados para que se vea mejor a simple vista la comparación con los observados. Para el contraste con las frecuencias observadas habría que dejar dos o tres decimales (esto es válido para cualquier ajuste). La prueba da $z=0,465$ que no es significativa. Por tanto el ajuste de esos datos a una DP es bueno

DISTRIBUCION NORMAL

Es la distribución típica de variables aleatorias cuantitativas continuas cuando el tamaño es grande (por consenso, cuando $N \geq 30$). Sus parámetros básicos son la media y la desviación estándar. Su desarrollo se debe fundamentalmente a Laplace y Gauss. Quetelet le dió el nombre de normal o natural porque observó que la gran mayoría de variables fisiológicas seguían este modelo. Es un nombre consagrado por el uso y no quiere decir que las otras distribuciones sean "anormales". Los norteamericanos usan y han exportado la denominación de "distribución gaussiana". Siguen la DN todo tipo de variables biológicas (como frecuencia cardíaca, tensión arterial, componentes químicos de la sangre y orina, medidas corporales...), duración o vida de objetos y seres vivos, etc

Notación : $N(\bar{x}, s)$

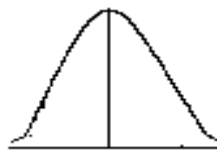
Fórmula

La fórmula para calcular las p asociadas a intervalos de valores (no se pueden calcular p de valores puntuales, ya que en el contexto de la DN son infinitésimos) es muy compleja y necesita integración. Pero afortunadamente no hay que utilizarla, pues se dispone de una tabla de fácil manejo, que nos da el cálculo ya hecho. A título informativo la fórmula es:

$$p(a \leq x \leq b) = \int_a^b f(x) dx, \text{ siendo } f(x) = \frac{1}{s\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\bar{x}}{s}\right)^2}$$

Representación gráfica

es la curva o campana de Gauss, en “chapeau de gendarme” (gorro de gendarme) de los tiempos napoleónicos. Es el límite de un histograma cuando la amplitud de las clase se hace infinitesimal y el n° de datos tiende a infinito.



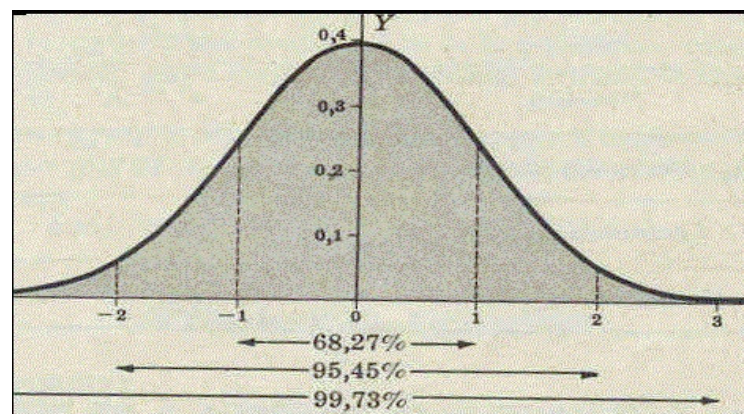
Es simétrica alrededor de un eje vertical que pasa por \bar{x} y asintótica al eje de abscisas (lo corta en el infinito por ambos lados, aunque a partir de $\bar{x} \pm 3s$ ya casi lo toca). La campana engloba todos los valores y por tanto la p de que un valor cualquiera esté en ella es 1 ó 100%. La superficie de campana delimitada por dos valores del eje de abscisas equivale a la probabilidad de que un valor cualquiera se encuentre en ese área. Cada distribución tiene su propia campana, hay infinitas curvas de DN. En estas condiciones su manejo sería muy difícil y complicado, ya que habría que aplicar cada vez la fórmula. Afortunadamente se ha encontrado un modelo único de distribución y por tanto de campana al que pueden ser adaptadas todas las DN. Es la llamada DN tipificada.

Tipificación

Consiste en transformar cualquier $N(\bar{x}, s)$ en otra $N(0, 1)$, es decir, en una DN de media 0 y desviación estándar 1. Para ello hay que transformar los valores originales x en puntuaciones estándar o valores tipificados, que aquí llamaremos c. (Otros nombres: z o SDS).

$$c = \frac{x - \bar{x}}{s}$$

Entre dos valores de c quedan delimitadas áreas (=probabilidad) que se pueden obtener a partir de la tabla de la DN tipificada. Ya se ha dicho al principio que no se pueden calcular p de valores aislados, sólo de intervalos más o menos grandes.



En esta campana están representadas las áreas o probabilidades entre valores de $c + 1$ y -1 , $+2$ y -2 , $+3$ y -3 . Pero es preferible expresar la p con números más “redondos”:

---Al intervalo entre $c = -1,96$ y $c = 1,96$ corresponde un 95% de la superficie de la campana.

$$p(-1,96 \leq c \leq 1,96) = 0,95 \text{ ó } 95\%$$

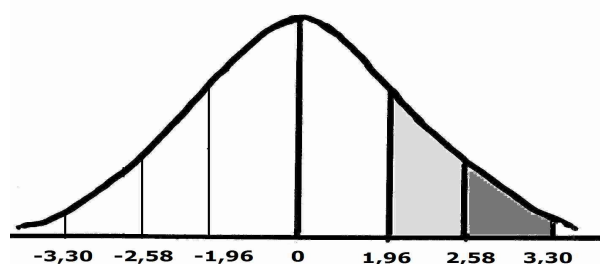
---Al intervalo entre $c = -2,58$ y $c = 2,58$ corresponde un 99% de la superficie de la campana.

$$p(-2,58 \leq c \leq 2,58) = 0,99 \text{ ó } 99\%$$

---Al intervalo entre $c = -3,30$ y $c = 3,30$ corresponde un 99,9% de la superficie de la campana.

$$p(-3,30 \leq c \leq 3,30) = 0,999 \text{ ó } 99,9\%$$

que son los que utilizaremos aquí.



Es imprescindible dibujar una campana y marcar en ella la media y el valor o valores de x .

Una vez tipificada se anotan el los valores de c . A la media le corresponde siempre por definición el valor de 0.

Tabla de la DN tipificada

El modelo que utilizamos es de media campana, va de 0 a $+\infty$. (Página 16). Hay otro con la campana entera, que abarca de $-\infty$ a $+\infty$. Nos da la p de que un valor cualquiera esté entre $c = 0$ y otro valor de c . Al ser la campana simétrica sirve por igual para valores de c positivos o negativos, siempre con dos decimales. Es una tabla de doble entrada. En la primera columna están valores de c con un decimal y en la primera fila está el segundo decimal. Donde confluyen ambos está la probabilidad buscada.

Problemas asociados a la DN

1---tipificar

p.e. $x=5$ y $x=3$ de una $B(4, 2)$

$$\rightarrow c = (5-4)/2 = 0,5 \quad \rightarrow c = (3-4)/2 = -0,5$$

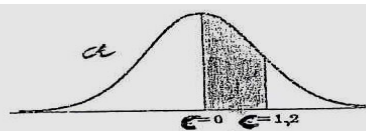
2---calcular la probabilidad de un intervalo,

p.e. entre $c = 0$ y $c = 0,46$

$$\rightarrow p(0 \leq c \leq 0,46) = 0,1772$$

CASOS POSIBLES

a) $p(0 \leq c \leq 1,2) = 0,3849 \text{ ó } 38,5\%$



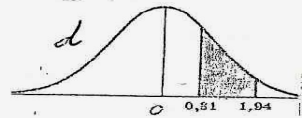
b) $p(-0,68 \leq c \leq 0) = 0,2518 \text{ ó } 25,2\%$



c) $p(-0,46 \leq c \leq 2,21) = 0,6636 \text{ ó } 66,4\%$
=área entre $-0,46$ y 0 más área entre 0 y $2,21$

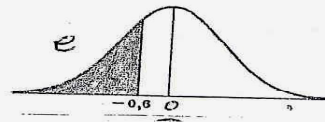


d) $p(0,81 \leq c \leq 1,94) = 0,1828 \text{ ó } 18,3\%$



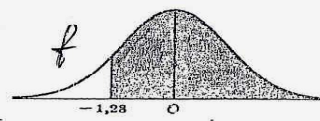
= área para $c=1,94$ menos
área para $c=0,81$

e) $p(c \leq -0,6) = 0,2742 \text{ ó } 27,4\%$



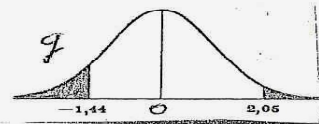
= 0,5 menos área para $c=-0,6$

f) $p(c \geq -1,28) = 0,8997 \text{ ó } 90\%$



= 0,5 más área para $c=-1,28$

g) $p(c \leq 1,44 \text{ y } c \geq 2,05) = 0,0951 \text{ ó } 9,5\%$



= $1 - \text{área para } c=-1,44 \text{ y } c=2,05$

Ejemplo:

La duración media de una bombilla es de 12 meses, con una varianza de 4. El fabricante garantiza que dura más de 8 meses. Calcular

- 1) la probabilidad de que se funda en el periodo de garantía
- 2) la probabilidad de que dure al menos 16 meses
- 3) la probabilidad de que dure entre 15 y 18 meses

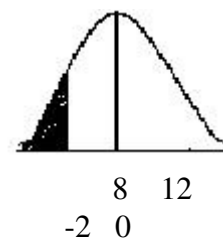
La variable "Vida de la bombilla" es una $N(12, 2)$

