

Apuntes de Estadística Descriptiva

Asignatura: Complementos de Matemáticas

Profesor: Dr. Manuel J. Galán Moreno

E.T.S.A.

ESTADÍSTICA DESCRIPTIVA

INTRODUCCIÓN.

FENÓMENOS DETERMINISTAS Y FENÓMENOS ALEATORIOS.

FENÓMENO DETERMINISTA.

Se denomina *fenómeno determinista* a toda experiencia que, al ser repetida en condiciones que nosotros consideramos similares, produce siempre el mismo resultado, dentro de unos límites razonables de precisión.

La mayoría de los fenómenos macroscópicos estudiados por las ciencias de la naturaleza y de la vida.

FENÓMENO ALEATORIO.

Las experiencias que, al ser repetidas en condiciones que estimamos similares, nos llevan a resultados diferentes constituyen los llamados *fenómenos aleatorios*.

- El lanzamiento de un dado.
- El número de usuarios que acuden a un banco en un determinado intervalo horario.
- Las consecuencias de la administración de una medicación.

CARACTERIZACIÓN DE LOS FENÓMENOS ALEATORIOS.

- Se pueden repetir en condiciones esencialmente análogas.
- Existe un conjunto que contiene todos los resultados posibles (**Universo o espacio muestral**).
- Antes de realizar el experimento no se puede predecir el resultado exacto.
- La frecuencia relativa de cada resultado tiende a estabilizarse al repetir indefinidamente el experimento.

¿PORQUE EXISTEN FENÓMENOS ALEATORIOS ?

- Imposibilidad de controlar todos los factores y condiciones iniciales que influyen en el resultado del experimento.
- Existencia de un número indeterminado de variables que pueden afectar al resultado
- Complejidad o desconocimiento de las leyes que rigen el fenómeno.

CONCEPTO DE ESTADÍSTICA.

Es el conjunto de procedimientos que nos permiten estudiar los fenómenos aleatorios.

La palabra Estadística se utiliza también como sinónimo de dato o resultado de la elaboración de un conjunto de datos mediante técnicas estadísticas.

OBJETO MATERIAL DE LA ESTADÍSTICA.

Los fenómenos aleatorios.

OBJETO FORMAL: EL MÉTODO ESTADÍSTICO.

Prescinde de lo individual y de los razonamientos de tipo causal, para considerar regularidades o propiedades aplicables a un conjunto de datos e inferir propiedades sobre la totalidad del fenómeno estudiado.

IMPORTANCIA DEL MÉTODO ESTADÍSTICO.

Permite la *inducción incompleta*. Es decir, permite la obtención de conclusiones acerca de un conjunto sin necesidad de estudiar todos y cada uno de los elementos que lo componen.

El método estadístico permite abordar el estudio de fenómenos abarcan casi todas las áreas del conocimiento y, al tener un carácter genérico, puede incorporarse como parte del método del resto de las ciencias.

EL MÉTODO CIENTÍFICO Y EL MÉTODO ESTADÍSTICO.

ETAPAS DEL MÉTODO CIENTÍFICO.

- Observación de un fenómeno.
- Planteamiento de hipótesis que expliquen el fenómeno observado.
- Deducción, a partir de las hipótesis propuestas, de consecuencias que puedan ser verificadas mediante experimentos.
- Verificación experimental de las consecuencias.
- Resolución sobre la idoneidad de las hipótesis según los resultados de la experimentación.

ETAPAS DEL MÉTODO ESTADÍSTICO.

- Definición de OBJETIVOS.
- Definición de UNIVERSO y MUESTRA.
- Definición de TÉRMINOS y UNIDADES de medida.
- Determinación de los DATOS necesarios.
- Recolección de los datos.
- Elaboración de los datos.
- Descripción, análisis e interpretación de los datos.

CONCEPTO DE POBLACIÓN Y MUESTRA.

POBLACIÓN O UNIVERSO.

Conjunto de elementos que poseen una característica o propiedad común, y que constituyen la totalidad de los individuos de interés para nuestro estudio.

MUESTRA.

Cualquier subconjunto de la población sobre el que se realizan los estudios para obtener conclusiones acerca de las características de la población.

INDIVIDUO.

Cada uno de los elementos de la muestra o de la población. No tienen por que ser objetos físicos. Vgr: el lanzamiento de un dado.

Al realizar un estudio estadístico, no solo es necesario definir la población de referencia y la muestra que se va a utilizar, también hay que especificar qué *características aleatorias* de los individuos vamos a tener en cuenta. Por ejemplo: intención de voto, resultado de un tratamiento, puntuación de un dado, ...

Las características aleatorias suelen corresponder a variables numéricas y si no es así, siempre pueden codificarse numéricamente. Una característica aleatoria definida numéricamente es lo que denominamos variable aleatoria.

FASES O NIVELES DEL MÉTODO ESTADÍSTICO.

ESTADÍSTICA DESCRIPTIVA.

Recopilación y análisis de los datos. Pueden ser datos referentes a toda la población o solo a una muestra, pero en este último caso no se pretende sacar conclusiones acerca de la población a la que pertenece la muestra.

TEORÍA DE LA PROBABILIDAD.

Es la teoría matemática en la que se basa la posibilidad de realizar inferencias acerca de las propiedades de la población partiendo de la información contenida en la muestra.

Históricamente se desarrollaron por separado la Estadística Descriptiva, que es la parte más antigua (hasta las civilizaciones más primitivas realizaban censos o recuentos de población), y la teoría de la probabilidad, que surgió como un *divertimento* matemático aplicable al estudio de los juegos de azar. Fue a finales del siglo pasado cuando se aprovechó el ímpetu que alcanzó el formalismo matemático para combinar ambas, obteniendo la posibilidad de realizar inferencias sobre toda una población a partir del estudio de una muestra.

ESTADÍSTICA INFERENCIAL O INDUCTIVA.

Utiliza la información que se desprende del análisis de una muestra para realizar una estimación de las propiedades de la población de la que se extrajo.

ORGANIZACIÓN DE DATOS

ESCALAS DE MEDIDA.

MEDIDA.

Medida es la asignación de números a los objetos o sucesos según ciertas reglas. Los fenómenos se presentan en distintas modalidades. Realizar una medida sobre un fenómeno equivale a asociar un valor numérico -y sólo uno- a cada una de las distintas modalidades en las que se puede presentar.

ESCALA DE MEDIDA.

Es una regla o patrón que permite asociar, de forma biunívoca, modalidades y números.

De acuerdo con las relaciones que pueden establecerse entre las distintas modalidades que presenta un fenómeno, tenemos los distintos tipos de escalas de medida que aparecen en el cuadro de la página siguiente.

ESCALAS NUMÉRICAS.

ESCALA NUMÉRICA DISCRETA O DISCONTINUA.

Es aquella en la que entre dos valores de la misma no siempre podemos situar otro. Vgr.: Número de intervenciones quirúrgicas en un día.

Con mayor rigor, es aquella que puede relacionarse mediante una aplicación *biyectiva* con el conjunto de los números Naturales o con un subconjunto de este.

ESCALA NUMÉRICA CONTINUA.

Es aquella en la que, dados dos valores cualesquiera, siempre podemos encontrar otro intermedio entre ellos. Vgr.: El peso, la estatura...

Con rigor, es aquella que puede ponerse en correspondencia biunívoca con el conjunto de los números Reales.

Tabla 1: Tipos de escalas de medida

| Escala | Operaciones posibles | Requisitos | Estadísticos válidos | Ejemplo. |
|---------------------|---|--------------------------------------|---------------------------------------|-----------------------------------|
| Nominal | Verificar la igualdad de dos modalidades. | Posibilidad de permutar modalidades. | Frecuencia, Moda. | Estado civil, Sexo, nacionalidad. |
| Ordinal | Verificar si una modalidad es mayor que otra. | Mantenimiento del orden. | Mediana, cuantiles. | Gravedad de una lesión. |
| De intervalo | Comparar las diferencias entre dos modalidades. | Unidad constante. | Media aritmética, desviación típica. | Temperatura. |
| De razón | Establecer razones entre modalidades. | Existencia de cero absoluto. | Media geométrica, coef. de variación. | Peso, altura... |

TABLAS Y GRÁFICOS ESTADÍSTICOS:

La organización de los datos recogidos en un trabajo estadístico es imprescindible para su aprovechamiento y mejor comprensión. La forma idónea de realizar este proceso es a través de tablas y gráficos. Para ello es preciso tener en cuenta las siguientes recomendaciones:

1. Gráficos y tablas deberán estar rotulados de forma clara, de modo que se expliquen por sí solos. Si se utilizan códigos o abreviaturas deben explicarse a pie de página. Todas las filas y columnas de las tablas estarán encabezadas, y los ejes de los gráficos rotulados.
2. Siempre que tenga sentido, las filas y columnas de una tabla deberán estar totalizadas.
3. Las unidades de medida han de estar claramente definidas.
4. Se deben evitar las tablas o gráficos excesivamente complejos. Es preferible realizar varias tablas simples antes que una compleja.

DISTRIBUCIÓN DE FRECUENCIAS.

FRECUENCIA ABSOLUTA.

La frecuencia absoluta de una modalidad es el número de veces que se repite esa modalidad como resultado de un experimento.

FRECUENCIA RELATIVA.

Es la frecuencia absoluta partida por el número total de observaciones.

FRECUENCIA ACUMULADA (Absoluta o relativa).

Igual que en cada uno de los anteriores casos pero sumando, no sólo, los resultados de la modalidad de que se trate, sino también los de todas las precedentes. No es válido para datos de escalas nominales, ya que en ellas no existe el orden.

DISTRIBUCIÓN DE FRECUENCIAS.

Es una disposición organizada de los datos recogidos en un estudio. Contiene un listado de las distintas modalidades del fenómeno considerado, con la frecuencia absoluta, relativa y acumulada de cada una. Cuando el número de modalidades es demasiado grande (esto ocurre siempre con las escalas continuas) se agrupan en *clases*.

DISTRIBUCIÓN DE FRECUENCIAS: REGLAS PRACTICAS.

1. Las clases han de ser excluyentes.
2. Los límites de cada clase deben tener más precisión que las medidas realizadas.
3. Aunque no tiene que ser necesariamente así, es conveniente que la amplitud de los intervalos sea constante.
4. Todos los datos de una clase quedan representados por la **marca de clase**, que es el valor medio de intervalo que forma la clase. De esta manera, todos los cálculos se realizan como si en lugar de tener N valores distintos en una clase, tuviéramos N veces la marca de clase.

MODELOS DE TABLAS ESTADÍSTICAS.

