INSPECTOR ENTIDADES DE CRÉDITO

Tema 1 Estadística descriptiva.

Nombre del logotipo

Tipos de datos y su representación gráfica

En este epígrafe introduciremos los conceptos básicos de la estadística descriptiva.

A la hora de comenzar a trabajar con estadística, lo primero que hay que tener claro es cuál es la población objeto de nuestro estudio.

> La población es el conjunto de individuos que se quiere estudiar, en el que los individuos pueden ser tanto objetos (yogures, ruedas, etc.) como personas, niños, programadores, etc.

También debemos tener claro si disponemos de datos sobre todo el conjunto de la población o sólo sobre una parte de dicha población, que llamaremos muestra.

> Una muestra es un subconjunto cualquiera de la población.

Muestra y población

Hay que distinguir entre población (la totalidad de los individuos) y muestra (una parte cualquiera de la población). En otro módulo se estudiará cómo extraer muestras de una población.

Ejemplos de poblaciones

Los ficheros almacenados en el disco duro de nuestro ordenador, todos los catalanes que ven alguna vez al día la televisión, los ordenadores que están en este preciso momento en el almacén de una distribuidora determinada, el conjunto de personas en el mundo que disponen de conexión a Internet, etc. son ejemplos de poblaciones.

Técnicas de muestreo

¿Os habéis preguntado alguna vez qué programas soléis ver? ¿Tenéis curiosidad por saber cómo es la muestra sobre la que se calculan las audiencias? Las técnicas de muestreo estudian varias alternativas para seleccionar muestras a partir de poblaciones.

Representatividad de una muestra

Imaginemos que queremos averiguar cuáles son los programas de la televisión más vistos por los catalanes durante un determinado mes. La población sería el conjunto de todos los catalanes que han visto la televisión alguna vez durante este mes. Como podéis imaginar fácilmente, es imposible acceder a todos los individuos de la población para preguntarles qué programas han visto: lo que se suele hacer es seleccionar una muestra y preguntar sólo a los individuos que la componen. Los resultados que obtendremos de estudiar la muestra serán más fiables cuanto más representativa sea la muestra, es decir, cuanto más reproduzca, a una escala más reducida, la estructura y composición de la población.

Una vez que se ha decidido la población o muestra objeto de estudio, escogeremos las variables que conviene estudiar.

> Una variable es una característica de los individuos objeto de nuestro estudio.

Ejemplos de variables estadísticas

De los ficheros de nuestro disco duro nos puede interesar el tamaño (en Kbytes), la fecha en que los salvamos por última vez, la aplicación que lo creó, etc. De las audiencias de televisión, las variables serán qué programas se ven, durante cuánto tiempo, cuál gusta más, etc.

Después recogeremos la información para obtener una serie de observaciones del valor de la variable sobre los individuos de la población o de la muestra. Observad que tendremos una observación para cada individuo, aunque, evidentemente, los valores se pueden repetir.

En este módulo os proporcionaremos las pautas básicas de lo que hay que hacer cuando nos encontremos con un conjunto de datos (valores de las variables) correspondientes a una característica determinada de un conjunto de individuos (población o muestra). De momento nos limitaremos a estudiar de forma aislada cada característica de los individuos, es decir, analizaremos cada variable por separado, y por eso diremos que hacemos un análisis univariante.

En esencia, habrá que explorar los datos para llegar a encontrar una descripción de la población (o la muestra) lo más detallada posible. En este sentido, se trata de realizar las funciones siguientes:

a) Averiguar la distribución de la variable: los valores que toma y cómo los toma (si algunos valores se repiten mucho, si algunos valores son "extraños", etc.).

b) Calcular algunos resúmenes numéricos que ayuden a entender cuál es el "centro" de los valores y cómo se distribuyen dichos valores en torno a este "centro".

c) Dibujar algunos gráficos que ayuden a visualizar los puntos mencionados anteriormente.

La importancia de los gráficos

Un buen gráfico siempre es de gran ayuda: ¡representad siempre los datos!