1) $p(x \leq 8) ?$

se dibuja la campana

se tipifica: $c = (8-12)/2 = -2$

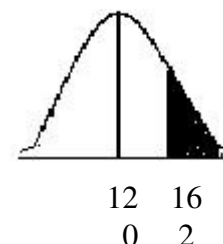
$p(c \leq -2) = 0,5 - p(-2 \leq c \leq 0) =$
 $0,5 - 0,4772 = 0,0228 \text{ ó } 2,28\%$



2) $p(x \geq 16) ?$

$c = (16-12)/2 = 2$

$p(c \geq 2) = 0,5 - p(0 \leq c \leq 2) =$
 $0,5 - 0,4772 = 0,0228 \text{ ó } 2,28\%$

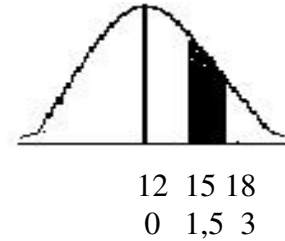


3) $p(15 \leq x \leq 18)$?

$$c_1 = (15-12)/2 = 1,5$$

$$c_2 = (18-12)/2 = 3$$

$$\begin{aligned} p(1,5 \leq c \leq 3) &= p(0 \leq c \leq 3) - p(0 \leq c \leq 1,5) \\ &= 0,4987 - 0,4332 = 0,0655 \text{ ó } 6,55\% \end{aligned}$$



3)---**calcular la frecuencia de un intervalo**, conocidos N y la p del intervalo.

Es similar a lo visto en la DB: $N_i = N * p$. Aquí para simplificar llamaremos al intervalo i (en vez de $a \leq x \leq b$ ó $x \in (a,b)$) y a su frecuencia N_i .

Supongamos que en una muestra de 6500 individuos en los que se hecho el análisis A hemos calculado una p de 0,2426 para el intervalo entre 7 y 10 mg/dl. ¿Cuántos individuos tendrán ese análisis entre 7 y 10 mg/dl?

Solución: $N_i = 6500 * 0,2426 = 1576,9 \approx 1577$

4)---**Calcular un valor de c a partir de una p y de un punto de referencia en la campana** (es decir, de otro valor de c)

Como en todos los problemas de campana es imprescindible dibujarla y situar en ella el punto c de referencia.

No olvidar que los de signo positivo se ponen a la derecha de la media (según vemos la campana) y los negativos a la izquierda.

Luego se busca en la tabla la p que nos dan y se ve a que valor de c corresponde. No olvidar el signo menos si le corresponde estar a la izquierda. Si el valor de p no está exactamente se toma el más próximo, siguiendo el mismo procedimiento que en el redondeo.

CALCULO DE UN VALOR c A PARTIR DE UNA PROBABILIDAD Y UN PUNTO DE REFERENCIA.

1) el área entre 0 y c es de 0.3770 ; $p \in (0 \div c) = 0.3770$

- dibujar campana



- buscar en la tabla. Vemos que le corresponde un c de 1.16

respuestas: hay dos $c = 1.16$ y $c = -1.16$

2) el área a la izquierda de c es 0.8621 ; $0.8621 = p \in (-\infty \div c)$

- dibujar la campana; al ser $p > 0.5$ c tiene que estar en el lado derecho

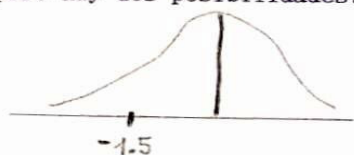


- como nuestra tabla es de sólo media campana restamos 0.5 $p = 0.3621$

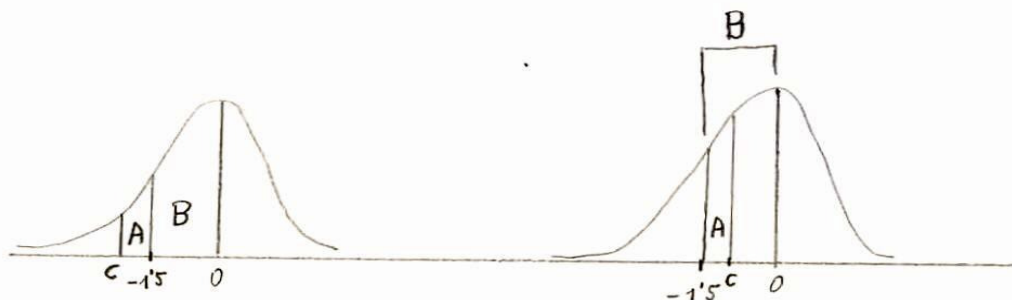
- buscamos en la tabla y encontramos una c de 1.09
respuesta: 1.09

3) el área entre -1.5 y c es de 0.0217 ; $0.0217 = p \in (-1.5 \div c)$

- dibujar campana; c tiene que estar por fuerza a la izquierda ya que si fuera + el área valdría más de 0.4332 que es la p que corresponde a $c = -1.5$ pero hay dos posibilidades: a la derecha y a la izquierda de -1.5



- $p = A+B = 0.0217+0.4332 = 0.4549$; le corresponde $c \approx -1.69$
- $p = B-A = 0.4332-0.0217 = 0.4115$; le corresponde $c \approx -1.35$



5)---Calcular una puntuación original, x , a partir de puntuaciones estándar c

Se utiliza la fórmula $c = \frac{x - \bar{x}}{s}$; puede ser necesario dibujar la campana si hay alguna duda.

Ejemplos:

a)---Calcular la puntuación original que corresponde a una $c = 1,6$ en una $N(6, 2)$

$$\rightarrow 1,6 = (x-6)/2 ; x = 9,2$$

b)---En esa misma distribución calcular la puntuación original que deja por debajo de ella el 86,21% de los valores.

\rightarrow 86,21% equivale a una p de 0,8621 , por lo que x tiene que estar situado en el lado derecho de la campana. Para poder utilizar la tabla le restamos 0,5 a 0,8621 y queda 0,3621 . Le corresponde una $c = 1,09$. Entonces $1,09 = (x-6)/2$; $x = 8,18$

6)---Calcular \bar{x} y s a partir de otros parámetros.

Se utiliza la misma fórmula: $c = \frac{x - \bar{x}}{s}$.

De sus 4 elementos hay que conocer 3. Puede ser conveniente dibujar la campana.

Ejemplo: Calcular la s de un DN cuya media es 5 y en la que $p(x \leq 6) = 0,6064$

\rightarrow x tiene que estar en el lado derecho de la campana al ser la $p > 0,5$

$0,6064 - 0,5 = 0,1064$ a quien corresponde una c de 0,27 .

$$0,27 = (6-5)/s \quad \text{y} \quad s = 3,70$$

7)---aproximar una DB o una DP a una DN

Ambas se aproximan de forma perfecta a la DN cuando np ó $\lambda \rightarrow \infty$.

Las condiciones para la aproximación de la DN de una DB, recordemos, son p y $q \geq 0,1$ (ó 10%) y np y $nq \geq 5$ (ó 500, si p se expresa como %).

La DB se transforma en una DN, que tenga la misma media y desviación estándar que la DB

La DP se aproxima de forma similar.

Hay que hacer una pequeña corrección, la llamada **corrección de continuidad**. La DB es discreta y por tanto discontinua y la DN es continua. No se toman los límites tabulados del intervalo sino el límite real que corresponda. Los límites tabulados deben quedar incluidos, por lo que en unos casos se tomará el límite real inferior y en otros el superior.

Así, si tiramos 300 monedas y queremos saber la p de obtener entre 90 y 120 caras, no calcularemos $p(90 \leq x \leq 120)$ sino $p(89,5 \leq x \leq 120,5)$.

Ejemplo: Esta misma tirada de las 300 monedas. Es una $B(300, 0,5)$. $\bar{x} = 300 * 0,5 = 150$

$s = \sqrt{npq} = 8,66$. Por tanto la transformamos en un $N(150, 8,66)$, en la que hay que calcular $p(89,5 \leq x \leq 120,5)$ por el procedimiento ya visto. (Es como el caso 2d, pero en el lado izquierdo de la campana. El resultado es 0,0003)

8)---Comprobar el ajuste de una distribución real (observada) a una DN.

Lo veremos con la distribución de la talla de sus compañeros del curso 1978/79.

$N = 47$ $\bar{x} = 167,9$ cm $s = 7,8$ cm

Talla de los alumnos de Bioestadística Curso 1978/79		
clases	p.m.	nº
152-161 cm	156,5	10
162-171 cm	166,5	23
172-181 cm	176,5	12
182-191 cm	186,5	2

Hay que construir una DN teórica que tenga los mismos parámetros que la real. Una vez conocidas las frecuencias teóricas de cada clase se contrastan con las reales, mediante la prueba correspondiente. Si no hay diferencias significativas, el ajuste es bueno.

El procedimiento es un tanto engorroso y conviene seguir una metodología clara para no equivocarse. Como la que se usa aquí.