Modelo de tabla para variables cuantitativas discretas.

| Variable. | Frecuencia absoluta. | Frecuencia absoluta acumulada. | Frecuencia relativa. | Frecuencia relativa acumulada. |
|-----------|----------------------|--------------------------------|----------------------|--------------------------------|
| X1 | n1 | n1 | $f1 = n1/N$ | $F1 = f1$ |
| X2 | n2 | $n1 + n2$ | $f2 = n2/N$ | $F2 = f1 + f2$ |
| ... | ... | ... | ... | ... |
| Xi | Ni | $\sum_{j=1}^i n_j$ | $f_i = n_i/N$ | $F_i = \sum_{j=1}^i f_j$ |
| ... | ... | ... | ... | ... |
| Xn | Nn | $\sum_{j=1}^n n_j$ $N =$ | $f_n = n_n/N$ | $F_n = \sum_{j=1}^n f_j = 1$ |
| N | | 1 | | |

Modelo de tabla para variables cuantitativas continuas.

| Clases o intervalos | Marca de clase | Frecuencia absoluta. | Frecuencia absoluta acumulada. | Frecuencia relativa. | Frecuencia relativa acumulada. |
|---------------------|---------------------------|----------------------|--------------------------------|----------------------|--------------------------------|
| $[a_0, a_1)$ | $X_1 = (a_0 + a_1)/2$ | n_1 | n_1 | $f_1 = n_1/N$ | $F_1 = f_1$ |
| $[a_1, a_2)$ | $X_2 = (a_1 + a_2)/2$ | n_2 | $n_1 + n_2$ | $f_2 = n_2/N$ | $F_2 = f_1 + f_2$ |
| ... | ... | ... | ... | ... | ... |
| $[a_{i-1}, a_i)$ | $X_i = (a_{i-1} + a_i)/2$ | N_i | $\sum_{j=1}^i n_j$ | $f_i = n_i/N$ | $F_i = \sum_{j=1}^i f_j$ |
| ... | ... | ... | ... | ... | ... |
| $[a_{n-1}, a_n]$ | $X_n = (a_{n-1} + a_n)/2$ | N_n | $N = \sum_{j=1}^n n_j$ | $f_n = n_n/N$ | $F_n = \sum_{j=1}^n f_j = 1$ |
| | | N | | 1 | |

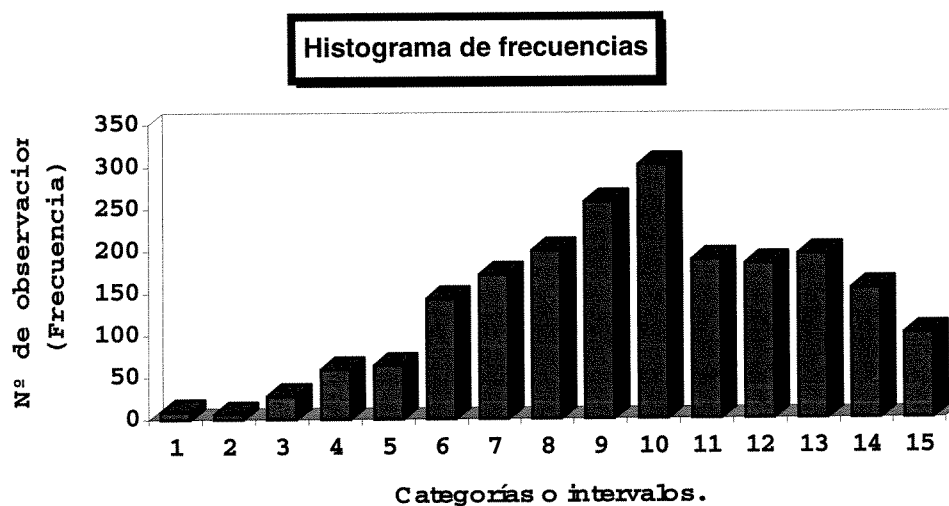
VARIABLES CUANTITATIVAS CONTINUAS: NORMAS PARA LA ELABORACIÓN DE TABLAS.

- 1º Obtener el **rango** o **amplitud** de los datos. Es la diferencia entre el valor máximo y el valor mínimo.
- 2º Determinar el **número de intervalos**. Pueden tomarse tantos intervalos como se quiera, pero es aconsejable que sea en torno a 10.
- 3º Calcular la **amplitud** de los intervalos, que es igual al rango dividido por el número de intervalos.
- 4º Determinar el **limite superior** del último intervalo y el **limite inferior** del primero.
- 5º Calcular la **marca de clase** de cada intervalo, que no es otra cosa que el punto medio del intervalo.

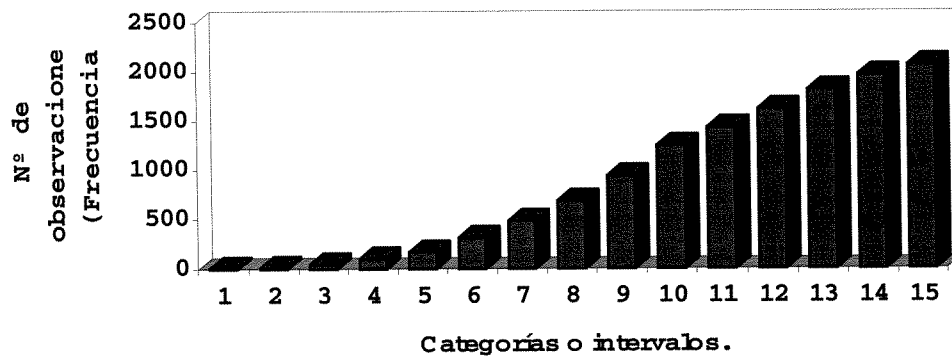
REPRESENTACIÓN GRÁFICA DE DATOS ESTADÍSTICOS.

La representación gráfica de datos tiene la ventaja de que es capaz de ofrecer de forma inmediata una perspectiva global de los resultados de un estudio.

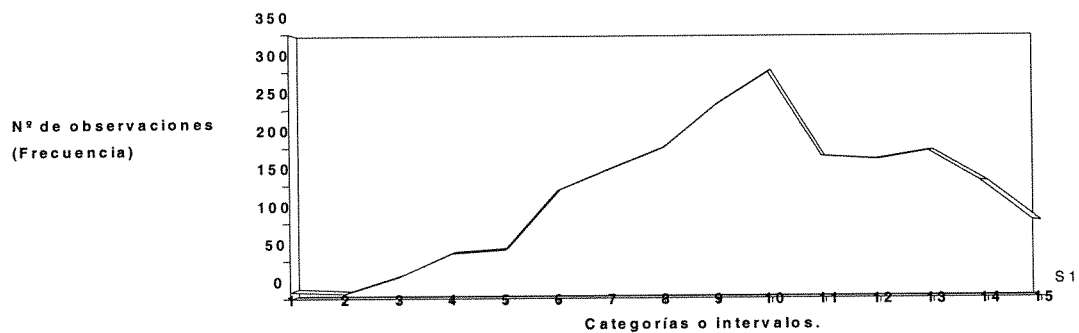
A continuación aparecen algunos de los formatos más utilizados



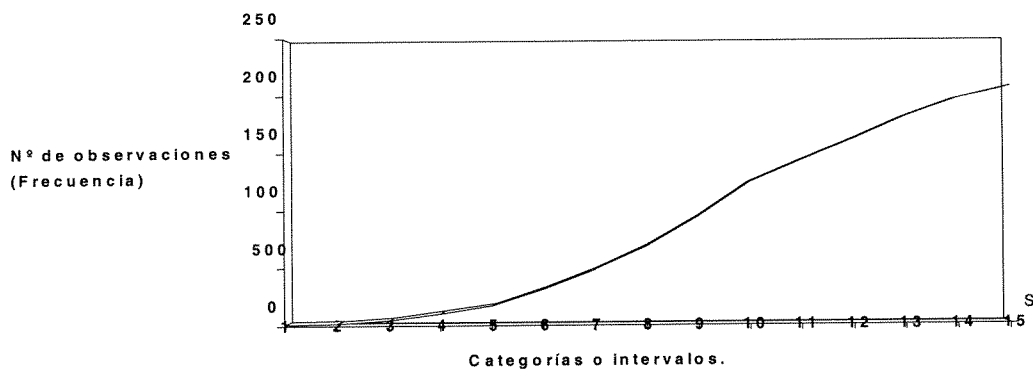
Histograma de frecuencias acumuladas



Polígono de frecuencias



Polígono de frecuencias acumuladas



ANÁLISIS DE DISTRIBUCIONES DE FRECUENCIAS UNIDIMENSIONALES.

INTRODUCCIÓN.

DEFINICIÓN.

Una serie de datos decimos que es unidimensional cuando se refiere solamente a una variable.

OBJETO DEL ANÁLISIS.

Reducir una serie de datos a unos pocos coeficientes que contengan la mayor parte de la información relevante, con el fin de descubrir regularidades estadísticas en el colectivo analizado.

COEFICIENTES MAS IMPORTANTES:

MEDIDAS DE POSICIÓN O CENTRALIZACIÓN:

Tratan de identificar el valor más representativo de la distribución. Es decir, si tenemos que comparar el salario de los españoles con el de los franceses, no tendría sentido estudiar individuo a individuo, sino que sería más práctico representar el salario de los españoles por un determinado valor y el de los franceses por otro, y comparar esos dos. El valor que se elija para representar el salario de todos los españoles será una medida de posición que, en función de nuestros intereses, puede ser la media, la mediana, la moda, etc.

MEDIDAS DE DISPERSION:

Determinan el grado de alejamiento de los datos respecto a una medida de posición que, generalmente, suele ser la media aritmética. Nos dan una idea acerca de lo agrupados que están los datos, y por lo tanto miden la homogeneidad de estos.

MEDIDAS DE FORMA:

Miden el grado de deformación respecto a una curva patrón (*distribución Normal*).

NOTACIÓN.

x_i = Valor asociado a la modalidad i .

Cuando las modalidades se agrupan en clases, denota la *marca de clase*.

n_i = *Frecuencia absoluta* de la modalidad i .

I = Número de modalidades o de clases..

N = Número total de observaciones:

$$N = \sum_{i=1}^I n_i$$

MOMENTOS.

CONCEPTO Y UTILIDAD.

Es un artificio que nos permite unificar el tratamiento matemático de los principales coeficientes de la Estadística Descriptiva.

MOMENTO DE ORDEN P RESPECTO DE UN PUNTO x_0 .

$$a_{px_0} = \frac{1}{N} \sum_{i=1}^I n_i (x_i - x_0)^p$$

MOMENTO ORDINARIO DE ORDEN P.

Es un momento en el que $x_0 = 0$.

$$a_p = \frac{1}{N} \sum_{i=1}^I n_i x_i^p$$

MOMENTO CENTRAL DE ORDEN P.

Es un momento en el que $x_0 =$ media aritmética.

$$m_p = \frac{1}{N} \sum_{i=1}^I n_i (x_i - \bar{x})^p$$

MEDIDAS DE POSICIÓN.

MEDIAS.

MEDIA GENERALIZADA DE ORDEN P.

$$M_p = \sqrt[p]{a_p} = \sqrt[p]{\frac{1}{N} \sum_{i=1}^I n_i x_i^p}$$

Propiedad: $x_p < x_{p+1}$

MEDIA ARMÓNICA.

$$H = M_{-1} = \frac{N}{\sum_{i=1}^I \frac{n_i}{x_i}}$$

MEDIA GEOMÉTRICA.

$$g = M_0 = \sqrt[n]{\prod_{i=1}^I x_i^{n_i}}$$

MEDIA ARITMÉTICA SIMPLE.

$$\bar{x} = M_1 = \frac{1}{N} \sum_{i=1}^I n_i x_i$$

MEDIA ARITMÉTICA PONDERADA.

$$\bar{x} = \frac{1}{N} \sum_{i=1}^I p_i x_i \quad \text{Siendo } \sum_{i=1}^I p_i = N$$

CUANTILES.

PERCENTIL (p_k).

Es el valor que deja por debajo de él al K% de las observaciones.
Lógicamente $0 < K < 100$.

B DECILES

Son los percentiles múltiplos de 10:

$$d_1 = p_{10}; d_2 = p_{20}; \dots$$

CUARTILES.

Son los percentiles múltiplos de 25:

$$Q_1 = p_{25}; Q_2 = p_{50} = d_5; Q_3 = p_{75}$$

MEDIANA.

Es el valor de la distribución que deja la mitad de las observaciones por debajo de ella y la otra mitad por encima.

$$M = Q_2 = d_5 = p_{50}$$

MODA.

Es el valor más frecuente, es decir, el que más se repite.

PROPIEDADES DE LAS MEDIDAS DE POSICIÓN.

MEDIAS

En su cálculo intervienen todos los datos, por lo tanto, se ven influidos por la variación de cualquiera de ellos.
En particular, tienen el inconveniente de que los valores extremos producen grandes modificaciones.

La media más utilizada es la *aritmética*.

La media *geométrica*, aunque poco utilizada, es más adecuada cuando se trata de promediar incrementos porcentuales que actúan de forma sucesiva.

La media *aritmética ponderada* es muy útil cuando se considera que los distintos valores promediados tienen una importancia desigual.

La relación sobre las distintas medias realizadas sobre una misma serie de datos es la siguiente:

$$H < g < \bar{x}$$

MEDIANA.

Utiliza menos información que la media, ya que sólo depende del orden de los datos, pero tiene la ventaja de que no se ve influida por los valores extremos.

La relación entre la media y la mediana depende de la simetría de la distribución, de modo que:

$M < \bar{x}$ si la distribución es asimétrica positiva.