Estas operaciones deben permitirnos elaborar una descripción global de los datos a partir de la cual debemos ser capaces de conseguir lo siguiente:

- Llegar a conclusiones sobre los datos, fundamentadas en los resúmenes numéricos y en los gráficos (como por ejemplo, "tal como se ve en este gráfico..." o "las medidas de centro indican que...").
- Comparar datos referidos a la misma característica sobre dos conjuntos diferentes de individuos (como por ejemplo, "en el gráfico correspondiente al primer colectivo se ve..., mientras que en el del segundo colectivo se ve..." o bien "los resúmenes numéricos del primer colectivo muestran como..., mientras que en el segundo colectivo vemos que...").
- Plantearnos preguntas más complejas (como por ejemplo, sobre relaciones entre variables –regresión– o sobre si hay diferencias entre dos colectivos).
- Ver, en una muestra, qué preguntas nos podemos hacer con respecto a las características globales de la población –inferencia.

Veréis la inferencia estadística y la regresión en otros módulos.

1. Tipos de variables

Cuando estudiamos una población o muestra determinada, seleccionamos unas cuantas variables relevantes. Estas variables pueden ser de diferentes tipos:

1) Variables cualitativas: son aquellas que no se expresan numéricamente, sino como categorías o características de los individuos. A veces reciben el nombre de variables categóricas.

2) Variables cuantitativas: son las que se expresan de forma numérica. De entre estas últimas podemos distinguir los tipos siguientes:

Variables cualitativas y cuantitativas

El color de los ojos, la profesión de una persona y el tipo de ordenador que utiliza son variables cualitativas.

El número de hijos de una persona (contamos hijos), los puntos obtenidos por un equipo en un partido de básquet (contamos puntos), etc., son variables cuantitativas (numéricas) discretas.

El peso de una persona, la cotización del euro con respecto al dólar, el tiempo de acceso a una base de datos, etc., son variables numéricas continuas.

a) Variables cuantitativas discretas: sólo toman valores enteros. Generalmente provienen de contar unidades de una clase determinada.

b) Variables cuantitativas continuas: pueden tomar cualquier valor en un intervalo. Acostumbran a ser el resultado de medir algún fenómeno.

En la práctica muchas variables cualitativas se codifican (se asigna un número a cada categoría) para facilitar su estudio.

Ejemplo de codificación de variables cualitativas

La variable "Tipo de ordenador que utiliza normalmente para conectarse a Internet" tiene un número finito de posibilidades diferentes. Si se asigna un número entero a cada tipo de ordenador posible (por ejemplo, 1 = PC, 2 = Mac, 3 = Otros), obtenemos una codificación numérica que describe la variable que queremos estudiar.

De la misma forma, una variable discreta que toma un número finito de valores se puede tratar como una variable cualitativa, en la que todos los individuos en los que la variable toma un valor determinado forman una categoría.

Ejemplo de categorización de una variable discreta

La variable "Número de hijos de una pareja" es cuantitativa discreta (y además toma pocos valores, quizá quince o dieciséis valores diferentes como máximo). Por otro lado, todas las parejas en las que la variable vale 1 constituyen la categoría de las parejas con hijo único, todas las parejas cuya variable vale 2 constituyen la categoría de las parejas con dos hijos, etc.

2. Variables cualitativas y variables numéricas discretas que toman un número pequeño de valores diferentes

Supongamos que tenemos que estudiar una variable categórica que puede tomar k valores posibles, o bien una variable cuantitativa discreta que puede tomar k valores diferentes, donde k es un número relativamente pequeño. Estos dos tipos de variables admiten un tratamiento similar, en el que hay que comenzar por averiguar las características siguientes:

1) El número total de individuos de los que se disponen datos; designaremos este número con N.

Frecuencias absoluta y relativa de una categoría

La frecuencia absoluta de una categoría es el número de individuos que pertenecen a la categoría.

La frecuencia relativa de una categoría es la proporción de individuos que pertenecen a ella.

2) La frecuencia absoluta de cada valor de la variable: es decir, el número de individuos para los cuales la variable toma este valor. Designaremos con n_i la frecuencia absoluta del valor i.

3) La frecuencia relativa de cada valor de la variable: es decir, la proporción de individuos en los que la variable toma este valor. Designaremos con f_i la frecuencia relativa del valor i, y la acostumbraremos a dar en porcentaje.

Con esto tendremos la distribución de frecuencias de la variable, que es el conjunto de los valores que adopta la variable y la frecuencia con que los adopta.

Utilidad de la frecuencia relativa

La frecuencia relativa nos será muy útil si debemos comparar la misma variable en dos poblaciones diferentes que tienen diferente número de individuos.