Pasos:

- 1) construirse una tabla auxiliar
- 2) comenzar a rellenarla por los Límites Reales

clases	L. reales	c	área entre c y 0	p de la clase	Ni teórico \approx		Ni real
	$-\infty$	$-\infty$					
	151,5						
	161,5						
	171,5						
	181,5						
	191,5						
	$+\infty$	$+\infty$					

3) situar las clases

clases	L. reales	c	área Entre c y 0	p de la clase	Ni teórico \approx	Ni real
	$-\infty$					

	151,5					
152-161						
	161,5					
162-171						
	171,5					
172-181						
	181,5					
182-191						
	191,5					

	$+\infty$					

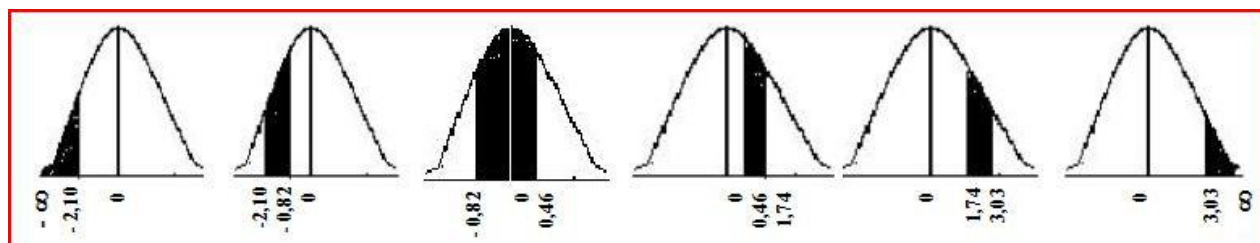
4) Calcular valores de c para cada L. real y el área entre c y 0

clases	L. reales	c	área entre c y 0	p de la clase	Ni teórico	Ni real
	$-\infty$	$-\infty$	0,5			

	151,5	-2,10	0,4821			
152-161						
	161,5	-0,82	0,2939			
162-171						
	171,5	0,46	0,1772			
172-181						
	181,5	1,74	0,4591			
182-191						
	191,5	3,03	0,4988			

	$+\infty$	$+\infty$	0,5			

5) calcular la p de cada clase (dibujar campana), pasarla a la tabla auxiliar y calcular Nr teórico



clases	L. reales	c	Área (p) entre c y 0	p de la clase	Ni teórico ≈		Ni real
	- ∞	- ∞	0,5				
-----				0,0179	0,9		--
152-161	151,5	-2,10	0,4821				
	161,5	-0,82	0,2939	0,1882	9		10
162-171	171,5	0,46	0,1772	0,4711	22		23
	181,5	1,74	0,4591	0,2819	13		12
182-191	191,5	3,03	0,4988	0,0397	2		2
-----				0,0012	0,1		--
	+ ∞	+ ∞	0,5				

6) aplicar prueba de contraste de frecuencias (fórmula nº 3; tema 16). Se obtiene $Z=1,233$, que es $< \chi^2(5, 0,05)=11,07$, n.s. Se concluye que el ajuste es bueno, como parece ya a simple vista.

Distribución de la t de Student

es la distribución teórica de las muestras pequeñas de una población que sigue la ley normal con datos cuantitativos continuos.

Gosset (que utilizaba el seudónimo de Student) comprobó que cuando disminuía el tamaño de las muestras, no valían del todo los normas de la DN, tanto más cuanto más pequeña sea la muestra. Hasta $N=30$ las diferencias son bastante acusadas. Por eso la mayoría de autores ponen a ese nivel la frontera de uso práctico entre DN y t de Student.. Otros lo ponen en 60 y algunos hasta en 120. Los programas estadísticos utilizan casi exclusivamente la t de Student para todas las variables continuas, ya que hasta el infinito no se produce una identidad plena entre ambas distribuciones. La DN está en vías de extinción, al menos en la práctica. Nosotros seguiremos el criterio de utilizar la t de Student para muestras pequeñas ($N<30$) y la DN para las grandes.

La **notación** es $t(gl, \alpha)$. α es el nivel de significación elegido y gl es el grado de libertad. Con este nombre se designa al número de observaciones independientes, que en general son $N-1$. Un ejemplo ayudará a entender este concepto. Si nos piden 5 valores que sumen 35, sólo podremos elegir libremente 4, pues el 5º es obligado: supongamos que elegimos 8 , 10 , 23 , -15 . El 5º número tiene que ser por fuerza 9 ; hay 4 grados de libertad.

Aquí no hay modelo tipificado y para cada valor de N hay una campana distinta (que no es preciso dibujar..).

La **tabla** sigue el modelo de las tablas de doble entrada. En la primera columna está el grado de libertad y en la primera fila hay tres niveles de significación.

$t(5, 0,05) = 2,571$; $t(26, 0,001) = 3,707$; $t(15, 0,01) = 2,947$

El término t se usa para designar varias cosas, lo que puede generar cierta confusión:

1---la distribución de la t de Student

2---los valores de la abscisa de la campana correspondiente, donde están los valores de referencia para valorar el resultado de las pruebas. Es el equivalente a la c de la DN

3---el resultado de las pruebas estadísticas que son valoradas por la t de Student. Esto lo obviamos llamando de una forma genérica **Z** a todos los resultados de las pruebas estadísticas, nombre arbitrario que puede ser sustituido por cualquier otro.

Distribución χ^2 (chi o ji cuadrado)

es la distribución que siguen las frecuencias de muestras obtenidas de una población.

También aquí hay grados de libertad y para cada grado de libertad hay un gráfico distinto.

Notación: χ^2 (gl, α)

La **tabla** es también de doble entrada, con una disposición similar, aunque nos ofrece un nivel de significación más, el de 0,02.

$\chi^2(1, 0,05) = 3,84$; $\chi^2(2, 0,01) = 9,21$; $\chi^2(5, 0,001) = 20,52$

Su uso es típico de las tablas de 2 por 2 (2×2) ó f por k ($f \times k$), siendo f el nº de filas y k el de columnas.

Con el nombre de χ^2 se pueden designar también dos cosas:

1---la distribución χ^2

2---los resultados de las pruebas que son valoradas por la χ^2 (lo que no seguimos aquí, pues a todos los resultados los llamamos **Z**, con independencia de cómo sean valorados).

Distribución de la F de Snedecor-Fisher

es la distribución de los posibles cocientes de dos varianzas, poniendo siempre la mayor de ellas en el numerador. Así F será siempre ≥ 1 , lo que supone un ahorro de espacio al confeccionar la tabla. Aquí también hay grados de libertad y gráficos distintos para cada grado de libertad (que no tenemos que dibujar).

Notación : $F(gl1, gl2, \alpha)$. Siendo $gl1 = k-1$ (k es el nº de muestras o grupos) y $gl2 = (N-1)(k-1)$. N es la frecuencia total, el tamaño total de todas las muestras o grupos .

Tablas: para cada nivel de significación hay una tabla distinta, que también es de doble entrada.

Se busca $gl1$ en la primera fila y $gl2$ en la primera columna.

$F(5, 9, 0,05) = 3,48$; $F(12, 10, 0,01) = 4,71$

Cuando la tabla no nos ofrece el valor exacto del gl , se aproxima al más cercano o si se es muy riguroso, siempre al inferior. Para $F(90, 30, 0,001)$ lo habitual es elegir 2,76, pero en función del rigor de la investigación se puede elegir también 2,92

Se usa para valorar la llamada “igualdad de varianzas” y los resultados de las pruebas de ANOVA.

Con F se pueden designar también dos cosas:

1---la distribución F

2---los resultados de las pruebas que son valoradas por la F (lo que no seguimos aquí, pues a todos los resultados los llamamos **Z**, con independencia de cómo sean valorados).

DISTRIBUCIÓN HIPERGEOMETRICA

Variante de Binomial cuando no hay reposición de efectivos y N es finita. Si N es muy grande, vale la Binomial. (La aproximación es ya buena, si $N_1/N \leq 0,1$ ó mejor si $\leq 0,05$). O sea, siempre que el tamaño de la muestra sea el 10% o menos del tamaño de la población, se puede usar –y de hecho se usa- la DB.

Notación: $H(n, N, N_1)$, siendo n como en la DB, N el n° total de individuos y N_1 los que presentan la característica. Se busca la p de r (que va de 0 a n, como en la DB).

$$\text{Fórmula: } p(r) = \frac{\binom{N_1}{r} \binom{N-N_1}{n-r}}{\binom{N}{n}} = \frac{\frac{N_1!}{r!(N_1-r)!} * \frac{(N-N_1)!}{(n-r)!(N-N_1-n+r)!}}{\frac{N!}{n!(N-n)!}}$$

$$\text{ó simplificando : } p(r) = \frac{N_1!(N-N_1)!n!((N-n)!}{r!(N_1-r)!(n-r)!(N-N_1-n+r)!N!}$$

$$\text{La varianza es menor que en la DB: } s^2 = npq \frac{N-n}{N-1}$$

Al intervenir tantas factoriales en la fórmula, las calculadoras e incluso muchos programas estadísticos de ordenador se ven sobrepasados fácilmente en su capacidad de cálculo. La hoja de cálculo Excel admite hasta $N = 170$, mientras otros programas más antiguos, basados en MS-Dos, no pasan de 33. Lo vemos aquí para completar el tema, ya que por este motivo no puede ser objeto de examen. En la práctica es habitual hacer los cálculos como si fuera una DB, ya que el error es en general muy pequeño.

Ejemplo 1: De 100 enfermos, 20 presentan una infección. Se toman 5 al azar y se pide la probabilidad de que sólo 1 presente la infección.