$M = \bar{x}$ si la distribución es simétrica

$M > \bar{x}$ si la distribución es asimétrica negativa.

Estos conceptos se comprenderán mejor al hablar de *medidas de forma*.

Cuando se trata de distribuciones que no son excesivamente asimétricas, se cumple la siguiente relación empírica entre la media, la moda y la mediana:

$$\bar{x} - Mo \approx 3(\bar{x} - M)$$

MEDIDAS DE DISPERSION.

VARIANZA.

Es el momento centrado de orden 2:

$$s^2 = m_2 = \frac{1}{N} \sum_{i=1}^l n_i (x_i - \bar{x})^2 = \frac{1}{N} \sum_{i=1}^l n_i x_i^2 - \bar{x}^2 = a_2 - \bar{x}^2$$

DESVIACIÓN TÍPICA.

$$s = \sqrt{s^2} = \sqrt{\frac{1}{N} \sum_{i=1}^l n_i (x_i - \bar{x})^2} = \sqrt{\frac{1}{N} \sum_{i=1}^l n_i x_i^2 - \bar{x}^2} = \sqrt{a_2 - \bar{x}^2}$$

3.DESVIACION MEDIA

$$DM = \frac{1}{N} \sum_{i=1}^l n_i |x_i - \bar{x}|$$

RANGO O RECORRIDO:

Es la distancia entre el mayor y el menor valor de la distribución y se halla calculando la diferencia entre ellos.

RANGO INTERCUARTILICO.

Es la diferencia entre Q_2 y Q_3 .

RANGO INTERDECILICO.

Es la diferencia entre d_9 y d_1 .

COEFICIENTE DE VARIACIÓN.

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

INTERPRETACIÓN DE LA DESVIACIÓN TÍPICA.

DESIGUALDAD DE TCHEBYCHEV:

$$f(|x_i - \bar{x}| > ks) \leq \frac{1}{k^2}$$

Para cualquier distribución de frecuencias, la frecuencia relativa de los valores que distan de la media más de k desviaciones típicas es menor de $1/k^2$. Por lo tanto, cuanto menor sea la desviación típica, mayor será el porcentaje de datos contenidos en un intervalo dado en torno a la media.

Por ejemplo, si la duración media de las intervenciones en un quirófano de urgencias es de 100 minutos, con una desviación típica de 8', entonces podemos decir que tendrán una duración entre 76' y 124' más del 89% de las intervenciones, ya que hemos tomado un intervalo de 3s en torno a la media y, por lo tanto, $1 - 1/3^2 = 1 - 1/9 = 0.889$

Una cuestión diferente es la de comparar la dispersión de dos series de medidas referentes a variables distintas o incluso a una misma variable pero con valores medios diferentes. Por ejemplo: estudiamos el tiempo que tardan en realizar una complicada intervención quirúrgica diversos equipos médicos encontramos que su promedio es de 210' con una desviación típica de 17'; por otro lado, estudiamos el tiempo promedio en la realización de una primera visita de consultas externas y constatamos que es de 14' con una desviación típica de 6'. En un primer análisis, podríamos pensar que la dispersión en los tiempos medidos es mayor en el primer caso, sin embargo, no es así, pues, a pesar de que la desviación típica es mayor, tiene menos importancia una variación de 17' frente a un total de 210', que una variación de 6' frente a un total de 14'. Por lo tanto podríamos concluir que la actuación de estos equipos es más homogénea en su actuación quirúrgica que en consulta.

El coeficiente de variación nos habría llevado a este resultado directamente, ya que precisamente calcula la importancia relativa de la desviación típica respecto de la media. Así, en el primer caso $C.V. = 0.08$ (8% expresado en porcentaje) mientras que en el segundo $C.V. = 0.43$ (43%). Es frecuente que, en lugar de la desviación típica, se utilice la denominada *Desviación Standard*, que se calcula a partir de aquella mediante la relación:

$$\sigma = s \sqrt{\frac{N}{N-1}} = \sqrt{\frac{1}{N-1} \sum_{i=1}^l n_i (x_i - \bar{x})^2} = \sqrt{\frac{1}{N-1} \sum_{i=1}^l n_i x_i^2 - \bar{x}^2}$$

En estadística Inferencial se utiliza σ preferentemente a s , ya que, como más tarde veremos, el valor de σ es el mejor estimador de la desviación típica poblacional partiendo de la muestra.

MEDIDAS RESISTENTES DE DISPERSION.

Son las que, igual que la mediana, no se ven afectadas por los datos extremos. La más resistente en este caso es el *rango intercuartílico*, y algo menos, la *desviación media*.

ANÁLISIS DE LA FORMA.

COEFICIENTES DE SIMETRÍA-ASIMETRÍA.

$$A) g_1 = \frac{m_3}{s^3}$$

$$B) \frac{\bar{x} - M}{s}$$

INTERPRETACIÓN:

$g_1 > 0 \implies$ Curva asimétrica positiva.

$g_1 = 0 \implies$ Curva simétrica.

$g_1 < 0 \implies$ Curva asimétrica negativa.

COEFICIENTE DE APUNTAMIENTO O CURTOSIS.

$$g_2 = \frac{m_4}{s^4}$$

INTERPRETACIÓN:

$g_2 > 1 \implies$ Leptocúrtica.

$g_2 = 1 \implies$ Mesocúrtica.

$g_2 < 1 \implies$ Platicúrtica.

SIGNIFICADO:

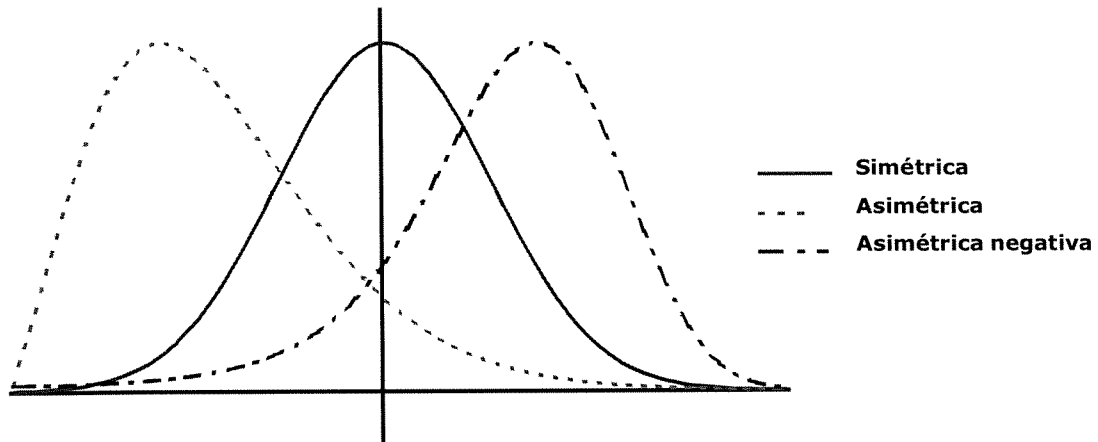
Leptocúrtica: Más apuntada que la normal.

Mesocúrtica: Simula a la normal.

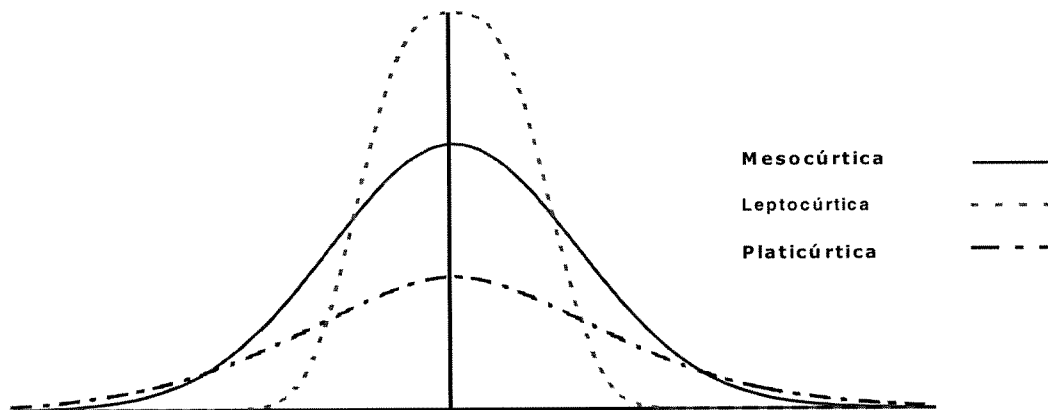
Platicúrtica: Más achatada que la normal.

Los gráficos de las páginas siguientes aclaran un poco estos conceptos.

Simetría



Curtosis



TRANSFORMACIONES.

CONCEPTO.

A veces los datos que nos interesan no son exactamente los que hemos recogido, sino otros que se derivan de éstos a través de una transformación; otras veces, una transformación puede simplificar los datos, haciendo más cómodo su manejo.

EFFECTOS DE UNA TRANSFORMACIÓN LINEAL $y = a + bx$

Media: $\bar{y} = a + b\bar{x}$
 Mediana: $M_y = a + bM_x$
 Desviación Típica: $s_y = |b|s_x$
 Medidas de forma: No varían.

TRANSFORMACIONES NO LINEALES $y = f(x)$

Mediana: $M_y = f(M_x)$.
 El resto sólo puede calcularse, en general, de forma aproximada:
 Media: $\bar{y} \approx f(\bar{x})$
 Desviación Típica: $s_y \approx s_x f'(\bar{x})$

ANÁLISIS DE DISTRIBUCIONES DE FRECUENCIAS MULTIDIMENSIONALES.

DEFINICIÓN.

Una serie de datos decimos que es MULTIDIMENSIONAL cuando para cada individuo recogemos información acerca de más de una variable.

TABULACIÓN. DISTRIBUCIÓN CONJUNTA. DISTRIBUCIONES MARGINALES Y CONDICIONADAS.

DISTRIBUCIÓN BIDIMENSIONAL: X, Y.

Si solamente estudiamos 2 variables X, Y, podemos representar los datos en una tabla de doble entrada, de modo que, en la cabecera de las filas ponemos las modalidades de una de las variables y en la cabecera de las columnas las de la otra. En las celdillas que se forman se anota el número de observaciones que presentan a la vez las características de la fila y la columna en la que se encuentran. A esta estructura se le denomina distribución conjunta o, en el caso de variables cualitativas, tabla de contingencia.

En la siguiente página aparece un modelo general para tablas de una distribución bivariable.

Distribución conjunta de frecuencias para dos variables.

| | | X | | | | |
|---|-------|----------|----------|---------|----------|-----------------------|
| | | X_1 | X_2 | \dots | X_n | |
| Y | Y_1 | n_{11} | n_{12} | \dots | n_{1n} | $\sum_{j=1}^n n_{1j}$ |
| | Y_2 | n_{21} | n_{22} | \dots | n_{2n} | $\sum_{j=1}^n n_{2j}$ |
| | | | | \dots | | |
| | Y_m | n_{m1} | n_{m2} | \dots | n_{mn} | $\sum_{j=1}^n n_{mj}$ |

Las distribuciones que aparecen en el margen inferior y en el derecho son, en realidad, las distribuciones *univariantes* de X y de Y consideradas por separado, y se denominan distribuciones marginales.

Se denomina distribución condicionada de, por ejemplo, Y para $X = x$ a la distribución que aparece en la columna i.

EJEMPLO:

Hemos recogido el número promedio de clientes que acuden al banco teniendo en cuenta 2 variables:

X = Resultado del negocio: Ingreso en efectivo, otras operaciones.

y = Día de la semana: Lunes-Jueves(L), Viernes (V), Sábados (S).