Dado que la variable debe tomar alguno de los k valores posibles, es evidente que $n_1 + n_2 + ... + n_k = N$. Además, f_i se obtiene de dividir el número de individuos en los que la variable toma el valor y por el total de individuos, es decir:

$$f_i = \frac{n_i}{N}$$

Es muy fácil ver que $f_1 + f_2 + ... + f_k = 1$, es decir, la suma de las frecuencias relativas de todos los valores es siempre igual a 1.

En caso de que tenga sentido ordenar los valores de la variable, para cada valor x_j se pueden definir la frecuencia absoluta acumulada, que es la suma de las frecuencias de los valores menores o iguales que x_j, y la frecuencia relativa acumulada, que es la suma de las frecuencias relativas de los valores menores o iguales que x_j. Más formalmente, si tenemos N observaciones de una variable que toma k valores diferentes $x_1, x_2, x_3, \ldots, x_k$, de manera que $x_1 < x_2 < x_3 < \ldots < x_k$, definimos:

- La frecuencia absoluta acumulada del valor x_j como $N_j = n_1 + n_2 + n_3 + \ldots + n_j$.
- La frecuencia relativa acumulada del valor x_j como $F_j = f_1 + f_2 + f_3 + \ldots + f_j$.

De estas definiciones, resulta evidente que $N_k = n_1 + n_2 + n_3 + \ldots + n_k = N$ y que $F_k = f_1 + f_2 + f_3 + \ldots + f_k = 1$.

Normalmente, las frecuencias se representan en forma de tabla. Lo veremos en seguida en el ejemplo de los tipos de ordenadores. Sin embargo, antes trataremos de representar gráficamente la distribución de la variable que se estudia; os sugerimos dos formas muy útiles y muy simples.

2.1. El diagrama de barras

Procedimiento para dibujar un diagrama de barras.

Para hacer una representación en forma de diagrama de barras, se dispone un eje horizontal sobre el que se sitúan tantas barras como categorías o valores toma la variable, separadas por pequeños espacios, y con un título claro en la base de cada barra que indique a qué categoría nos referimos.

En el eje vertical del diagrama de barras marcaremos una escala que permita leer bien la altura de cada barra. Esta altura será o bien la frecuencia absoluta o bien la frecuencia relativa de la categoría correspondiente a la barra. En el eje vertical hay que indicar qué frecuencia se representa. En el eje horizontal también marcaremos, si es preciso, una escala con las unidades correspondientes.

2.2. El diagrama de sectores

Procedimiento para dibujar un diagrama de sectores.

Un diagrama de sectores consiste en un círculo que se reparte en diferentes sectores (o porciones) de manera que las superficies de los sectores sean proporcionales a las frecuencias de cada una de las categorías.

Si tenemos que hacerlo a mano, para obtener el ángulo que corresponde a cada sector tenemos que multiplicar los 360° de la circunferencia por la frecuencia relativa de cada clase; después hay que utilizar el transportador de ángulos.

Ejemplo del tipo de ordenadores

Hemos hecho una encuesta entre estudiantes de la UOC en la que se pregunta el tipo de ordenador que utilizan habitualmente en casa para conectarse al campus virtual. A continuación damos las respuestas abreviadas:

PC, PC, MAC, PCP, PCP, PCP, MAC, O, O, PC
PC, PC, PC, MACP, O, MAC, O, PCP, PCP, PC

donde:

- 1 = PC = "PC compatible de sobremesa"
- 2 = PCP = "PC portátil"
- 3 = MAC = "MAC"
- 4 = MACP = "MAC portátil"
- 5 = O = "Otros"

Con estos datos podemos calcular fácilmente la distribución de frecuencias de la variable. Primero contamos las respuestas (20), con lo que N = 20. A partir de aquí confeccionamos una tabla en la que se indiquen las frecuencias absolutas y relativas de cada categoría y con una hilera adicional para las sumas totales, para comprobar que no nos hemos equivocado.