$N=100$; $N_1=20$; $n=5$; $r=1$ Es $H(5, 100, 20)$

Haciendo las operaciones sale $p(r=1) = 0,420144...$

Como binomial sería $B(5, 0.2)$ y $p(r=1)=0,4096$

Ejemplo 2:

p de que sacando 4 cartas de una baraja española de 40 cartas, las 4 sean ases.

Es $H(4, 40, 4)$ y $p(r=4) = 1,0942 \cdot 10^{-5}$

Como $B(4, 4/40) = B(4, 0.1)$, $p(r=4) = 0,0001$

Por cálculo elemental (que es exacto) : $4/40 * 3/39 * 2/38 * 1/37 = 24/2193369 = 1,0942 \cdot 10^{-5}$

Por Poisson , $P(0,4)$: $p(r=4) = 0,0007$

Distribución Binomial B(n , p)

$$\bar{X} = np \quad s = \sqrt{npq} \quad N = \sum N_r \quad N_r = Np(r)$$

$$\bar{X} = \frac{\sum (rN_r)}{N} \quad p(r) = \frac{n!}{r! * (n-r)!} p^r q^{(n-r)}$$

n	r	p								
		0,10	0,15	0,20	0,25	0,30	0,35	0,40	0,45	0,50
1	0	0,9000	0,8500	0,8000	0,7500	0,7000	0,6500	0,6000	0,5500	0,5000
	1	0,1000	0,1500	0,2000	0,2500	0,3000	0,3500	0,4000	0,4500	0,5000
2	0	0,8100	0,7225	0,6400	0,5625	0,4900	0,4225	0,3600	0,3025	0,2500
	1	0,1800	0,2550	0,3200	0,3750	0,4200	0,4550	0,4800	0,4950	0,5000
	2	0,0100	0,0225	0,0400	0,0625	0,0900	0,1225	0,1600	0,2025	0,2500
3	0	0,7290	0,6141	0,5120	0,4219	0,3430	0,2746	0,2160	0,1664	0,1250
	1	0,2430	0,3251	0,3840	0,4219	0,4410	0,4436	0,4320	0,4084	0,3750
	2	0,0270	0,0574	0,0960	0,1406	0,1890	0,2389	0,2880	0,3341	0,3750
	3	0,0010	0,0034	0,0080	0,0156	0,0270	0,0429	0,0640	0,0911	0,1250
4	0	0,6561	0,5220	0,4096	0,3164	0,2401	0,1785	0,1296	0,0915	0,0625
	1	0,2916	0,3685	0,4096	0,4219	0,4116	0,3845	0,3456	0,2995	0,2500
	2	0,0486	0,0975	0,1536	0,2109	0,2646	0,3105	0,3456	0,3675	0,3750
	3	0,0036	0,0115	0,0256	0,0469	0,0756	0,1115	0,1536	0,2005	0,2500
	4	0,0001	0,0005	0,0016	0,0039	0,0081	0,0150	0,0256	0,0410	0,0625
5	0	0,5905	0,4437	0,3277	0,2373	0,1681	0,1160	0,0778	0,0503	0,0313
	1	0,3281	0,3915	0,4096	0,3955	0,3602	0,3124	0,2592	0,2059	0,1563
	2	0,0729	0,1382	0,2048	0,2637	0,3087	0,3364	0,3456	0,3369	0,3125
	3	0,0081	0,0244	0,0512	0,0879	0,1323	0,1811	0,2304	0,2757	0,3125
	4	0,0005	0,0022	0,0064	0,0146	0,0284	0,0488	0,0768	0,1128	0,1563
	5	0,0000	0,0001	0,0003	0,0010	0,0024	0,0053	0,0102	0,0185	0,0313
6	0	0,5314	0,3771	0,2621	0,1780	0,1176	0,0754	0,0467	0,0277	0,0156
	1	0,3543	0,3993	0,3932	0,3560	0,3025	0,2437	0,1866	0,1359	0,0938
	2	0,0984	0,1762	0,2458	0,2966	0,3241	0,3280	0,3110	0,2780	0,2344
	3	0,0146	0,0415	0,0819	0,1318	0,1852	0,2355	0,2765	0,3032	0,3125
	4	0,0012	0,0055	0,0154	0,0330	0,0595	0,0951	0,1382	0,1861	0,2344
	5	0,0001	0,0004	0,0015	0,0044	0,0102	0,0205	0,0369	0,0609	0,0938
	6	0,0000	0,0000	0,0001	0,0002	0,0007	0,0018	0,0041	0,0083	0,0156
7	0	0,4783	0,3206	0,2097	0,1335	0,0824	0,0490	0,0280	0,0152	0,0078
	1	0,3720	0,3960	0,3670	0,3115	0,2471	0,1848	0,1306	0,0872	0,0547
	2	0,1240	0,2097	0,2753	0,3115	0,3177	0,2985	0,2613	0,2140	0,1641
	3	0,0230	0,0617	0,1147	0,1730	0,2269	0,2679	0,2903	0,2918	0,2734
	4	0,0026	0,0109	0,0287	0,0577	0,0972	0,1442	0,1935	0,2388	0,2734
	5	0,0002	0,0012	0,0043	0,0115	0,0250	0,0466	0,0774	0,1172	0,1641
	6	0,0000	0,0001	0,0004	0,0013	0,0036	0,0084	0,0172	0,0320	0,0547
	7	0,0000	0,0000	0,0000	0,0001	0,0002	0,0006	0,0016	0,0037	0,0078
8	0	0,4305	0,2725	0,1678	0,1001	0,0576	0,0319	0,0168	0,0084	0,0039
	1	0,3826	0,3847	0,3355	0,2670	0,1977	0,1373	0,0896	0,0548	0,0313
	2	0,1488	0,2376	0,2936	0,3115	0,2965	0,2587	0,2090	0,1569	0,1094
	3	0,0331	0,0839	0,1468	0,2076	0,2541	0,2786	0,2787	0,2568	0,2188
	4	0,0046	0,0185	0,0459	0,0865	0,1361	0,1875	0,2322	0,2627	0,2734
	5	0,0004	0,0026	0,0092	0,0231	0,0467	0,0808	0,1239	0,1719	0,2188
	6	0,0000	0,0002	0,0011	0,0038	0,0100	0,0217	0,0413	0,0703	0,1094
	7	0,0000	0,0000	0,0001	0,0004	0,0012	0,0033	0,0079	0,0164	0,0313
	8	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002	0,0007	0,0017	0,0039

Distribución de Poisson $P(\lambda)$

$$p(r) = \frac{\lambda^r}{r!} e^{-\lambda}$$

$$\bar{X} = \lambda = np \quad \bar{X} = \frac{\sum (rN_r)}{N} \quad s^2 = \lambda \quad s = \sqrt{\lambda}$$

Valores de $e^{-\lambda}$

(De *Statistics*, por M. R. SPIEGEL. Schaum Publishing Company. Nueva York, 1961.)

λ	0	1	2	3	4	5	6	7	8	9
0,0	1,0000	0,9900	0,9802	0,9704	0,9608	0,9512	0,9418	0,9324	0,9231	0,9139
0,1	0,9048	0,8958	0,8869	0,8781	0,8694	0,8607	0,8521	0,8437	0,8353	0,8270
0,2	0,8187	0,8106	0,8025	0,7945	0,7866	0,7788	0,7711	0,7634	0,7558	0,7483
0,3	0,7408	0,7334	0,7261	0,7189	0,7118	0,7047	0,6977	0,6907	0,6839	0,6771
0,4	0,6703	0,6636	0,6570	0,6505	0,6440	0,6376	0,6313	0,6250	0,6188	0,6126
0,5	0,6065	0,6005	0,5945	0,5886	0,5827	0,5770	0,5712	0,5655	0,5599	0,5543
0,6	0,5488	0,5434	0,5379	0,5326	0,5273	0,5220	0,5169	0,5117	0,5066	0,5016
0,7	0,4966	0,4916	0,4868	0,4819	0,4771	0,4724	0,4677	0,4630	0,4584	0,4538
0,8	0,4493	0,4449	0,4404	0,4360	0,4317	0,4274	0,4232	0,4190	0,4148	0,4107
0,9	0,4066	0,4025	0,3985	0,3946	0,3906	0,3867	0,3829	0,3791	0,3753	0,3716

($\lambda = 1, 2, 3, \dots, 10$)

λ	1	2	3	4	5	6	7	8	9	10
$e^{-\lambda}$	0,36788	0,13534	0,04979	0,01832	0,006738	0,002479	0,000912	0,000335	0,000123	0,000045