La distribución conjunta es la siguiente:

| | Resultado | | |
|-------------|-----------|-------|-----|
| | Ingreso | Otras | |
| Lunes-Juev. | 30 | 157 | 187 |
| Viernes | 61 | 228 | 289 |
| Sabados | 57 | 187 | 244 |
| | 148 | 572 | 720 |

Distribución marginal de X

| | Frecuencia | Frecuencia relativa |
|---------|------------|---------------------|
| Ingreso | 148 | 0.206 |
| Otras | 572 | 0.794 |
| Total | 720 | 1 |

Distribución marginal de Y

| | Frecuencia | Frecuencia relativa |
|-------|------------|---------------------|
| L | 187 | 0.260 |
| V | 289 | 0.401 |
| S | 244 | 0.339 |
| Total | 720 | 1 |

Distribución marginal de Y para X=Ingreso

| | Frecuencia | Frecuencia relativa |
|-------|------------|---------------------|
| L | 30 | 0.203 |
| V | 61 | 0.412 |
| S | 57 | 0.385 |
| Total | 148 | 1 |

DISTRIBUCIONES CON MAS DE 2 VARIABLES: X_1, X_2, \dots, X_m .

Los conceptos son similares a los presentados en el caso de dos variables, pero con la diferencia de que no es posible tabular los datos de la misma forma, ya que la única posibilidad es detallar los resultados para cada individuo:

Distribución conjunta de frecuencias para m variables.

| Variable | | | | | | |
|-----------|----------|----------|--|----------|--|----------|
| Individuo | X_1 | X_2 | | X_i | | X_m |
| 1 | x_{11} | x_{12} | | x_{1i} | | x_{1m} |
| 2 | x_{21} | x_{22} | | x_{2i} | | x_{2m} |
| | | | | | | |
| i | x_{i1} | x_{i2} | | x_{ij} | | x_{im} |
| | | | | | | |
| n | x_{n1} | x_{n2} | | x_{ni} | | x_{nm} |

La estructura resultante es una matriz de n filas por m columnas, denominada matriz de datos.

VECTOR DE MEDIAS.

Si consideramos cada una de las variables por separado (DISTRIBUCIÓN MARGINAL), podemos tratarla como una distribución univariable, calculando su media. Con el resultado obtenido, podemos construir un vector de dimensión m, que se denomina VECTOR DE MEDIAS y que contiene la media de cada variable:

$$\bar{X} = [\bar{X}_1, \bar{X}_2, \dots, \bar{X}_n]$$

MATRIZ DE VARIANZAS Y COVARIANZAS. COEFICIENTES DE CORRELACIÓN.**VARIANZA.**

Igual que sucede con la media, podemos calcular la varianza de cada variable por separado, denotando S_i^2 la varianza de X_i .

COVARIANZA.

Se define la covarianza entre dos variables X e Y como:

$$S_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n}$$

Nótese que $S_{xx} = S_x^2$.

La covarianza no tiene porque ser necesariamente positiva, cosa que sí sucede con la varianza.

SIGNIFICADO DE LA COVARIANZA:

La covarianza expresa el grado de variación conjunta de dos variables.

En este sentido puede suceder que:

COVARIANZA > 0 \implies Cuando aumenta una de ellas, también aumenta la otra.

COVARIANZA < 0 \implies Cuando aumenta una, la otra disminuye

COVARIANZA $= 0$ \implies No hay relación entre los aumentos de una y otra.

Estas relaciones pueden ser de mayor o menor intensidad, según la magnitud de la COVARIANZA. El mayor inconveniente de la covarianza es que su magnitud no sólo depende del grado de variación conjunta de las variables, sino también de la dispersión de cada variable. Para eliminar la influencia de este último factor, se utiliza el denominado coeficiente de correlación.

COEFICIENTE DE CORRELACIÓN:

$$r = \frac{S_{xy}}{S_x S_y}$$

Es un coeficiente adimensional cuyo valor es siempre mayor o igual que -1 y menor o igual que 1.

Si $|r| = 1 \implies$ La variación conjunta es máxima, de modo que existe una relación lineal perfecta entre las variables que puede expresarse mediante una ecuación del tipo $Y = a + bX$, por lo que podemos prescindir de una de ellas.

VARIANZA GENERALIZADA: MATRIZ DE VARIANZAS Y COVARIANZAS.

Dada una distribución *multidimensional* con m variables, se define la varianza generalizada:

$$S = \begin{bmatrix} S_{11}^2 & S_{12}^2 & S_{13}^2 & S_{1n}^2 \\ S_{21}^2 & S_{22}^2 & S_{23}^2 & S_{2n}^2 \\ S_{31}^2 & S_{32}^2 & S_{33}^2 & S_{3n}^2 \\ S_{m1}^2 & S_{m2}^2 & S_{m3}^2 & S_{mn}^2 \end{bmatrix}$$

S es una matriz simétrica semidefinida positiva.

CORRELACIÓN Y REGRESIÓN ENTRE DOS VARIABLES.**CORRELACIÓN.**

Decimos que existe una asociación, concordancia o correlación entre 2 variables cuando cierta o ciertas modalidades de una de ellas están ligadas a cierta o ciertas modalidades de la otra. También podríamos decir que, en tal caso, la modificación, en determinado sentido, de una de las variables tiende a asociarse con modificaciones en la otra.

CORRELACIÓN Y CAUSALIDAD:

La correlación entre variables no implica una relación causal, sino solamente una variación conjunta. Existen numerosos ejemplos que ilustran esta idea. Vgr: G.M. Jenkins encontró un coeficiente de correlación $r = 0.995$ entre el número de nacimientos y el número de cigüeñas en Baviera.

En muchas ocasiones la correlación se debe al influjo de una tercera variable. Vgr.: Existe una correlación positiva entre el número de teléfonos y el número de accidentes de tráfico. Esto no significa que la abundancia de teléfonos origine más accidentes de tráfico. Lo que sucede es que el número de teléfonos está relacionado con un mayor nivel de vida, y a su vez, este implica una mayor abundancia de automóviles, que es la que realmente está relacionada con el número de accidentes.

VARIABLES CUANTITATIVAS.**COEFICIENTE DE CORRELACIÓN LINEAL DE PEARSON.****DEFINICIÓN:**

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

PROPIEDADES:

a) $-1 \leq r_{xy} \leq 1$

b) $r_{xy} = r_{yx}$

c) Es invariable bajo transformaciones lineales:
$$\left. \begin{array}{l} x' = a + bx \\ y' = c + dy \end{array} \right\} \Rightarrow r_{x'y'} = r_{xy}$$

INTERPRETACIÓN:

Los valores extremos no plantean duda:

a) $|r_{xy}|=1 \rightarrow$ Hay una relación lineal perfecta, por lo que podemos calcular exactamente que valor de la segunda variable se asocia con cada uno de los de la primera, o viceversa.

b) $r_{xy} = 0 \rightarrow$ No existe ninguna relación entre las variables.

La interpretación de otros valores es muy relativa y depende de cada estudio. En general suele aceptarse la siguiente clasificación:

$0 \leq |r_{xy}| < 0.30 \rightarrow$ Relación baja entre las variables.

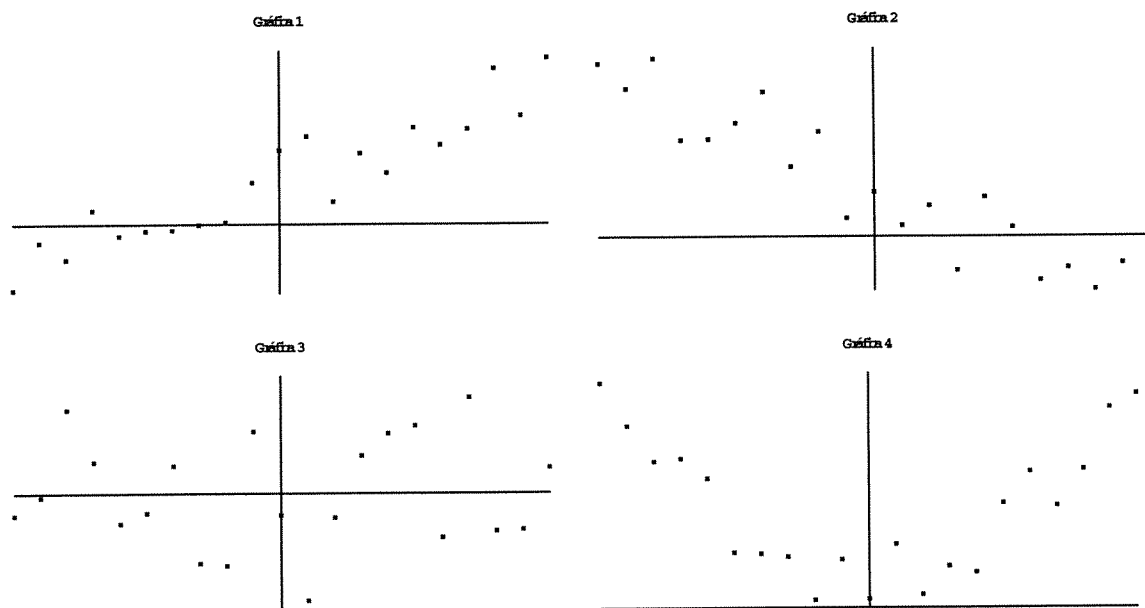
$0.30 \leq |r_{xy}| < 0.70 \rightarrow$ Relación media.

$0.70 \leq |r_{xy}| \leq 1 \rightarrow$ Relación alta.

Es importante, no obstante, comparar los resultados obtenidos con los que han publicado otros investigadores, aplicando una buena dosis de *sentido común* antes de hacer una afirmación disparatada (ejemplo de las cigüeñas).

Por otro lado, no hay que olvidar que r_{xy} mide la **correlación lineal** entre variables. Esto significa que tendrá un valor alto cuando la relación entre las variables X, Y sea tal que variaciones similares en el valor de X produzcan modificaciones aproximadamente iguales en el valor de Y.

Las gráficas siguientes corresponden a distintos valores de r_{xy} .



El coeficiente de correlación de Pearson parece funcionar bastante bien en las gráficas 1, 2 y 3, pero falla en 4, ya que en este caso $r_{xy} \approx 0$, lo cual significa que no existe relación entre las variables, cuando es obvio que hay una clara variación conjunta. Esto es debido a que $r_{xy} \approx 0$ significa, realmente, que no existe relación *lineal* entre X e Y, y esto es cierto tanto en 3 como en 4. Lo que sucede es que dos variables pueden tener una relación no necesariamente lineal. Vgr.: La mortalidad por enfermedades neoplásicas y la edad son dos variables con una clara relación curvilínea cuya gráfica es similar a 4. En este caso, el aumento de la edad produce, según el tramo en el que nos movamos, unas veces disminución de la mortalidad y otras aumento. Para medir este grado de asociación se utiliza la **razón de correlación**.

LA RAZÓN DE CORRELACIÓN DE Y SOBRE X: η_{yx} .

Suponemos que hemos realizado un estudio sobre N individuos midiendo, para cada uno de ellos los valores de dos variables: X e Y . Agrupamos las distintas modalidades de X en m clases, de modo que a cada clase le correspondan varios valores de Y , tal y como se expresa en la siguiente tabla:

| X | | | | | | |
|-----------|-------------|-------------|-----|-----------|-----|-----------|
| Individuo | Clase 1 | Clase 2 | ... | Clase j | ... | Clase m |
| 1 | Y_{11} | Y_{12} | ... | Y_{1j} | ... | Y_{1m} |
| 2 | Y_{21} | Y_{22} | ... | Y_{2j} | ... | Y_{2m} |
| ... | ... | ... | ... | ... | ... | ... |
| i | Y_{i1} | Y_{i2} | ... | Y_{ij} | ... | Y_{im} |
| ... | ... | ... | ... | ... | ... | ... |
| | $Y_{n+1,1}$ | $Y_{n+2,2}$ | ... | $Y_{n,j}$ | ... | $Y_{n,m}$ |

Por supuesto $N = \sum_{j=1}^m n_j$

El número de valores de Y que le corresponden a la clase j es n_j y la media de estos será:

$$\bar{Y}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} Y_{ij}$$

La media de todos los valores de Y es: $\bar{Y} = \frac{1}{N} \sum_{j=1}^m \sum_{i=1}^{n_j} Y_{ij}$

Si queremos asociar a cada clase de X un valor de Y , tenemos dos posibilidades:

1. Asociar a la clase j el valor \bar{Y}_j . Para evaluar el error cometido podemos calcular la suma de los errores cometidos al sustituir cada puntuación Y_{ij} por \bar{Y}_j , esto es:

$$\sum_{i=1}^{n_j} (Y_{ij} - \bar{Y}_j)$$

Como esta suma sería cero, hacemos la suma de los mismos términos al cuadrado, de forma que todos los términos sean positivos:

$$E_j^2 = \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y}_j)^2$$

A este tipo de suma se le denomina **suma de errores cuadráticos**.