	Frecuencia	Frecuencia relativa
Tipo de ordenador	n_i	f_i
1 = PC	$n_1 = 7$	$f_1 = 7/20 = 0{,}35 = 35\%$
2 = PCP	$n_2 = 5$	$f_2 = 5/20 = 0{,}25 = 25\%$
3 = MAC	$n_3 = 3$	$f_3 = 3/20 = 0{,}15 = 15\%$
4 = MACP	$n_4 = 1$	$f_4 = 1/20 = 0{,}05 = 5\%$
5 = Otros	$n_5 = 4$	$f_5 = 4/20 = 0{,}2 = 20\%$
Totales	N = 20	100%

Tabla de distribución de frecuencias

Esta tabla muestra la distribución de frecuencias de la variable.

Observad que la suma de la columna n_i es:

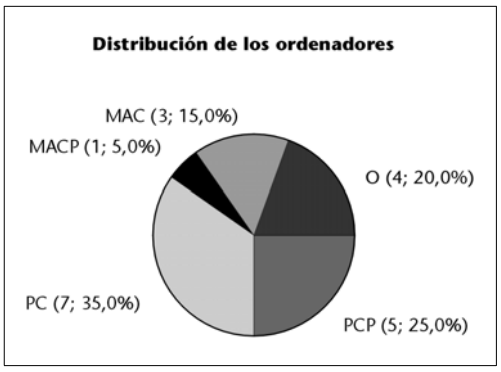
$7 + 5 + 3 + 1 + 4 = 20 = N$

y que la suma de las frecuencias relativas es:

$100\% = 1$

Si haciendo las sumas no se obtienen estos valores, seguro que hay algún error (salvo problemas de redondeo).

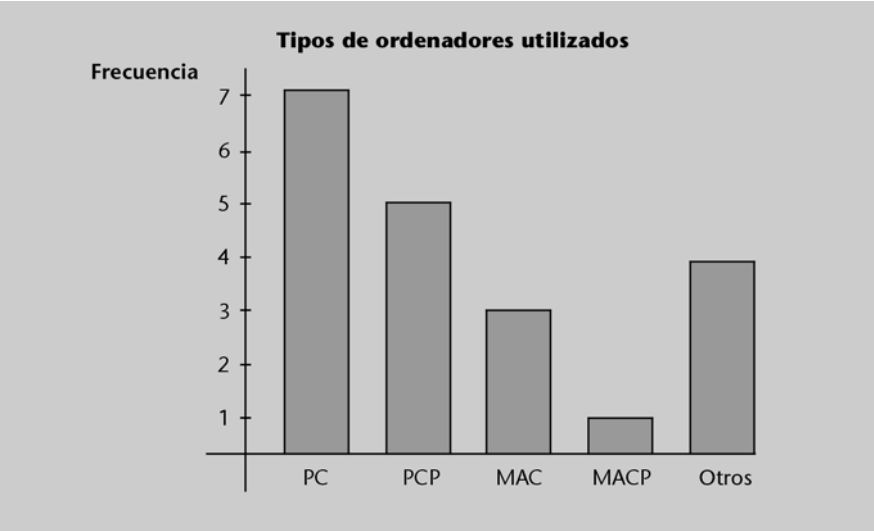
Representamos en primer lugar un diagrama de sectores, en el que vemos que el "pastel" correspondiente a la población está repartido en cinco categorías y que las porciones más grandes corresponden a ordenadores personales (PC) y personales portátiles (PCP), que en total incluyen al 60% de los usuarios.



Cómo dibujar las secciones circulares

¿Recordáis el transportador de ángulos? La mayoría de las veces haremos el gráfico con ayuda de algún programa informático, así que no tendremos que preocuparnos por él.

A continuación dibujamos el diagrama de barras con la frecuencia absoluta de cada clase. Observamos que las barras más altas corresponden a los PC, bien sean de sobremesa, bien sean portátiles. Con este gráfico obtenemos una visión global del reparto de individuos por categorías.



Ejemplo de las notas de estadística

En el semestre anterior las notas de la asignatura de Estadística de un grupo de alumnos fueron las siguientes:

4, 5, 5, 5, 6, 7, 8, 4, 9, 9, 10, 4, 7, 7 ,7 ,8, 6, 7, 8

A continuación calcularemos la distribución de frecuencias de la variable y añadiremos las frecuencias acumuladas:

	Frecuencia absoluta	Frecuencia relativa	Frecuencia absoluta acumulada	Frecuencia relativa acumulada
Nota	n_i	f_i	N_i	F_i
0	0	0	0	0
1	0	0	0	0
2	0	0	0	0
3	0	0	0	0
4	3	0,15 = 15%	3	0,15 = 15%
5	3	0,15 = 15%	6	0,3 = 30%
6	2	0,1 = 10%	8	0,4 = 40%
7	6	0,3 = 30%	14	0,7 = 70%
8	3	0,15 = 15%	17	0,85 = 85%
9	2	0,1 = 10%	19	0,95 = 95%
10	1	0,05 = 5%	20	1 = 100%
Totales	20			

A partir de la tabla se pueden deducir los hechos siguientes:

- Un 15% de los alumnos ha sacado un 5 (frecuencia relativa del valor 5).
- 17 alumnos han aprobado (suma de frecuencias absolutas de los valores 5 a 10).
- El 85% de los alumnos ha aprobado (suma de frecuencias relativas de los valores 5 a 10).

3. Variables numéricas

Disponemos ahora de un conjunto de datos correspondientes a una variable numérica. Aparte de las técnicas introducidas para el caso de las variables discretas

con pocos valores diferentes y especialmente en caso de que tengamos una variable discreta con muchos valores diferentes o bien una variable continua, lo primero que hay que averiguar es:

- Cuántos individuos aparecen en el estudio (N).
- El máximo (máx) y el mínimo (mín) de los valores que toma la variable.
- El rango de la variable, es decir, la diferencia entre el valor máximo y el mínimo.

A continuación veremos diferentes gráficos que se pueden utilizar para representar este tipo de variables.

3.1. El diagrama de puntos

El diagrama de puntos consta de un único eje horizontal con una escala fijada en la que los individuos se representan por puntos dibujados encima del valor que les corresponde. En caso de valores repetidos, situamos un punto sobre el otro.

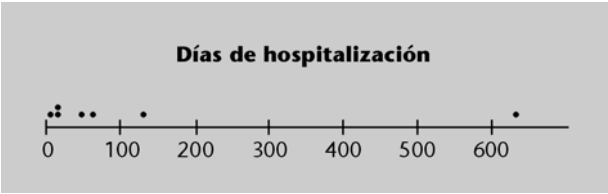
Procedimiento para dibujar un diagrama de puntos.

Ejemplo de los días de hospitalización

Se ha elaborado un estudio sobre los días de hospitalización de un grupo de pacientes sometidos a un mismo tratamiento, estudio del que se han obtenido los datos siguientes: 15, 15, 3, 46, 623, 126, 64 días.

Así pues, tenemos N = 7 datos, que van desde mín = 3 hasta máx = 623.

Dibujamos un eje que contenga estos valores y representamos a cada paciente con un punto un poco por encima de los días que ha estado hospitalizado, de forma que obtenemos el gráfico siguiente, en el que podemos ver que hay una concentración de los datos entre los valores 0 y 70 y un punto muy alejado de los otros (el 623):



Diremos que un dato alejado de los otros o que está fuera del patrón previsible en el gráfico es un dato atípico o dato extremo.

Nomenclatura

Utilizaremos de forma indistinta los términos anómala, extrema, insólita, atípica o su expresión inglesa: outlier.

Si nos encontramos un dato con estas características, de entrada hay que comprobar que no se trate de un error en la introducción de los datos. Una vez comprobado que no lo es, conviene averiguar la causa de este comportamiento anómalo.

3.2. El diagrama de tallo y hojas

Procedimiento para dibujar un diagrama de tallo y hojas.

Para hacer un diagrama de tallo y hojas, hay que llevar a cabo las operaciones siguientes:

1) Partimos cada una de las observaciones de la variable en dos partes: la primera contiene todos los dígitos menos el de más a la derecha (ésta será el tallo), y la segunda contiene el último dígito (ésta será la hoja).

Nomenclatura

En inglés el diagrama de tallo y hojas se llama steam and leaf diagram.

2) Situamos los tallos uno debajo de otro, ordenados de manera creciente.

3) Colocamos al lado de cada tallo las hojas que le correspondan, también en orden creciente.

Para separar el tallo de las hojas, se puede dibujar una línea vertical. Como veremos a continuación, este diagrama tiene un aspecto similar a un diagrama de barras colocado en posición vertical. En este diagrama se encuentran representadas todas las observaciones de la variable y, por tanto, debemos escribir tantas hojas repetidas como sea necesario. Si el gráfico nos parece poco descriptivo (porque tiene pocos niveles, por ejemplo), podemos optar por desdoblar algún nivel.