NOTA. Para obtener valores de $e^{-\lambda}$ para otros valores de λ basta tener en cuenta las reglas del producto de potencias, por ejemplo:

$$e^{-3,48} = e^{-3,00} \cdot e^{-0,48} = 0,04979 \cdot 0,6188 = 0,03081.$$

ejemplos:

$$e^{-0,28} = 0,7558$$

$$e^{-0,95} = 0,3867$$

$$e^{-0,50} = 0,6065$$

$$e^{-5} = 0,001832$$

$$e^{-3,48} = e^{-3} \cdot e^{-0,48} = 0,04979 \cdot 0,6188 = 0,03081$$

Distribución normal N (0 , 1)

$$c = \frac{x - \bar{x}}{s}$$



la tabla da la probabilidad de que un valor cualquiera esté entre $c = 0$ y otro valor de c

a esta c se la llama hoy día mayoritariamente **z**

c	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,0000	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239	0,0279	0,0319	0,0359
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675	0,0714	0,0753
0,2	0,0793	0,0832	0,0871	0,0910	0,0948	0,0987	0,1026	0,1064	0,1103	0,1141
0,3	0,1179	0,1217	0,1255	0,1293	0,1331	0,1368	0,1406	0,1443	0,1480	0,1517
0,4	0,1554	0,1591	0,1628	0,1664	0,1700	0,1736	0,1772	0,1808	0,1844	0,1879
0,5	0,1915	0,1950	0,1985	0,2019	0,2054	0,2088	0,2123	0,2157	0,2190	0,2224
0,6	0,2257	0,2291	0,2324	0,2357	0,2389	0,2422	0,2454	0,2486	0,2517	0,2549
0,7	0,2580	0,2611	0,2642	0,2673	0,2704	0,2734	0,2764	0,2794	0,2823	0,2852
0,8	0,2881	0,2910	0,2939	0,2967	0,2995	0,3023	0,3051	0,3078	0,3106	0,3133
0,9	0,3159	0,3186	0,3212	0,3238	0,3264	0,3289	0,3315	0,3340	0,3365	0,3389
1,0	0,3413	0,3438	0,3461	0,3485	0,3508	0,3531	0,3554	0,3577	0,3599	0,3621
1,1	0,3643	0,3665	0,3686	0,3708	0,3729	0,3749	0,3770	0,3790	0,3810	0,3830
1,2	0,3849	0,3869	0,3888	0,3907	0,3925	0,3944	0,3962	0,3980	0,3997	0,4015
1,3	0,4032	0,4049	0,4066	0,4082	0,4099	0,4115	0,4131	0,4147	0,4162	0,4177
1,4	0,4192	0,4207	0,4222	0,4236	0,4251	0,4265	0,4279	0,4292	0,4306	0,4319
1,5	0,4332	0,4345	0,4357	0,4370	0,4382	0,4394	0,4406	0,4418	0,4429	0,4441
1,6	0,4452	0,4463	0,4474	0,4484	0,4495	0,4505	0,4515	0,4525	0,4535	0,4545
1,7	0,4554	0,4564	0,4573	0,4582	0,4591	0,4599	0,4608	0,4616	0,4625	0,4633
1,8	0,4641	0,4649	0,4656	0,4664	0,4671	0,4678	0,4686	0,4693	0,4699	0,4706
1,9	0,4713	0,4719	0,4726	0,4732	0,4738	0,4744	0,4750	0,4756	0,4761	0,4767
2,0	0,4772	0,4778	0,4783	0,4788	0,4793	0,4798	0,4803	0,4808	0,4812	0,4817
2,1	0,4821	0,4826	0,4830	0,4834	0,4838	0,4842	0,4846	0,4850	0,4854	0,4857
2,2	0,4861	0,4864	0,4868	0,4871	0,4875	0,4878	0,4881	0,4884	0,4887	0,4890
2,3	0,4893	0,4896	0,4898	0,4901	0,4904	0,4906	0,4909	0,4911	0,4913	0,4916
2,4	0,4918	0,4920	0,4922	0,4925	0,4927	0,4929	0,4931	0,4932	0,4934	0,4936
2,5	0,4938	0,4940	0,4941	0,4943	0,4945	0,4946	0,4948	0,4949	0,4951	0,4952
2,6	0,4953	0,4955	0,4956	0,4957	0,4959	0,4960	0,4961	0,4962	0,4963	0,4964
2,7	0,4965	0,4966	0,4967	0,4968	0,4969	0,4970	0,4971	0,4972	0,4973	0,4974
2,8	0,4974	0,4975	0,4976	0,4977	0,4977	0,4978	0,4979	0,4979	0,4980	0,4981
2,9	0,4981	0,4982	0,4982	0,4983	0,4984	0,4984	0,4985	0,4985	0,4986	0,4986
3,0	0,4987	0,4987	0,4987	0,4988	0,4988	0,4989	0,4989	0,4989	0,4990	0,4990
3,1	0,4990	0,4991	0,4991	0,4991	0,4992	0,4992	0,4992	0,4992	0,4993	0,4993
3,2	0,4993	0,4993	0,4994	0,4994	0,4994	0,4994	0,4994	0,4995	0,4995	0,4995
3,3	0,4995	0,4995	0,4995	0,4996	0,4996	0,4996	0,4996	0,4996	0,4996	0,4997
3,4	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4998
3,5	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998
3,6	0,4998	0,4998	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999
3,7	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999
3,8	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999
3,9	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000

Tabla de χ^2

p

g. l.	0,05	0,02	0,01	0,001
1	3,84	5,41	6,64	10,83
2	5,99	7,82	9,21	13,82
3	7,81	9,84	11,34	16,27
4	9,49	11,77	13,28	18,47
5	11,07	13,39	15,09	20,52
6	12,59	15,03	16,81	22,46
7	14,07	16,62	18,48	24,32
8	15,51	18,17	20,09	26,13
9	16,92	19,68	21,67	27,88
10	18,31	21,16	23,21	29,59

Tabla de la t de Student

p

p

g. l.	0,05	0,01	0,001		g. l.	0,05	0,01	0,001
1	12,71	63,66	636,6		26	2,056	2,779	3,707
2	4,303	9,925	31,60		27	2,052	2,771	3,690
3	3,182	5,841	12,94		28	2,048	2,763	3,674
4	2,776	4,604	8,610		29	2,045	2,756	3,659
5	2,571	4,032	6,859		30	2,042	2,750	3,646
6	2,447	3,707	5,959		35	2,030	2,724	3,592
7	2,365	3,499	5,405		40	2,021	2,704	3,551
8	2,306	3,355	5,041		45	2,014	2,689	3,521
9	2,262	3,250	4,781		50	2,008	2,678	3,496
10	2,228	3,169	4,587		55	2,004	2,669	3,476
11	2,201	3,106	4,437		60	2,000	2,660	3,460
12	2,179	3,055	4,318		70	1,994	2,648	3,435
13	2,160	3,012	4,221		80	1,989	2,638	3,416
14	2,145	2,977	4,140		90	1,986	2,631	3,402
15	2,131	2,947	4,073		100	1,982	2,626	3,390
16	2,120	2,921	4,015		120	1,980	2,617	3,373
17	2,110	2,898	3,965		130	1,977	2,612	3,361
18	2,101	2,878	3,922		140	1,975	2,607	3,352
19	2,093	2,861	3,883		150	1,974	2,605	3,349
20	2,086	2,845	3,850		160	1,973	2,603	3,346
21	2,080	2,831	3,819		200	1,972	2,601	3,340
22	2,074	2,819	3,792		300	1,968	2,592	3,340
23	2,069	2,807	3,767		400	1,966	2,588	3,315
24	2,064	2,797	3,745		500	1,965	2,586	3,310
25	2,060	2,787	3,725		∞	1,960	2,576	3,291