2. Asociar a todas las clases el valor \bar{Y} . En este caso, la suma de errores cuadráticos será:

$$E^2 = \sum_{j=1}^m \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y})^2$$

Estas dos sumas están relacionadas de la siguiente manera:

$$E^2 = \sum_{j=1}^m \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y})^2 = \sum_{j=1}^m \sum_{i=1}^{n_j} [(Y_{ij} - \bar{Y}_j) + (\bar{Y}_j - \bar{Y})]^2 = \sum_{j=1}^m \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y}_j)^2 + \sum_{j=1}^m n_j (\bar{Y}_j - \bar{Y})^2$$

Con lo que
$$\sum_{j=1}^m n_j (\bar{Y}_j - \bar{Y})^2 = E^2 - \sum_{j=1}^m E_j^2$$

Es decir, el término $\sum_{j=1}^m n_j (\bar{Y}_j - \bar{Y})^2$ representa la diferencia entre el error cometido al actuar según la opción 2 y

el error cometido al actuar según la opción 1.

Por lo tanto, el cociente:

$$\eta_{XY}^2 = \frac{\sum_{j=1}^m n_j (\bar{Y}_j - \bar{Y})^2}{\sum_{j=1}^m \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y})^2}$$

representa la proporción de error que se elimina al tener la opción 1 en lugar de la opción 2.

A $\eta = \sqrt{\eta_{XY}^2}$ se le denomina **Razón de Correlación**.

PROPIEDADES

1. Al tratarse de una proporción $0 \leq \eta_{YX} \leq 1$.
2. Para unos mismos datos $r_{XY} \leq \eta_{YX}$.
3. Como r_{XY} mide la relación lineal entre X e Y y η_{YX} mide la relación tanto lineal como no lineal, entonces $\eta_{YX} - r_{XY}$ mide el alejamiento de linealidad en la relación entre X e Y.
4. $\eta_{YX} \neq \eta_{XY}$.
5. η_{YX} varía según el número de clases que tenemos. Si tomamos una sola clase $\rightarrow \eta_{YX} = 0$; si tomamos tantas clases como modalidades tenemos de Y $\rightarrow \eta_{YX} = 1$.

VARIABLES ORDINALES.

SIGNIFICADO DE LA RELACIÓN ENTRE 2 VARIABLES ORDINALES.

La relación entre 2 variables ordinales mide el grado de acuerdo entre 2 ordenaciones diferentes de una misma serie de valores. Vgr.: Solicitamos a dos jefes de servicio que establezcan un orden de prioridades entre 5 medidas de actuación propuestas por la Dirección de un hospital. En tal caso tendremos dos variables:

X = orden asignado por el jefe de servicio número 1.

Y = orden asignado por el jefe de servicio número 2.

| Propuesta | X | Y |
|-----------|----|----|
| A | 2º | 1º |
| B | 1º | 2º |
| C | 3º | 4º |
| D | 5º | 3º |
| E | 4º | 5º |

COEFICIENTE DE SPEARMAN: r_s .

Si llamamos $d_i = x_i - y_i$, es decir la diferencia entre el orden que ocupa el sujeto i en la ordenación X con el que ocupa en la ordenación Y, entonces:

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

En realidad $r_s = r_{xy}$, tratando los ordenes como si fueran números: 1º = 1, 2º = 2 ...

PROPIEDADES:

Son las mismas que las de r_{xy} ya que en realidad r_s es un caso particular de r_{xy} .

COEFICIENTE DE CORRELACIÓN τ DE KENDALL.

Dados dos sujetos, si el orden que tienen entre si en la ordenación X es distinto del que tienen en la Y decimos que se da una inversión. Vgr.: En el caso anterior, se da una inversión entre los elementos 1 y 2, ya que para X el 2 está antes que el 1 y para Y el 1 está antes que el 2; Sin embargo no hay inversión entre 1 y 3, ya que tanto X como Y consideran que la propuesta 1 es prioritaria sobre la 3.

Si denominamos:

P = número de no inversiones.

Q = número de inversiones.

Se define:

$$\tau = \frac{P - Q}{P + Q}$$

Nótese que $P + Q = n(n - 1)/2$

De acuerdo con la definición $-1 \leq \tau \leq 1$.

COEFICIENTE DE CORRELACIÓN γ DE GOODMAN Y KRUSKAL.

Tanto r_s como τ son útiles cuando no hay empates entre las ordenaciones. En caso de que si los haya, es mucho mas apropiado el coeficiente γ .

En este supuesto, al analizar la situación relativa de 2 elementos hay 3 resultados posibles:

No inversión: Idem al caso anterior.

Inversión: Idem al caso anterior.

Empate: Si tienen el mismo orden de preferencias en Y en X o en ambos.

Si denominamos, respectivamente, P, Q, S al número de pares que se encuentran en cada una de las situaciones anteriores, se define:

$$\gamma = \frac{P}{P + Q} - \frac{Q}{P + Q} = \frac{P - Q}{P + Q}$$

En realidad el cálculo es similar a τ , pero descartando del total los pares empatados.

También en este caso $-1 \leq \gamma \leq 1$.

VARIABLES NOMINALES.

CONCEPTOS PREVIOS.

Decimos que una variable nominal es DICOTÓMICA cuando sólo tiene 2 modalidades. Vgr.: Sexo, nacionalidad.

A veces interesa tratar variables que no son dicotómicas como si lo fueran. Para ello basta con agrupar todas las modalidades en dos clases. En este caso diremos que la variable está *dicotomizada*.

La tabulación general de una variable dicotómica queda recogida en la siguiente tabla de contingencia:

| Variable Y | Variable X | | |
|----------------|----------------|----------------|-----------|
| | X ₁ | X ₂ | |
| Y ₁ | a | b | a+b |
| Y ₂ | c | d | c+d |
| | a+c | b+d | n=a+b+c+d |

RELACIÓN ENTRE VARIABLES DICOTÓMICAS.

COEFICIENTE Q DE YULE.

Si no hay relación entre X e Y $\Rightarrow \frac{a}{c} = \frac{b}{d} \rightarrow ad = bc \rightarrow ad - bc = 0$.

Cuanto mayor sea $ad - bc$ mayor será la relación. Para evitar que esta diferencia pueda crecer ilimitadamente formamos el cociente:

$$Q = \frac{ad - bc}{ad + bc}$$

Nuevamente se verifica $-1 \leq Q \leq 1$

COEFICIENTE DE CORRELACIÓN ϕ .

Se obtiene al calcular r_{xy} para dos variables dicotómicas, y vale:

$$\phi = \frac{cb - ad}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$$

RELACIÓN ENTRE VARIABLES NOMINALES NO DICOTÓMICAS.

En este caso la tabla de contingencia será del tipo

| Y | X | | | | | | |
|-------|----------|----------|-----|----------|-----|----------|----------|
| | X_1 | X_2 | ... | X_j | ... | X_m | |
| Y_1 | n_{11} | n_{12} | ... | n_{1j} | ... | n_{1m} | $n_{1.}$ |
| Y_2 | n_{21} | n_{22} | ... | n_{2j} | ... | n_{2m} | $n_{2.}$ |
| ... | ... | ... | ... | ... | ... | ... | ... |
| Y_i | n_{i1} | n_{i2} | ... | n_{ij} | ... | n_{im} | $n_{i.}$ |
| ... | ... | ... | ... | ... | ... | ... | ... |
| Y_n | n_{n1} | n_{n2} | ... | n_{nj} | ... | n_{nm} | $n_{n.}$ |
| | $n_{.1}$ | $n_{.2}$ | ... | $n_{.j}$ | ... | $n_{.m}$ | |

NOTA: La notación $n_{i.}$ y $n_{.j}$ es muy común para denotar las frecuencias marginales de X e Y.

COEFICIENTE χ^2 .

No debe confundirse con la distribución χ^2 , aunque cuando n es muy grande sus distribuciones se aproximan bastante.

Su cálculo se basa en la comparación de frecuencias teóricas (las que cabe esperar supuesto que no exista ninguna relación entre las variables) y frecuencias experimentales (las realmente obtenidas).

Así, para la celda ij tenemos:

Frecuencia experimental: $f_{ij} = n_{ij}/N$

Si no hubiese relación entre las variables, entonces $n_{ij}/n_{.j} = n_{i.}/N$ o lo que es lo mismo, $n_{ij}/n_{i.} = n_{.j}/N$

A partir de cualquiera de ellas, llegamos a la conclusión de que, para que no haya relación entre las variables, debe verificarse:

$$n_{ij} = n_{.j} \cdot n_{i.} / N$$

De esta forma definimos la

Frecuencia Teórica: $F_{ij} = n_{.j} \cdot n_{i.} / N$

Con estas premisas, podemos definir:

$$\chi^2 = \sum_{i=1}^n \sum_{j=1}^m \frac{(f_{ij} - F_{ij})^2}{F_{ij}} = \sum_{i=1}^n \sum_{j=1}^m \frac{n_{ij}^2}{F_{ij}} - N$$

Un inconveniente de este coeficiente es que no está acotado. Es decir, puede tener cualquier valor mayor que 0, y este valor será tanto mayor cuanto más crezca N. Para evitarlo definimos el *coeficiente de contingencia C*.

COEFICIENTE DE CONTINGENCIA C.

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

De este modo $0 < C < 1$.

El valor de C depende de n y m, de modo que, para unos mismos datos, es máximo si $m = n$.

CORRELACIÓN ENTRE VARIABLES CONTINUAS DICOTOMIZADAS.

Todos los coeficientes que se utilizan derivan del coeficiente r_{xy} de Pearson, aplicado en condiciones particulares, y por lo tanto tienen sus mismas propiedades.

COEFICIENTE DE CORRELACIÓN BISERIAL PUNTUAL r_{bp} .

Se utiliza cuando tenemos una variable X continua y otra Y dicotómica. Para calcularlo asignamos a una de las modalidades de esta última el valor 0 y a otra el 1, y aplicamos la definición de r_{xy} . El valor obtenido será:

$$r_{bp} = \frac{\bar{X}_p - \bar{X}_q}{s_y} \sqrt{pq}$$

p = proporción de personas que tienen la modalidad 1 de las dos posibles que tiene Y.

q = Proporción de personas que tienen la modalidad 2 de las dos posibles que tiene Y.

X = Media de los valores de X que tienen la modalidad 1 de Y.

\bar{X} = Media de los valores de X que tienen la modalidad 2 de Y.

Otros coeficientes son el coeficiente de correlación biserial r_{by} y el coeficiente de correlación tetracórica.

COEFICIENTE DE CORRELACIÓN BISERIAL PUNTUAL r_{bp} .

Se utiliza cuando tenemos una variable X continua y otra Y dicotómica. Para calcularlo asignamos a una de las modalidades de esta última el valor 0 y a otra el 1, y aplicamos la definición de r_{xy} . El valor obtenido será:

$$r_{bp} = \frac{\bar{X}_p - \bar{X}_q}{s_y} \sqrt{pq}$$

p = proporción de personas que tienen la modalidad 1 de las dos posibles que tiene Y.

q = Proporción de personas que tienen la modalidad 2 de las dos posibles que tiene Y.

\bar{X}_p = Media de los valores de X que tienen la modalidad 1 de Y.

\bar{X}_q = Media de los valores de X que tienen la modalidad 2 de Y.

Otros coeficientes son el coeficiente de correlación biserial r_{by} y el coeficiente de correlación tetracórica.

REGRESIÓN LINEAL.

CONCEPTO.

La simple constatación de la existencia de una asociación entre variables no permite realizar predicciones sobre los valores que adoptará una variable al asignar valores a la otra.

Para ello es necesario establecer una relación funcional entre los mismos, encontrando una ecuación que las ligue. El término regresión debe ser interpretado en este contexto como predicción, pronóstico o estimación.