Ejemplo del jugador de ajedrez

Consideremos el número de minutos que un jugador de ajedrez de gran nivel ha necesitado para ganar al programa de ordenador Deep Yellow en quince partidas consecutivas:

54, 59, 35, 41, 46, 25, 47, 60, 54, 46, 49, 46, 41, 34, 22

El último dígito es la hoja, de manera que, en este caso, en el tallo habrá representadas las decenas:

2	25
3	45
4	1166679
5	449
6	0

En este gráfico podemos ver que los datos se concentran en el "nivel" de los 40 minutos y no hay datos extremos; parece un jugador muy regular.

Si consideramos que el gráfico es poco descriptivo, podemos desdoblar cada decena en dos niveles: el primero contendrá las hojas del 0 al 4 y el segundo, las hojas del 5 al 9, de manera que resultará un diagrama como éste:

2	2
2	5
3	4
3	5
4	11
4	66679
5	44
5	9
6	0

Aquí se aprecia mejor que la acumulación de datos se da en la zona que va de 45 a 49.

4. Variables continuas e histograma

En caso de que nuestra variable numérica tenga muchos valores diferentes o que haya muchos individuos (más de cien, por ejemplo), el diagrama de tallo y hojas puede ser demasiado cargado y difícil de interpretar. También podemos encontranos con tablas de distribución de frecuencias muy "aburridas", en las que todas las frecuencias sean iguales a 1 (éste sería el caso si todos los valores de la variable fuesen diferentes). Para simplificar estas situaciones, se suelen agrupar los datos y representar no las frecuencias de cada valor, sino las de las diferentes agrupaciones.

Estas consideraciones conducen a la necesidad de definir lo que se entiende por distribución de frecuencias de una variable numérica continua o bien discreta con muchos valores diferentes.

Para obtener una expresión de la distribución de frecuencias, haremos lo siguiente:

Procedimiento para calcular la distribución de frecuencias en el caso continuo.

1) Agrupamos las observaciones en intervalos (generalmente todos con el mismo ancho) llamados clases. Los intervalos deben ser adyacentes (¡no vale dejar agujeros!) y deben cubrir, como mínimo, todo el rango, desde el mínimo hasta el máximo.

2) Calculamos el punto medio de cada intervalo, llamado marca de clase. La marca de clase será la "representante" de todas las observaciones que caen en el intervalo.

3) Una vez que tenemos las clases, calculamos la frecuencia absoluta de cada una (el número de observaciones que caen en el intervalo) y su frecuencia relativa (la proporción de observaciones que caen en el intervalo).

4) Si es preciso, también podemos calcular la frecuencia absoluta acumulada de la clase (el número de observaciones que caen en el intervalo más las que caen en intervalos anteriores) y la frecuencia relativa acumulada de la clase (suma de las frecuencias relativas de la clase más la de todas las clases anteriores).

En este caso, una manera muy eficiente de representar la distribución de frecuencias es por medio de los llamados histogramas.

Un histograma es más que un diagrama de barras: de hecho, es un diagrama de rectángulos. Cada rectángulo del diagrama representará una de las clases en las que tenemos distribuidos los valores de la variable, de manera que la proporción sobre el área total del área de cada rectángulo es precisamente la fre-

cuencia relativa de la clase que representa. En el caso de que la suma del área de todos los rectángulos sea 1, el área de cada rectángulo será igual a la frecuencia relativa de la clase que representa, y diremos que hemos construido un histograma de densidad.

Para construir un histograma de densidad, debemos dar los pasos siguientes:

Procedimiento para dibujar un histograma de densidad.

1) Seleccionamos una escala adecuada en el eje de las x.

2) Marcamos en el eje x todas las clases en las que se distribuye la variable.

3) Seleccionamos una escala adecuada en el eje de las y.

4) Para cada clase representamos un rectángulo que tiene como base la clase misma y como altura, la frecuencia relativa de la clase dividida por el ancho de ésta, que, evidentemente, es la longitud de la clase.

En este caso es fácil ver que la suma de las áreas de los rectángulos es 1.

Procedimiento para dibujar un histograma de frecuencias relativas.