F de Snedecor-Fisher $\alpha = 0,05$

g.l. 2 ↓	g.l. 1 →																						∞
	3	4	5	6	7	8	9	10	11	12	13	14	15	20	25	30	35	40	50	100	500		
5	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74	4,70	4,68	4,66	4,64	4,62	4,56	4,52	4,50	4,48	4,46	4,44	4,41	4,37	4,36	
6	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	4,03	4,00	3,98	3,96	3,94	3,87	3,83	3,81	3,79	3,77	3,75	3,71	3,68	3,67	
7	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64	3,60	3,57	3,55	3,53	3,51	3,44	3,40	3,38	3,36	3,34	3,32	3,27	3,24	3,23	
8	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35	3,31	3,28	3,26	3,24	3,22	3,15	3,11	3,08	3,06	3,04	3,02	2,97	2,94	2,93	
9	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14	3,10	3,07	3,05	3,03	3,01	2,94	2,89	2,86	2,84	2,83	2,80	2,76	2,72	2,71	
10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98	2,94	2,91	2,89	2,86	2,85	2,77	2,73	2,70	2,68	2,66	2,64	2,59	2,55	2,54	
11	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85	2,82	2,79	2,76	2,74	2,72	2,65	2,60	2,57	2,55	2,53	2,51	2,46	2,42	2,40	
12	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75	2,72	2,69	2,66	2,64	2,62	2,54	2,50	2,47	2,44	2,43	2,40	2,35	2,31	2,30	
13	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67	2,63	2,60	2,58	2,55	2,53	2,46	2,41	2,38	2,36	2,34	2,31	2,26	2,22	2,21	
14	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60	2,57	2,53	2,51	2,48	2,46	2,39	2,34	2,31	2,28	2,27	2,24	2,19	2,14	2,13	
15	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54	2,51	2,48	2,45	2,42	2,40	2,33	2,28	2,25	2,22	2,20	2,18	2,12	2,08	2,07	
16	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49	2,46	2,42	2,40	2,37	2,35	2,28	2,23	2,19	2,17	2,15	2,12	2,07	2,02	2,01	
17	3,20	2,96	2,81	2,70	2,61	2,55	2,49	2,45	2,41	2,38	2,35	2,33	2,31	2,23	2,18	2,15	2,12	2,10	2,08	2,02	1,97	1,96	
18	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41	2,37	2,34	2,31	2,29	2,27	2,19	2,14	2,11	2,08	2,06	2,04	1,98	1,93	1,92	
19	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38	2,34	2,31	2,28	2,26	2,23	2,16	2,11	2,07	2,05	2,03	2,00	1,94	1,89	1,88	
20	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35	2,31	2,28	2,25	2,22	2,20	2,12	2,07	2,04	2,01	1,99	1,97	1,91	1,86	1,84	
21	3,07	2,84	2,68	2,57	2,49	2,42	2,37	2,32	2,28	2,25	2,22	2,20	2,18	2,10	2,05	2,01	1,98	1,96	1,94	1,88	1,83	1,81	
22	3,05	2,82	2,66	2,55	2,46	2,40	2,34	2,30	2,26	2,23	2,20	2,17	2,15	2,07	2,02	1,98	1,96	1,94	1,91	1,85	1,80	1,78	
23	3,03	2,80	2,64	2,53	2,44	2,37	2,32	2,27	2,24	2,20	2,18	2,15	2,13	2,05	2,00	1,96	1,93	1,91	1,88	1,82	1,77	1,76	
24	3,01	2,78	2,62	2,51	2,42	2,36	2,30	2,25	2,22	2,18	2,15	2,13	2,11	2,03	1,97	1,94	1,91	1,89	1,86	1,80	1,75	1,73	
25	2,99	2,76	2,60	2,49	2,40	2,34	2,28	2,24	2,20	2,16	2,14	2,11	2,09	2,01	1,96	1,92	1,89	1,87	1,84	1,78	1,73	1,71	
26	2,98	2,74	2,59	2,47	2,39	2,32	2,27	2,22	2,18	2,15	2,12	2,09	2,07	1,99	1,94	1,90	1,87	1,85	1,82	1,76	1,71	1,69	
27	2,96	2,73	2,57	2,46	2,37	2,31	2,25	2,20	2,17	2,13	2,10	2,08	2,06	1,97	1,92	1,88	1,86	1,84	1,81	1,74	1,69	1,67	
28	2,95	2,71	2,56	2,45	2,36	2,29	2,24	2,19	2,15	2,12	2,09	2,06	2,04	1,96	1,91	1,87	1,84	1,82	1,79	1,73	1,67	1,65	
29	2,93	2,70	2,55	2,43	2,35	2,28	2,22	2,18	2,14	2,10	2,08	2,05	2,03	1,94	1,89	1,85	1,83	1,81	1,77	1,71	1,65	1,64	
30	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16	2,13	2,09	2,06	2,04	2,01	1,93	1,88	1,84	1,81	1,79	1,76	1,70	1,64	1,62	
40	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08	2,04	2,00	1,97	1,95	1,92	1,84	1,78	1,74	1,72	1,69	1,66	1,59	1,53	1,51	
50	2,79	2,56	2,40	2,29	2,20	2,13	2,07	2,03	1,99	1,95	1,92	1,89	1,87	1,78	1,73	1,69	1,66	1,63	1,60	1,52	1,46	1,44	
75	2,73	2,49	2,34	2,22	2,13	2,06	2,01	1,96	1,92	1,88	1,85	1,83	1,80	1,71	1,65	1,61	1,58	1,55	1,52	1,44	1,36	1,34	
100	2,70	2,46	2,31	2,19	2,10	2,03	1,97	1,93	1,89	1,85	1,82	1,79	1,77	1,68	1,62	1,57	1,54	1,52	1,48	1,39	1,31	1,29	
200	2,65	2,42	2,26	2,14	2,06	1,98	1,93	1,88	1,84	1,80	1,77	1,74	1,72	1,62	1,56	1,52	1,48	1,46	1,41	1,32	1,22	1,19	
1000	2,61	2,38	2,22	2,11	2,02	1,95	1,89	1,84	1,80	1,76	1,73	1,70	1,68	1,58	1,52	1,47	1,43	1,41	1,36	1,26	1,13	1,08	
∞	2,60	2,37	2,21	2,10	2,01	1,94	1,88	1,83	1,79	1,75	1,72	1,69	1,67	1,57	1,51	1,46	1,42	1,40	1,35	1,24	1,11	1	

F de Snedecor-Fisher $\alpha = 0,01$

g.l. 1 →		3	4	5	6	7	8	9	10	11	12	13	14	15	20	25	30	35	40	50	100	500	∞
g.l. 2 ↓	5	12,06	11,39	10,97	10,67	10,46	10,29	10,16	10,05	9,96	9,89	9,82	9,77	9,72	9,55	9,45	9,38	9,33	9,29	9,24	9,13	9,04	9,02
	6	9,78	9,15	8,75	8,47	8,26	8,10	7,98	7,87	7,79	7,72	7,66	7,60	7,56	7,40	7,30	7,23	7,18	7,14	7,09	6,99	6,90	6,88
	7	8,45	7,85	7,46	7,19	6,99	6,84	6,72	6,62	6,54	6,47	6,41	6,36	6,31	6,16	6,06	5,99	5,94	5,91	5,86	5,75	5,67	5,65
	8	7,59	7,01	6,63	6,37	6,18	6,03	5,91	5,81	5,73	5,67	5,61	5,56	5,52	5,36	5,26	5,20	5,15	5,12	5,07	4,96	4,88	4,86
	9	6,99	6,42	6,06	5,80	5,61	5,47	5,35	5,26	5,18	5,11	5,05	5,01	4,96	4,81	4,71	4,65	4,60	4,57	4,52	4,41	4,33	4,31
10		6,55	5,99	5,64	5,39	5,20	5,06	4,94	4,85	4,77	4,71	4,65	4,60	4,56	4,41	4,31	4,25	4,20	4,17	4,12	4,01	3,93	3,91
11		6,22	5,67	5,32	5,07	4,89	4,74	4,63	4,54	4,46	4,40	4,34	4,29	4,25	4,10	4,01	3,94	3,89	3,86	3,81	3,71	3,62	3,60
12		5,95	5,41	5,06	4,82	4,64	4,50	4,39	4,30	4,22	4,16	4,10	4,05	4,01	3,86	3,76	3,70	3,65	3,62	3,57	3,47	3,38	3,36
13		5,74	5,21	4,86	4,62	4,44	4,30	4,19	4,10	4,02	3,96	3,91	3,86	3,82	3,66	3,57	3,51	3,46	3,43	3,38	3,27	3,19	3,17
14		5,56	5,04	4,69	4,46	4,28	4,14	4,03	3,94	3,86	3,80	3,75	3,70	3,66	3,51	3,41	3,35	3,30	3,27	3,22	3,11	3,03	3,00
15		5,42	4,89	4,56	4,32	4,14	4,00	3,89	3,80	3,73	3,67	3,61	3,56	3,52	3,37	3,28	3,21	3,17	3,13	3,08	2,98	2,89	2,87
16		5,29	4,77	4,44	4,20	4,03	3,89	3,78	3,69	3,62	3,55	3,50	3,45	3,41	3,26	3,16	3,10	3,05	3,02	2,97	2,86	2,78	2,75
17		5,19	4,67	4,34	4,10	3,93	3,79	3,68	3,59	3,52	3,46	3,40	3,35	3,31	3,16	3,07	3,00	2,96	2,92	2,87	2,76	2,68	2,65
18		5,09	4,58	4,25	4,01	3,84	3,71	3,60	3,51	3,43	3,37	3,32	3,27	3,23	3,08	2,98	2,92	2,87	2,84	2,78	2,68	2,59	2,57
19		5,01	4,50	4,17	3,94	3,77	3,63	3,52	3,43	3,36	3,30	3,24	3,19	3,15	3,00	2,91	2,84	2,80	2,76	2,71	2,60	2,51	2,49
20		4,94	4,43	4,10	3,87	3,70	3,56	3,46	3,37	3,29	3,23	3,18	3,13	3,09	2,94	2,84	2,78	2,73	2,69	2,64	2,54	2,44	2,42
21		4,87	4,37	4,04	3,81	3,64	3,51	3,40	3,31	3,24	3,17	3,12	3,07	3,03	2,88	2,79	2,72	2,67	2,64	2,58	2,48	2,38	2,36
22		4,82	4,31	3,99	3,76	3,59	3,45	3,35	3,26	3,18	3,12	3,07	3,02	2,98	2,83	2,73	2,67	2,62	2,58	2,53	2,42	2,33	2,31
23		4,76	4,26	3,94	3,71	3,54	3,41	3,30	3,21	3,14	3,07	3,02	2,97	2,93	2,78	2,69	2,62	2,57	2,54	2,48	2,37	2,28	2,26
24		4,72	4,22	3,90	3,67	3,50	3,36	3,26	3,17	3,09	3,03	2,98	2,93	2,89	2,74	2,64	2,58	2,53	2,49	2,44	2,33	2,24	2,21
25		4,68	4,18	3,85	3,63	3,46	3,32	3,22	3,13	3,06	2,99	2,94	2,89	2,85	2,70	2,60	2,54	2,49	2,45	2,40	2,29	2,19	2,17
26		4,64	4,14	3,82	3,59	3,42	3,29	3,18	3,09	3,02	2,96	2,90	2,86	2,81	2,66	2,57	2,50	2,45	2,42	2,36	2,25	2,16	2,13
27		4,60	4,11	3,78	3,56	3,39	3,26	3,15	3,06	2,99	2,93	2,87	2,82	2,78	2,63	2,54	2,47	2,42	2,38	2,33	2,22	2,12	2,10
28		4,57	4,07	3,75	3,53	3,36	3,23	3,12	3,03	2,96	2,90	2,84	2,79	2,75	2,60	2,51	2,44	2,39	2,35	2,30	2,19	2,09	2,08
29		4,54	4,04	3,73	3,50	3,33	3,20	3,09	3,00	2,93	2,87	2,81	2,77	2,73	2,57	2,48	2,41	2,36	2,33	2,27	2,16	2,06	2,03
30		4,51	4,02	3,70	3,47	3,30	3,17	3,07	2,98	2,91	2,84	2,79	2,74	2,70	2,55	2,45	2,39	2,34	2,30	2,25	2,13	2,03	2,01
40		4,31	3,83	3,51	3,29	3,12	2,99	2,89	2,80	2,73	2,66	2,61	2,56	2,52	2,37	2,27	2,20	2,15	2,11	2,06	1,94	1,83	1,80
50		4,20	3,72	3,41	3,19	3,02	2,89	2,78	2,70	2,63	2,56	2,51	2,46	2,42	2,27	2,17	2,10	2,05	2,01	1,95	1,82	1,71	1,68
75		4,05	3,58	3,27	3,05	2,89	2,76	2,65	2,57	2,49	2,43	2,38	2,33	2,29	2,13	2,03	1,96	1,91	1,87	1,81	1,67	1,55	1,52
100		3,98	3,51	3,21	2,99	2,82	2,69	2,59	2,50	2,43	2,37	2,31	2,27	2,22	2,07	1,97	1,89	1,84	1,80	1,74	1,60	1,47	1,43
200		3,88	3,41	3,11	2,89	2,73	2,60	2,50	2,41	2,34	2,27	2,22	2,17	2,13	1,97	1,87	1,79	1,74	1,69	1,63	1,48	1,33	1,29
1000		3,80	3,34	3,04	2,82	2,66	2,53	2,43	2,34	2,27	2,20	2,15	2,10	2,06	1,90	1,79	1,72	1,66	1,61	1,54	1,38	1,19	1,11
∞		3,78	3,32	3,02	2,80	2,64	2,51	2,41	2,32	2,25	2,18	2,13	2,08	2,04	1,88	1,77	1,70	1,64	1,59	1,52	1,36	1,15	1