La relación funcional más simple, desde el punto de vista del análisis matemático, es la relación *lineal*. Es decir aquella que viene dada por la ecuación del tipo : $Y = a + bX$, que corresponde a la ecuación de una recta en dos dimensiones.

CRITERIO DE MÍNIMOS CUADRADOS.

Dadas dos variables cuantitativas X, Y, vamos a tratar de encontrar una ecuación del tipo $Y=a+bX$ que nos permita aproximar los valores de Y a partir de los de X.

Existen infinitas soluciones capaces de satisfacer la condición expresada en el párrafo anterior, por lo que es necesario añadir una restricción adicional para que la solución quede perfectamente determinada. Con esta finalidad, exigimos también que el error cometido al realizar la predicción de Y sea mínimo.

Si llamamos Y_i al valor de la variable Y que tiene asociado el valor X_i de la variable X, e Y' al valor que resulta de sustituir en la ecuación $Y = a + bX$, el valor de X_i , es decir: $Y' = a + bX_i$, entonces el error cometido en la predicción será: $e_i = Y - Y' = Y - a - bX_i$.

Para evaluar el error total no podemos sumar e_i simplemente, ya que obtendríamos un valor lejano a la realidad al acumular desviaciones positivas y negativas.

Para evitar este inconveniente sumamos e_i^2 , es decir, tomamos la *suma de errores cuadráticos*:

$$e = \sum_{i=1}^n (Y_i - Y'_i)^2 = \sum_{i=1}^n (Y_i - a - bX_i)^2$$

El criterio de mínimos cuadrados consiste en tomar la suma de errores cuadráticos como medida del error. Esto significa que debemos buscar los valores a y b que hacen que la suma de errores cuadráticos sea mínima.

ECUACIONES DE REGRESIÓN LINEAL.

Traduciendo el criterio anterior a términos matemáticos nos queda:

$$\frac{\partial}{\partial a} \sum_{i=1}^n (Y_i - a - bX_i)^2 = 0 \quad \text{y} \quad \frac{\partial}{\partial b} \sum_{i=1}^n (Y_i - a - bX_i)^2 = 0$$

Operando llegamos a:

$$b = \frac{N \sum_{i=1}^N X_i Y_i - \sum_{i=1}^N X_i \sum_{i=1}^N Y_i}{N \sum_{i=1}^N X_i^2 - \left(\sum_{i=1}^N X_i \right)^2} = \frac{s_{xy}}{s_x^2} \quad \text{y} \quad a = \bar{Y} - b\bar{X}$$

Luego la recta será:

$$Y - \bar{Y} = \frac{s_{xy}}{s_x^2} (X - \bar{X})$$

Del mismo modo podríamos habernos planteado la ecuación de regresión de X sobre Y .

En este caso, si suponemos que $X = a' + b'Y$, tenemos: $b' = \frac{S_{xy}}{S_y^2}$ y $a' = \bar{X} - b'Y$

Nótese que el coeficiente de correlación lineal r_{xy} es la media geométrica de los coeficientes b y b' :

$$r_{xy} = \frac{S_{xy}}{S_x S_y} = \sqrt{\frac{S_{xy}^2}{S_x^2 S_y^2}} = \sqrt{\frac{S_{xy}}{S_x} \frac{S_{xy}}{S_y}} = \sqrt{bb'}$$

En general, los valores que obtenemos al sustituir en la ecuación de regresión los X_i no coinciden exactamente con los valores Y_i que están asociados en la distribución original. Para diferenciarlos, llamaremos $Y_i' = a + bX_i$.

Queda así definida una nueva variable Y' cuyos estadísticos principales son:

$$\bar{Y}' = \bar{Y} \quad \text{y} \quad S_{Y'}^2 = S_{a+bX}^2 = b^2 S_x^2 = \left(\frac{S_{xy}}{S_x^2} \right)^2 S_x^2 = \frac{S_{xy}^2}{S_x^2} = \frac{S_{xy}}{S_x} S_{xy} = b S_{xy} = r_{xy}^2 S_y^2$$

La última relación nos conduce también a: $r_{xy}^2 = \frac{S_{Y'}^2}{S_y^2}$

cuyo significado comprenderemos más tarde.

BONDAD DEL AJUSTE. COEFICIENTE DE DETERMINACIÓN.

La definición de Y' nos lleva de forma inmediata a considerar otra variable más: la diferencia entre Y e Y' , es decir, la diferencia entre el valor experimental y el valor estimado. La denotaremos $e = Y - Y'$ y representa el error cometido en cada predicción.

Sus principales características son:

- 1) $\bar{e} = \bar{Y} - \bar{Y}' = 0$
- 2) $S_{eY'} = 0$
- 3) $S_e^2 = S_Y^2 - S_{Y'}^2$

De 1) se desprende que $\sum e_i = 0$, por lo que, tal y como habíamos anticipado, no podemos tomar $\sum e_i$ como medida de la bondad del ajuste.

La suma de errores cuadráticos $\sum e_i^2$ no presenta este inconveniente, pero si el de depender del número de observaciones. Para evitar esta dependencia tomamos el *error cuadrático medio* (ECM), que se calcula:

$$ECM = \frac{\sum e_i^2}{N} \geq 0$$

El ECM o su raíz cuadrada, que se denomina *error de regresión*, son inversamente proporcionales a la bondad del ajuste: cuanto mayor es ECM peor será el ajuste de los datos a una recta, mientras que cuanto más se aproxime a cero, más perfecta será la relación lineal entre las variables.

De 3) se deduce una relación fundamental: $S_Y^2 = S_{Y'}^2 + S_e^2$

Es decir la varianza de Y tiene dos componentes: una debida a la relación lineal entre las variables y que está contenida en el término $S_{Y'}^2$; y otra que es la *varianza residual* (S_e^2) y que contiene la variabilidad que no es capaz de explicar el modelo lineal.

Como $\bar{e} = 0$, entonces $S_e^2 = ECM$, y de ahí que el ECM sea un error estimado de la bondad del ajuste, ya que equivale a la varianza residual. Cuanto mayor sea la varianza residual, mayor será la parte de la variabilidad de Y que es incapaz de explicarse por la relación lineal entre X e Y . Esto puede deberse a que no existe ninguna relación entre las variables o a que ésta no sea lineal.

Las últimas consideraciones acerca de la relación $S_Y^2 = S_{Y'}^2 + S_e^2$ nos lleva a una interpretación del coeficiente de correlación r_{xy} como una medida de la raíz cuadrada de la proporción de la varianza de Y que es capaz de explicar el modelo. En efecto, recordemos que:

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

por lo que, teniendo en cuenta el significado atribuido a $S_{Y'}^2$ y S_Y^2 queda clara la interpretación realizada de r_{xy} .

Como la proporción es más directa al tomar $(r_{xy})^2$ que al tomar r_{xy} , definimos el *coeficiente de determinación* R como:

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

Lógicamente $0 \leq R \leq 1$, ya que $S_Y^2 \geq S_{Y'}^2$. De esta forma, podemos decir que si hemos encontrado un valor de $r_{xy} = 0.9$, entonces, aproximadamente el 81% de la variación de Y puede explicarse por su relación lineal con X .

REGRESIÓN NO LINEAL.

REGRESIÓN DE LA MEDIA.

Un valor de r_{xy} bajo no excluye una fuerte asociación entre las variables, ni tampoco la posibilidad de encontrar una relación funcional que permita predecir los valores de una de ellas conociendo los de la otra.

Al estudiar una distribución bivariable podemos encontrar que cada valor X tiene asociados varios valores de Y : $(Y_{1i}, Y_{2i}, \dots, Y_{nii})$, bien de forma natural, o bien porque, intencionadamente hemos agrupado las modalidades de X en clases, tal y como se explicó al hablar de la razón de correlación de Pearson. Bajo estas circunstancias, existe un criterio para asociar a cada X_i un valor Y'_i , de modo que la varianza residual sea la menor posible. Este procedimiento consiste en asociar a cada X_i la media de los valores de Y que tiene asociados.

Utilizando la notación usada al hablar de η_{yx} , podemos decir que establecemos como modelo de regresión la relación funcional:

$$Y'_j = f(X_j) = \bar{Y}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} Y_{ij}$$

La proporción de la varianza de Y que queda explicada por el modelo de regresión viene dada por $(\eta_{yx})^2$ y ya dijimos que era siempre mayor o igual que r_{xy} . A este modelo se le denomina *regresión de la media*.

Podemos preguntarnos porqué si la regresión de la media es la que produce un mejor ajuste, no se utiliza siempre, olvidándonos de los demás modelos. La respuesta es que la relación $Y'_j = f(X_j)$ no siempre es analítica, es decir, no siempre puede expresarse mediante una ecuación. De hecho, la regresión de la media, por si sola no aporta ningún procedimiento para determinar la función analítica que más se aproxima a $f(X_j)$.

Podemos establecer una estrategia óptima para buscar la relación funcional más adecuada entre dos variables:

1.- Determinar $(\eta_{yx})^2$. Si tiene un valor bajo, es mejor no probar con ningún modelo, ya que cualquiera de ellos nos dará un coeficiente de determinación todavía menor.

2.- Determinar R^2 . Si hay poca diferencia entre R^2 y $(\eta_{yx})^2$, entonces el modelo lineal es satisfactorio. Si la diferencia es grande, es necesario buscar un modelo no lineal. Esta búsqueda puede facilitarse si se tiene en cuenta el *diagrama de dispersión* de los datos, ya que la forma en que se distribuyen los puntos en el plano puede sugerirnos la curva más apropiada. Una vez elegido el modelo no lineal más conveniente es preciso volver a evaluar el ajuste, hasta conseguir un coeficiente de determinación lo más aproximado posible a $(\eta_{yx})^2$.

MODELOS DERIVADOS DEL LINEAL.

FUNCIÓN POLINÓMICA: $Y = a_0 + a_1 X + a_2 X^2 + \dots + a_n X^n$

Puede resolverse aplicando el criterio de mínimos cuadrados de forma similar a la ecuación lineal.

También puede resolverse como un caso *multilineal* con $X_i = X^i$.

FUNCIÓN POTENCIAL: $Y = aX^b$

Podemos tomar logaritmos con lo que: $\lg(Y) = \lg(a) + b \cdot \lg(X)$. Entonces, Construimos las variables: $Y' = \lg(Y)$ y $X' = \lg(X)$ y buscamos la regresión lineal entre ellas: $Y' = a' + b'X'$.

Una vez determinadas a' y b' , podemos calcular a y b teniendo en cuenta que: $a' = \lg(a) \rightarrow a = 10^{a'}$

FUNCIÓN EXPONENCIAL: $Y = ab^X$

Operando como en el caso anterior: $\lg(Y) = \lg(a) + X \cdot \lg(b)$. Haciendo, en este caso, $Y' = \lg(Y)$ calculamos a' y b' de modo que $Y' = a' + b'X$. Una vez calculados despejamos a y b mediante la relación: $a = 10^{a'}$ y $b = 10^{b'}$.

FUNCIÓN LOGARÍTMICA: $Y = a + b \lg(X)$

Basta con hacer el cambio $X' = \lg(X)$ y tratarlo como una ecuación lineal.

MODELOS NO DERIVABLES DEL LINEAL.

Son ecuaciones que no pueden reducirse al modelo lineal de forma analítica. Esto obliga a que el tratamiento matemático sea diferente, utilizando métodos distintos al de mínimos cuadrados.

Los modelos de este tipo más utilizados son los que consideran una relación funcional de este tipo:

$$1. Y = a + bc^X$$

$$2. Y = \frac{c}{1 + be^{-aX}}$$

EL PROBLEMA DE LA PREDICCIÓN.

La búsqueda de un modelo de regresión que se ajuste a los datos estudiados tiene, según dijimos al comienzo, una clara justificación: poder realizar predicciones fiables sobre los valores que adoptará la *variable explicada* (Y) cuando la *variable explicativa* (X) tome un valor determinado. Sin embargo, hay que hacer dos importantes aclaraciones:

1. La construcción del modelo la realizamos basándonos en un grupo de datos que en ningún caso constituye la totalidad de la población (si fuera así no tendría sentido la predicción), sino sólo una muestra de ésta.