F de Snedecor-Fisher $\alpha = 0,001$

g.l. 2 ↓ g.l. 1 →	3	4	5	6	7	8	9	10	11	12	13	14	15	20	25	30	35	40	50	100	500	∞
	33,2	31,1	29,8	28,8	28,2	27,6	27,2	26,9	26,6	26,4	26,2	26,1	25,9	25,4	25,1	24,9	24,7	24,6	24,4	24,1	23,9	23,79
5																						
6	23,7	21,9	20,8	20,0	19,5	19,0	18,7	18,4	18,2	18,0	17,8	17,7	17,6	17,1	16,9	16,7	16,5	16,4	16,3	16,0	15,8	15,75
7	18,8	17,2	16,2	15,5	15,0	14,6	14,3	14,1	13,9	13,7	13,6	13,4	13,3	12,9	12,7	12,5	12,4	12,3	12,2	12,0	11,7	11,70
8	15,8	14,4	13,5	12,9	12,4	12,0	11,8	11,5	11,4	11,2	11,1	10,9	10,8	10,5	10,3	10,1	10,0	9,9	9,8	9,6	9,4	9,33
9	13,90	12,56	11,71	11,13	10,70	10,37	10,11	9,89	9,72	9,57	9,44	9,33	9,24	8,90	8,69	8,55	8,45	8,37	8,26	8,04	7,86	7,81
10	12,55	11,28	10,48	9,93	9,52	9,20	8,96	8,75	8,59	8,45	8,32	8,22	8,13	7,80	7,60	7,47	7,37	7,30	7,19	6,98	6,81	6,76
11	11,56	10,35	9,58	9,05	8,65	8,35	8,12	7,92	7,76	7,63	7,51	7,41	7,32	7,01	6,81	6,68	6,59	6,52	6,42	6,21	6,04	6,00
12	10,80	9,63	8,89	8,38	8,00	7,71	7,48	7,29	7,14	7,00	6,89	6,79	6,71	6,40	6,22	6,09	6,00	5,93	5,83	5,63	5,46	5,42
13	10,21	9,07	8,35	7,86	7,49	7,21	6,98	6,80	6,65	6,52	6,41	6,31	6,23	5,93	5,75	5,63	5,54	5,47	5,37	5,17	5,01	4,97
14	9,73	8,62	7,92	7,44	7,08	6,80	6,58	6,40	6,26	6,13	6,02	5,93	5,85	5,56	5,38	5,25	5,17	5,10	5,00	4,81	4,65	4,60
15	9,34	8,25	7,57	7,09	6,74	6,47	6,26	6,08	5,94	5,81	5,71	5,62	5,54	5,25	5,07	4,95	4,86	4,80	4,70	4,51	4,35	4,31
16	9,01	7,94	7,27	6,80	6,46	6,20	5,98	5,81	5,67	5,55	5,44	5,35	5,27	4,99	4,82	4,70	4,61	4,54	4,45	4,26	4,10	4,06
17	8,73	7,68	7,02	6,56	6,22	5,96	5,75	5,58	5,44	5,32	5,22	5,13	5,05	4,78	4,60	4,48	4,40	4,33	4,24	4,05	3,89	3,85
18	8,49	7,46	6,81	6,35	6,02	5,76	5,56	5,39	5,25	5,13	5,03	4,94	4,87	4,59	4,42	4,30	4,22	4,15	4,06	3,87	3,71	3,67
19	8,28	7,27	6,62	6,18	5,85	5,59	5,39	5,22	5,08	4,97	4,87	4,78	4,70	4,43	4,26	4,14	4,06	3,99	3,90	3,71	3,55	3,51
20	8,10	7,10	6,46	6,02	5,69	5,44	5,24	5,08	4,94	4,82	4,72	4,64	4,56	4,29	4,12	4,00	3,92	3,86	3,77	3,58	3,42	3,38
21	7,94	6,95	6,32	5,88	5,56	5,31	5,11	4,95	4,81	4,70	4,60	4,51	4,44	4,17	4,00	3,88	3,80	3,74	3,64	3,46	3,30	3,26
22	7,80	6,81	6,19	5,76	5,44	5,19	4,99	4,83	4,70	4,58	4,49	4,40	4,33	4,06	3,89	3,78	3,69	3,63	3,54	3,35	3,19	3,15
23	7,67	6,70	6,08	5,65	5,33	5,09	4,89	4,73	4,60	4,48	4,39	4,30	4,23	3,96	3,79	3,68	3,60	3,53	3,44	3,25	3,10	3,05
24	7,55	6,59	5,98	5,55	5,24	4,99	4,80	4,64	4,51	4,39	4,30	4,21	4,14	3,87	3,71	3,59	3,51	3,45	3,36	3,17	3,01	2,97
25	7,45	6,49	5,89	5,46	5,15	4,91	4,71	4,56	4,42	4,31	4,22	4,13	4,06	3,79	3,63	3,52	3,43	3,37	3,28	3,09	2,93	2,89
26	7,36	6,41	5,80	5,38	5,07	4,83	4,64	4,48	4,35	4,24	4,14	4,06	3,99	3,72	3,56	3,44	3,36	3,30	3,21	3,02	2,86	2,82
27	7,27	6,33	5,73	5,31	5,00	4,76	4,57	4,41	4,28	4,17	4,08	3,99	3,92	3,66	3,49	3,38	3,30	3,23	3,14	2,96	2,80	2,75
28	7,19	6,25	5,66	5,24	4,93	4,69	4,50	4,35	4,22	4,11	4,01	3,93	3,86	3,60	3,43	3,32	3,24	3,18	3,09	2,90	2,74	2,69
29	7,12	6,19	5,59	5,18	4,87	4,64	4,45	4,29	4,16	4,05	3,96	3,88	3,80	3,54	3,38	3,27	3,18	3,12	3,03	2,84	2,68	2,64
30	7,05	6,12	5,53	5,12	4,82	4,58	4,39	4,24	4,11	4,00	3,91	3,82	3,75	3,49	3,33	3,22	3,13	3,07	2,98	2,79	2,63	2,59
40	6,59	5,70	5,13	4,73	4,44	4,21	4,02	3,87	3,75	3,64	3,55	3,47	3,40	3,15	2,98	2,87	2,79	2,73	2,64	2,44	2,28	2,23
50	6,34	5,46	4,90	4,51	4,22	4,00	3,82	3,67	3,55	3,44	3,35	3,27	3,20	2,95	2,79	2,68	2,60	2,53	2,44	2,25	2,07	2,03
75	6,01	5,16	4,62	4,24	3,96	3,74	3,56	3,42	3,30	3,19	3,10	3,03	2,96	2,71	2,55	2,44	2,35	2,29	2,19	1,99	1,81	1,75
100	5,86	5,02	4,48	4,11	3,83	3,61	3,44	3,30	3,18	3,07	2,99	2,91	2,84	2,59	2,43	2,32	2,24	2,17	2,08	1,87	1,67	1,62
200	5,63	4,81	4,29	3,92	3,65	3,43	3,26	3,12	3,00	2,90	2,82	2,74	2,67	2,42	2,26	2,15	2,07	2,00	1,90	1,68	1,46	1,39
1000	5,46	4,65	4,14	3,78	3,51	3,30	3,13	2,99	2,87	2,77	2,69	2,61	2,54	2,30	2,14	2,02	1,94	1,87	1,77	1,53	1,27	1,15
∞	5,42	4,62	4,10	3,74	3,47	3,27	3,10	2,96	2,84	2,74	2,66	2,58	2,51	2,27	2,10	1,99	1,90	1,84	1,73	1,49	1,21	1

Tema 11 : Planificación de estudios estadísticos. Clases de estudios.