Por tanto se tratará de una estimación del modelo real, y la fiabilidad o grado de aproximación a este modelo, dependerá de la metodología empleada en el tratamiento de los datos.

2. Toda predicción supone un proceso inferencial que, en consecuencia, tendrá asociada una determinada precisión o fiabilidad. Esta depende de la seguridad o margen de confianza con el que estimemos los parámetros del modelo, es decir, a y b .

En particular, por tratarse de un modelo lineal, es muy importante la incertidumbre con la que estimemos el valor de b , ya que esta se multiplica por el valor de X , y por lo tanto, cuanto mayor sea X , mayor será el error cometido en la predicción.

Esto se traduce en la práctica en que cuando la predicción se realiza sobre datos que se encuentran en el mismo rango que los estudiados (*interpolación*), obtenemos valores mas fiables que cuando se trata de datos que exceden ese rango (*extrapolación*).

Aunque es problema de la Estadística Inferencial, un modelo de regresión no está completo si no se especifica junto a los parámetros encontrados, cual es su margen de confianza. Este margen se mide con la desviación típica y, aunque no lo deduzcamos, vale para b :

$$S_b = \sqrt{\frac{N}{N-2}} \frac{S_e}{\sqrt{\sum_{i=1}^N (X_i - \bar{X})^2}} = \frac{1}{\sqrt{N-2}} \frac{S_e}{S_X}$$

Siendo:

N = número de pares de datos.

$$S_e = \frac{\sum_{i=1}^N (e_i - \bar{e})^2}{N} = \frac{\sum_{i=1}^N (Y_i - \hat{Y}_i)^2}{N}$$

Coeficientes de correlación.

Correlación entre variables cuantitativas

Lineal de Pearson

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

Razón de Correlación

$$\eta_{XY}^2 = \frac{\sum_{j=1}^m n_j (\bar{Y}_j - \bar{Y})^2}{\sum_{j=1}^m \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y})^2}$$

Correlación entre variables cualitativas. Variables ordinales.

Spearman

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

Goodman y Kruskal

$$\gamma = \frac{P}{P+Q} - \frac{Q}{P+Q} = \frac{P-Q}{P+Q}$$

Kendall

$$\tau = \frac{P-Q}{P+Q}$$

Variables nominales dicotómicas.

Yule

$$Q = \frac{ad - bc}{ad + bc}$$

Phi

$$\varphi = \frac{cb - ad}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$$

No dicotómicas

Chi Cuadrado

$$\chi^2 = \sum_{i=1}^n \sum_{j=1}^m \frac{(f_{ij} - F_{ij})^2}{F_{ij}} = \sum_{i=1}^n \sum_{j=1}^m \frac{n_{ij}^2}{F_{ij}} - N$$

Coeficiente de contingencia:

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

Correlación entre variables cualitativas y cuantitativas.

$$r_{bp} = \frac{\bar{X}_p - \bar{X}_q}{s_y} \sqrt{pq}$$

Apuntes de Probabilidad

Asignatura: Complementos de Matemáticas

Profesor: Dr. Manuel J. Galán Moreno

E.T.S.A.

CONCEPTOS FUNDAMENTALES DEL CÁLCULO DE PROBABILIDADES.

I. CONCEPTOS PREVIOS.

1. SUCESOS.

1. SUCESO ELEMENTAL.

Se denomina Suceso Elemental de un experimento aleatorio a cada uno de los posibles resultados de dicho experimento que no pueden descomponerse en resultados más simples. Al realizar un experimento aleatorio siempre ocurre uno de los sucesos elementales. Al ocurrir un suceso elemental, quedan excluidos todos los demás.

Por ejemplo, si consideramos el experimento aleatorio: *resultado de sacar una carta de un baraja* y los sucesos:

S_1 = Sacar as de oros.

S_2 = Sacar un as.

Sólo el S_1 es elemental, ya que S_2 puede descomponerse en sacar el as de oros, el as de copas, el de espadas o el de bastos.

2. SUCESO.

Un suceso de un experimento aleatorio es cualquier composición de los sucesos elementales de dicho experimento. Vgr.: El experimento aleatorio *Estado civil de una persona* tiene como sucesos elementales:

S_1 = Soltero.

S_2 = Casado.

S_3 = Divorciado.

S_4 = Viudo.

A partir de éstos podemos formar:

$S_5 = \{S_1, S_2\}$ = Soltero o casado.

$S_6 = \{S_1, S_3\}$ = Soltero o divorciado.

$S_7 = \{S_1, S_4\}$ = Soltero o viudo.

Y de la misma forma:

$S_8 = \{S_2, S_3\}$; $S_9 = \{S_2, S_4\}$; $S_{10} = \{S_3, S_4\}$; $S_{11} = \{S_1, S_2, S_3\}$; $S_{12} = \{S_1, S_2, S_4\}$; $S_{13} = \{S_1, S_3, S_4\}$; $S_{14} = \{S_2, S_3, S_4\}$.

Hay además dos sucesos triviales que son:

Suceso imposible: $\emptyset = 0$ (que no suceda nada).

Suceso seguro: $\Omega = \{S_1, S_2, S_3, S_4\}$ (que suceda cualquier cosa)

2. ESPACIO MUESTRAL.

Se denomina **Espacio Muestral** (Ω) de un experimento aleatorio al conjunto de *todos los sucesos elementales* del mismo. Equivale al suceso seguro definido anteriormente: $\Omega = \{S_1, S_2, S_3, S_4\}$. Puede ser:

- a) **Finito**: Si el número de elementos que tiene Ω está acotado. Vgr.: Cualquiera de los dos casos anteriores.
- b) **Infinito numerable**: Cuando, a pesar de tener infinitos elementos, **no** siempre es posible intercalar uno entre dos dados. Vgr.: N° de veces que hay que lanzar un dado hasta que salga un 6. Este número, en teoría es ilimitado, pero nunca puede estar entre 5 y 6 (siempre será entero).
- c) **Infinito no numerable**: Cuando Ω tiene infinitos elementos, y además siempre se puede intercalar uno entre dos cualesquiera de ellos. Vgr.: Tiempo de espera hasta que un paciente que acude a urgencias es atendido.

Otra clasificación que podríamos realizar sería:

- a) **Discreto**: Si es finito o infinito numerable.
- b) **Continuo**: Si es infinito no numerable.

El conjunto de todos los posibles sucesos de un experimento aleatorio será, lo que en teoría de conjuntos se denomina $\mathbf{P}(\Omega)$ (Conjunto de las partes del espacio muestral). En el ejemplo del estado civil será: $\mathbf{P}(\Omega) = \{\emptyset, S_1, S_2, S_3, \dots, S_{13}, S_{14}, \Omega\}$. En consecuencia, si Ω consta de n sucesos elementales, podemos definir un total de 2^n posibles sucesos.

3. ÁLGEBRA DE SUCESOS.

Dado un espacio muestral Ω y dados $S_1, S_2, \dots, S_i \in \Omega$, podemos definir las siguientes operaciones entre sucesos:

1. SUMA O UNIÓN DE SUCESOS: $S_1 + S_2$.

Es el suceso compuesto que resulta de que ocurra S_1 o bien S_2 .

2. PRODUCTO O INTERSECCIÓN DE SUCESOS: $S_1 \cdot S_2$.

Es el suceso que resulta al exigir que ocurra S_1 y S_2 simultáneamente.

3. COMPLEMENTARIO DE UN SUCESO: \bar{S}

Es el suceso que se verifica *si y sólo si* no se verifica S .

Muchas veces no estamos interesados en todos los sucesos posibles ($\mathbf{P}(\Omega)$) y preferimos limitarnos a una parte de ellos ($\mathbf{F} \subset \mathbf{P}(\Omega)$). Este subconjunto de sucesos no lo podemos elegir de forma totalmente arbitraria, sino que, para poder definir una medida de probabilidad sobre él, tal y como veremos mas adelante, es necesario que se cumplan 2 requisitos:

1. Si $S_1, S_2 \in \mathbf{F} \Rightarrow (S_1 + S_2) \in \mathbf{F}$
2. Si $S \in \mathbf{F} \Rightarrow \bar{S} \in \mathbf{F}$

Es decir, es necesario que \mathbf{F} sea lo que se denomina un **Álgebra aditiva o σ -Álgebra**. Por ejemplo, en el experimento del Estado Civil pueden interesarnos solamente los sucesos:

A = Casado.

B = No casado.

Con lo que $\mathbf{F} = \{\emptyset, A, B, \Omega\}$, ya que \emptyset y Ω se deben incluir siempre para que pueda ser un álgebra aditiva. Efectivamente, podemos comprobar que \mathbf{F} cumple los requisitos anteriores, ya que:

1. $A + B = \Omega \in \mathbf{F}$
2. $\bar{A} = B \in \mathbf{F}$ y $\bar{B} = A \in \mathbf{F}$

II. DISTINTAS DEFINICIONES DE PROBABILIDAD.

1. DEFINICIÓN DE LAPLACE E INTERPRETACIÓN FRECUENCIALISTA.

Históricamente es la primera que surgió y corresponde a la idea intuitiva de:

$$\text{Probabilidad} = \frac{\text{Casos favorables}}{\text{Casos posibles}}$$

Se aplica fácilmente cuando Ω es *finito y homogéneo o simétrico*, es decir: cuando todos los sucesos elementales de los que se compone Ω tienen la misma probabilidad de ocurrir. Vgr.: lanzamiento de una moneda o de un dado no cargados.

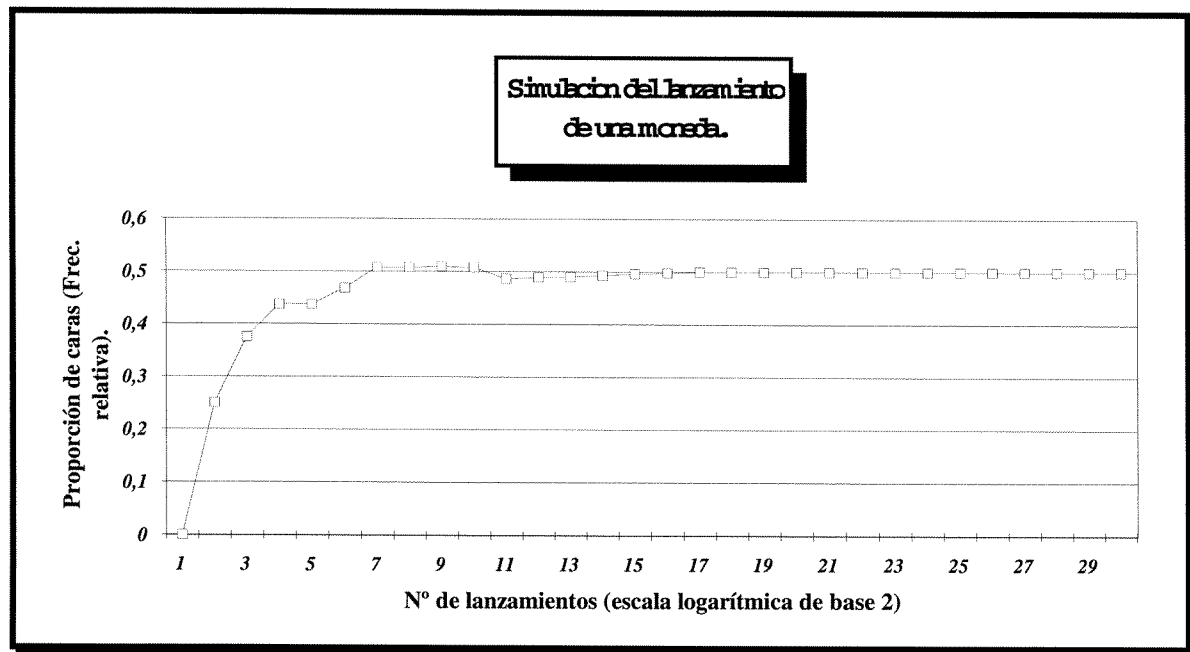
Sin embargo falla cuando Ω no es homogéneo o no es finito. Este último caso corresponde a preguntas como: ¿Cual es la probabilidad de que al nacer un niño éste sea varón? ¿Cual es la probabilidad de contraer determinada enfermedad? ¿Cual es la probabilidad de que acudan más de 100 pacientes a urgencias?...