Los descubrimientos o avances científicos pueden ser fruto de

- 1) la casualidad, muy a menudo unida a una intuición genial. Por ejemplo, el descubrimiento de los Rx, la penicilina, el yodo, la ley de la gravedad....
- 2) la búsqueda de soluciones a problemas, como la necesidad de nuevos medicamentos o nuevos combustibles.
- 3) la curiosidad teórica, con Einstein como uno de los mejores ejemplos.

El primer camino es excepcional, no porque no se den ocasiones, sino porque la mayoría de las personas no reconocen la trascendencia de la observación. La suerte sólo favorece a los preparados (Pasteur). Los otros dos caminos son los habituales y requieren un estudio planificado.

Etapas fundamentales de un estudio

En un estudio planificado se pueden distinguir 5 etapas fundamentales: 1 planteamiento, 2 información, 3 formulación de la hipótesis, 4 realización u obtención de datos y 5 análisis de resultados y conclusiones.

Esta distinción se hace a efectos teóricos y didácticos, pues en la práctica al comienzo del trabajo se imbrican las tres primeras etapas y sólo al cabo de un tiempo quedan claramente definidas, cosa que inexcusablemente debe de ocurrir antes de iniciar el paso 4º, la realización. Veamos estas etapas con más detalle:

- 1) **PLANTEAMIENTO** : qué se va a estudiar, por qué, para qué, cómo, etc

El “cómo” incluye

- a) el diseño de la investigación: lo que habitualmente se conoce en los trabajos científicos como material y métodos, p.e. el nº de individuos a estudiar, las características que deben reunir, el procedimiento de elección, tratamiento aplicado, variables a medir, etc
- b) las necesidades de material, personal y dinero.

Como ya se ha dicho el planteamiento inicial es provisional, pudiendo ser modificado en función de los pasos 2 y 3.

- 2) **INFORMACION** : es preciso saber lo máximo posible sobre el tema de la investigación, consultando libros y revistas especializadas. Es lo que se llama “revisión bibliográfica” o “revisión de la literatura”.

Este material debe ser valorado críticamente. Ante cada trabajo concreto hay que hacerse una serie de preguntas. ¿quien lo ha escrito? , ¿donde? , ¿cuando? , ¿el material y el método utilizados son correctos? , ¿están justificadas las conclusiones? , etc... El motivo de esta valoración crítica es que es muy, muy difícil hacer bien un trabajo científico, por lo que la inmensa mayoría tienen errores y deficiencias más o menos trascendentes.

Tras este examen habrá cosas claras y generalmente aceptadas, mientras que otras serán inciertas, dudosas o controvertidas. Se tomará buena nota de los fallos observados en otros investigadores para no incurrir en ellos.

- 3) **HIPOTESIS** : es la explicación provisional de unos hechos. Al concluir la investigación se verá si es o no cierta (“verificación” de la hipótesis). Los estudios puramente descriptivos no tienen hipótesis, aunque pueden servir de base para formular hipótesis.
- 4) **REALIZACION U OBTENCION DE DATOS (RECOGIDA DE LA INFORMACION)**
Para ello se va cumpliendo exactamente lo previsto en el punto “Material y métodos” del paso nº 1. Una vez recogidos todos los datos se clasifican y ordenan siguiendo las normas de la Estadística Descriptiva. Es importante buscar posibles errores de ejecución y desechar todo lo que no se ajuste exactamente al método previsto.

- 5) **ANALISIS DE LOS RESULTADOS Y CONCLUSIONES**

Se aplica el método de análisis estadístico que corresponda al tipo de datos y al objetivo de la investigación. Así se verifica la hipótesis de trabajo, es decir se confirma o se desecha. Las

hipótesis no confirmadas también tiene su valor. Así, puede concluirse que un nuevo medicamento no es más eficaz que los que había, que una nueva técnica no mejora la actual, etc. Todo ello permitirá sacar **CONCLUSIONES**. Hay que distinguir entre las conclusiones estadísticas, que como se verá en su momento llevan anejo un juicio de significación y si es posible un juicio de causalidad, y las conclusiones del estudio que se basan en las anteriores. Es conveniente recordar que las conclusiones estadísticas lo son a nivel de grupo, no a nivel individual. Son válidas para la inmensa mayoría de los individuos, no para todos. “La estadística no es una ciencia exacta”.

Un error frecuente es sacar conclusiones basadas en la información previa, no en el estudio

Clases de estudios estadísticos

Se pueden clasificar desde distintos puntos de vista:

■ en función del nº de variables:

- ❖ E. de **INFORMACION**: estudio de una variable
 - **DESCRIPTIVOS**: tabulación, representación gráfica, índices estadísticos...
 - de **ESTIMACION**: estimar parámetros de una población a partir de una muestra
 - de **CONFORMIDAD**: valorar si una muestra puede proceder de una población determinada
- ❖ E. de **INVESTIGACION O COMPARATIVOS**: diferencias o relaciones entre dos o más variables
 - **EXPERIMENTALES**
 - Clásicos: 1 variable controlada y el resto aleatorias
 - Factoriales: 2 ó más variables controladas y el resto aleatorias
 - de **OBSERVACION**: todas las variables son aleatorias.

Sólo los estudios experimentales permiten una interpretación causal

■ en función del momento en que se generan los datos:

- ❖ Estudios **RETROSPECTIVOS** o históricos. Los datos ya se han generado cuando se planifica, por lo que los métodos previstos en “material y métodos” pueden no haber sido observados exactamente. p.e. se revisan las historias clínicas de 1000 pacientes que tomaron el medicamento M para ver los efectos secundarios que presentaron.
A este grupo pertenecen los estudios caso-control: un grupo de individuos afectados se compara con otro u otros no afectados para investigar el nivel de exposición a determinados factores que podrían ser causales o protectores. Cada caso se empareja con uno o más controles, que por lo demás deben ser lo más parecidos posible a los casos (sexo, edad, etc). Es la herramienta de trabajo clásica de los estudios epidemiológicos, p.e., en el caso de una intoxicación alimenticia en una boda. Su parámetro típico es la razón de probabilidad u **ODDS RATIO (OR)**, que veremos en otro tema.
- ❖ Estudios **PROSPECTIVOS** o de futuro. Los datos se generan después de la planificación del estudio y como consecuencia del mismo. p.e. a partir de hoy se van a recoger los efectos secundarios en mil pacientes consecutivos que toman el medicamento M.
A este grupo pertenecen los estudios de cohortes, típicos de estudios epidemiológicos, mucho menos usados que los de caso control. Son difíciles y caros y llevan más tiempo. Se seleccionan individuos expuestos y no expuestos a un factor y a lo largo del tiempo se ve si enferman o no. Su parámetro típico es el cociente de riesgo o riesgo relativo (**RR**), que veremos también en otro tema.

■ **en función de los individuos:**

- ❖ Estudios con datos independientes. Los individuos están repartidos en dos o más grupos o muestras; cada individuo sólo forma parte de un grupo. p.e. se prueba el medicamento A en 100 individuos y el B en otros 100.
- ❖ Estudios con datos apareados. todos los individuos forman parte de todos los grupos. El orden por el que entran en cada uno de los grupos se determina al azar. p.e. 100 pacientes reciben en momento dado el medicamento A y en otro momento el B y se comparan sus efectos. Los 100 pacientes forman parte del grupo medicamento A y también del grupo medicamento B.

■ **en función del conocimiento de los detalles y resultado del estudio:**

- ❖ Abiertos. Los que realizan el estudio, los que lo valoran y, si son conscientes, también los individuos conocen los grupos y el tratamiento que reciben.
- ❖ Ciegos. Quien valora los resultados desconoce a que grupo pertenecen los individuos y por tanto el tratamiento recibido.
- ❖ Doble ciegos. Ese desconocimiento se extiende a los que realizan el estudio, a los que lo valoran y a los individuos, si son conscientes. Sólo el director del estudio, que no hace la valoración, revela al final todos los detalles.

■ **en función del lugar en que se realiza el estudio:**

- ❖ unicéntricos : todo el estudio se realiza por el mismo equipo investigador
- ❖ multicéntricos: el estudio se realiza simultáneamente en diversos sitios por diversos investigadores siguiendo un diseño común.

■ **en función del método experimental:**

- ❖ con tratamiento activo. Se da el producto que se investiga.
- ❖ con placebo. Se aplica un tratamiento inactivo, sin efecto, con el mismo aspecto externo que el tratamiento activo. Esto se aplica sólo a humanos y lógicamente el individuo no sabe lo que está tomando.

En los últimos años las revistas científicas más prestigiosas han introducido de forma obligatoria la “Declaración de intereses”: los autores declaran si tienen o han tenido alguna relación laboral, comercial, de asesoría o de mecenazgo con personas, empresas o instituciones que tengan algo que ver con el estudio. Es decir, si hay o no hay “conflicto de intereses”.

Los mejores estudios son los unicéntricos, experimentales, prospectivos, doble ciegos, incluyendo placebo y si es posible con datos apareados.