Para resolverlas nos apoyamos en la **Ley de la Regularidad Estadística** o **Ley de los Grandes Números**, que afirma que cuando un experimento aleatorio se repite indefinidamente, la frecuencia relativa con la que se da un determinado suceso tiende a estabilizarse en torno a un valor. Ese valor en el que se estabiliza la frecuencia relativa es el que tomaremos como probabilidad del suceso.

Para ilustrar este concepto, aparece a continuación un resumen de los resultados obtenidos al simular mediante un ordenador el lanzamiento de una moneda y contar el número de veces que sale cara. Cuando el número de lanzamientos es pequeño, la frecuencia relativa del suceso *salir cara* oscila entorno al valor 0.5 con un margen muy amplio, pero cuando crece el número de tiradas este margen se va reduciendo hasta llegar a ser despreciable. En consecuencia, la diferencia entre la frecuencia registrada y la frecuencia que cabría esperar puede hacerse tan pequeño como queramos, con solo aumentar el número de tiradas.

Tabla 1: Simulación por ordenador del lanzamiento de una moneda.

| Nº de lanzamientos | Nº de caras | Frecuencia relativa | Diferencia con la frecuencia teórica | Nº de lanzamientos | Nº de caras | Frecuencia relativa | Diferencia con la frecuencia teórica |
|--------------------|-------------|---------------------|--------------------------------------|--------------------|-------------|---------------------|--------------------------------------|
| 2 | 0 | 0 | 0.5 | 65.536 | 32.645 | 0.498 | 0.002 |
| 4 | 1 | 0.25 | 0.25 | 131.072 | 65.535 | 0.499992 | 0.000008 |
| 8 | 3 | 0.38 | 0.125 | 262.144 | 131.071 | 0.499996 | 0.000004 |
| 16 | 7 | 0.437 | 0.063 | 524.288 | 262.143 | 0.499998 | 0.000002 |
| 32 | 14 | 0.437 | 0.063 | 1.048.576 | 524.287 | 0.4999990 | 0.000001 |
| 64 | 30 | 0.47 | 0.03 | 2.097.152 | 1.048.575 | 0.4999995 | 0.0000005 |
| 128 | 65 | 0.508 | 0.008 | 4.194.304 | 2.097.151 | 0.4999998 | 0.0000002 |
| 256 | 130 | 0.508 | 0.008 | 8.388.608 | 4.194.303 | 0.49999990 | 0.0000001 |
| 512 | 261 | 0.51 | 0.01 | 16.777.216 | 8.388.607 | 0.49999994 | 0.00000006 |
| 1.024 | 520 | 0.508 | 0.008 | 33.554.432 | 16.777.215 | 0.49999997 | 0.00000003 |
| 2.048 | 996 | 0.486 | 0.014 | 67.108.864 | 33.554.431 | 0.499999985 | 0.000000015 |
| 4.096 | 2.005 | 0.49 | 0.01 | 134.217.728 | 67.108.863 | 0.499999993 | 0.000000007 |
| 8.192 | 4.016 | 0.49 | 0.01 | 268.435.456 | 134.217.727 | 0.499999996 | 0.000000004 |
| 16.384 | 8.075 | 0.493 | 0.007 | 536.870.912 | 268.435.455 | 0.499999998 | 0.000000002 |
| 32.768 | 16.273 | 0.497 | 0.003 | 1.073.741.824 | 536.870.911 | 0.4999999990 | 0.000000001 |



2. DEFINICIÓN AXIOMÁTICA.

La definición frecuentista no deja de presentar problemas en su aplicación práctica, ya que, al no ser posible la repetición ilimitada de un experimento, ¿Cómo saber cuando hemos realizado un número suficiente de repeticiones como para haber encontrado la frecuencia relativa correcta?.

Como consecuencia del desarrollo del formalismo matemático y con la intención de soslayar los inconvenientes anteriores, se planteó en los años 30 la formulación axiomática de la probabilidad. Este enfoque evita dar una definición conceptual y sólo se refiere a las propiedades elementales que debe cumplir la definición de una **medida de probabilidad** sobre un conjunto (estas propiedades son, precisamente, las que, de forma natural, caracterizan a la frecuencia relativa) y una vez aceptadas como axiomas, nos conducen, a partir de un tratamiento matemático riguroso, a un conjunto mucho mayor de propiedades y consecuencias que, a primera vista, nunca hubieran parecido evidentes.

1. DEFINICIÓN AXIOMÁTICA DE PROBABILIDAD.

Dado un Espacio Muestral Ω y un conjunto de sucesos del mismo $\mathbf{F} \subset \mathbf{P}(\Omega)$, que constituyen un álgebra aditiva, se denomina **Probabilidad** a cualquier forma de asignar a cada suceso $S_I \in \mathbf{F}$ un valor numérico $P(S_I)$, siempre que se verifiquen las siguientes propiedades:

1. $P(S_I) \geq 0$
2. $P(\Omega) = 1$
3. Dados S_I y $S_2 \in \mathbf{F}$, tales que $S_I \cdot S_2 = \emptyset$, entonces: $P(S_I + S_2) = P(S_I) + P(S_2)$

Partiendo de estos axiomas, podemos deducir, como primeras consecuencias, las siguientes propiedades:

1. $P(S) = 1 - P(\bar{S})$
2. $P(S) \leq 1$
3. $S_I, S_2 \in \mathbf{F} \Rightarrow P(S_I + S_2) = P(S_I) + P(S_2) - P(S_I \cdot S_2)$

3. INTERPRETACIÓN BAYESIANA O SUBJETIVA.

Es una forma mucho más operativa de definir la probabilidad, que consiste en utilizar la información de la que un sujeto dispone, para realizar una apreciación personal de la probabilidad de un suceso. Obviamente, si el sujeto conoce la frecuencia relativa del suceso, y su actuación es coherente, utilizará esta como probabilidad. Si no conoce la frecuencia relativa, utilizará cualquier otro tipo de información para dar una estimación de la probabilidad. Es lo que hacemos cuando, por ejemplo, decimos: Hay una probabilidad del 60% de que llueva este fin de semana; tenemos un 50% de posibilidades de que salga adelante este proyecto...

Evidentemente, la estimación de la probabilidad variará en función del sujeto y de la información de la que éste disponga. Por eso sería más correcto decir $P(A/I)$, que se lee: Probabilidad de que suceda A dado que disponemos de la información I. (Este tipo de notación se usa para la probabilidad condicionada, tal como veremos más adelante).

La Estadística clásica se apoya exclusivamente en los datos para estimar las características de la población, mientras que la Estadística Bayesiana utiliza además la información basada en el grado de creencia que tiene el experimentador acerca de esas características. El análisis de los datos permite variar esa creencia y el resultado puede servir de base para una nueva estimación.

4. CALCULO PRÁCTICO DE PROBABILIDADES.

La raíz del problema está en la asignación de probabilidades a los sucesos elementales, ya que a partir de éstos se puede calcular la probabilidad de cualquier suceso compuesto, aplicando las reglas que hemos visto en el apartado 2. La determinación de la probabilidad de un suceso elemental puede hacerse:

1. Estudiando la frecuencia relativa mediante la repetición del experimento hasta ver que ésta se estabiliza.
2. Deduciéndolo de la naturaleza del experimento. El caso más simple es el de Ω homogéneo o simétrico, ya que todos los sucesos elementales serán equiprobables, por lo que todos tendrán probabilidad $1/N$, siendo N el número de sucesos elementales de Ω .
3. En este caso la probabilidad de un suceso compuesto obedece a la conocida fórmula de:

$$P(S) = \frac{\text{Casos favorables}}{\text{Casos posibles}} = \frac{N^{\circ} \text{ de sucesos elementales de } S}{N^{\circ} \text{ de sucesos elementales de } \Omega}$$

4. Combinando la información acerca de la naturaleza del experimento con los resultados de éste.

III. TEOREMAS FUNDAMENTALES DEL CALCULO DE PROBABILIDADES.

1. PROBABILIDAD CONDICIONADA.

1. CONCEPTO:

La probabilidad de A dado B, o condicionada a la ocurrencia de B, es la frecuencia relativa con la que se da el suceso A cuando se ha dado B. Se denota $P(A/B)$ y no debe confundirse con $P(AB)$.

$P(AB)$ es la probabilidad de que se den simultáneamente A y B referida al espacio muestral Ω .

$P(A/B)$ es lo mismo, pero tomando B como espacio muestral de referencia.

Por ejemplo, si consideramos el experimento aleatorio resultado de lanzar un dado, entonces $\Omega = \{1, 2, 3, 4, 5, 6\}$.

Podemos tomar los sucesos:

$$A = \{5, 6\} \text{ (salir 5 o más)}$$

$$B = \{2, 4, 6\} \text{ (salir par)}$$

En consecuencia, $AB = \{6\}$ y de aquí $P(AB) = 1/6$ (N° de elementos de AB / N° de elementos de Ω)

Si sabemos que ha sucedido B, el espacio muestral se reduce a B ya que los únicos resultados posibles son $\{2, 4, 6\}$. En tal caso: $A = \{6\}$, puesto que el 5 no existe en el nuevo espacio muestral. y entonces $P(A/B) = 1/3$ (N° de elementos del nuevo A / N° de elementos de B).

2. DEFINICIÓN.

El razonamiento anterior nos lleva a la relación: $P(A/B) = \frac{P(AB)}{P(B)}$

que también puede escribirse como: $P(AB) = P(B)P(A/B) = P(A)P(B/A)$

Aplicada de forma iterativa nos da:

$$P(A_1 A_2 \dots A_n) = P(A_1) P(A_2 \dots A_n / A_1) = P(A_1) P(A_2 / A_1) P(A_3 \dots A_n / A_1 A_2) = \dots$$

Con lo que al final obtenemos:

$$P(A_1 A_2 \dots A_n) = P(A_1) P(A_2 / A_1) P(A_3 / A_1 A_2) \dots P(A_n / A_1 A_2 \dots A_{n-1})$$

2. SUCEOS INDEPENDIENTES.

Decimos que los sucesos A y B son independientes cuando el conocimiento de que uno de ellos ha ocurrido no modifica la probabilidad de que ocurra el otro.

Teniendo en cuenta lo dicho al hablar de probabilidad condicionada: A y B son independientes es equivalente a cualquiera de las 3 afirmaciones siguientes:

- a) $P(A/B) = P(A)$
- b) $P(B/A) = P(B)$
- c) $P(AB) = P(A)P(B)$

Este resultado es generalizable a un número cualquiera de sucesos:

$$A_1 A_2 \dots A_n \text{ son independientes} \Leftrightarrow P(A_1 A_2 \dots A_n) = P(A_1) P(A_2) \dots P(A_n)$$

3. TEOREMA DE LA PROBABILIDAD TOTAL.

Dado un espacio muestral Ω y un conjunto de sucesos $B = \{B_1, B_2, \dots, B_n\}$ de modo que:

- a) $B_1 + B_2 + \dots + B_n = \Omega$
- b) $B_i \cdot B_j = \emptyset \quad \forall i \neq j$ ¹

para cualquier suceso $A \in \mathcal{P}(\Omega)$ se verifica:

$$P(A) = \sum_{i=1}^n P(B_i) P(A/B_i)$$

Efectivamente:

¹ Es decir, B constituye lo que se denomina una *partición* de Ω .

$A = A \cdot \Omega = A \sum B_i = \sum AB_i \Rightarrow P(A) = P(\sum AB_i)$ y por ser los B_i disjuntos dos a dos: $P(A) = \sum P(AB_i) \Rightarrow P(A) = \sum P(B_i)P(A/B_i)$.

4. TEOREMA DE BAYES.

Con las mismas premisas del caso anterior:

$$P(B_1/A) = P(AB_1)/P(A) = P(B_1)P(A/B_1)/P(A)$$

y utilizando el resultado del teorema de la probabilidad total:

$$P(B_1 / A) = \frac{P(B_1) P(A / B_1)}{P(B_1) P(A / B_1) + \dots + P(B_n) P(A / B_n)}$$