En el caso bastante habitual de que todas las clases tengan el mismo ancho, podemos confeccionar otro tipo de histograma que consiste simplemente en poner la frecuencia relativa de la clase como altura de cada rectángulo. De esta manera la suma de todas las áreas de los rectángulos es precisamente el ancho de las clases, y si dividimos el área de uno de los rectángulos por el área total, obtenemos precisamente la frecuencia relativa de la clase. Entonces decimos que se trata de un histograma de frecuencias relativas.

Normalmente, y siempre que sea posible, construiremos histogramas de frecuencias relativas, ya que son más fáciles de interpretar y, de hecho, tienen el mismo aspecto visual que un histograma de densidad (si las clases presentan el mismo ancho). En todo caso, siempre hay que dejar claro qué tipo de histograma se construye e indicar si en el eje de las y se representan frecuencias relativas o densidad.

Recomendaciones

Especificad siempre las unidades en las que medís la escala de cada eje. Indicad también los títulos que convenga.

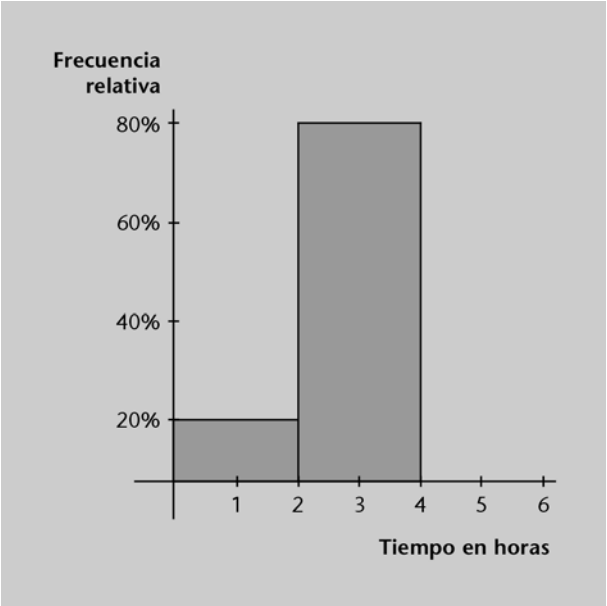
Si las clases tienen anchos diferentes, es aconsejable trabajar con histogramas de densidad.

Ejemplo de elaboración de un diagrama de frecuencias relativas

Se ha llevado a cabo un estudio sobre el número de horas diarias que los estudiantes de la UOC ven la televisión y se ha constatado que el 20% la ve menos de 2 horas y que el 80% restante la ve 2 horas o más, pero menos de 4 horas. Representamos los datos en forma de tabla:

Clases	f_i
[0,2)	20%
[2,4)	80%

Si dibujamos el histograma correspondiente (por ejemplo, de frecuencias relativas, ya que el ancho de todas las clases es igual a 2) obtenemos el gráfico siguiente:



En este gráfico la suma de las áreas de todos los rectángulos es:

$$2 \times 20\% + 2 \times 80\% = 2.$$

El área del primer rectángulo es $2 \times 20\% = 0{,}4$, que es precisamente el 20% de 2 (el área total), y el área del segundo rectángulo es $2 \times 80\% = 1{,}6$, que es precisamente el 80% del área total (que es 2).

Ejemplo de elaboración de un diagrama de densidad

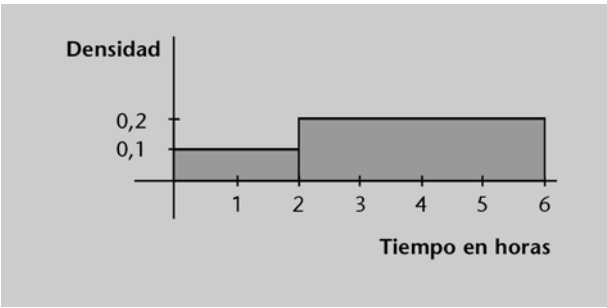
Otro estudio sobre el número de horas diarias que los estudiantes de la UOC ven la televisión ha constatado que el 20% de los estudiantes la ve menos de 2 horas y que el 80% restante la ve 2 horas o más, pero menos de 6. Cuando representamos los datos en una tabla, obtenemos lo siguiente:

Clases	f_i
[0,2)	20%
[2,6)	80%

En este caso tenemos que dibujar un histograma de densidad, en el que la altura de cada rectángulo viene dada por los datos siguientes:

Clases	f_i	Altura del rectángulo
[0,2)	20%	20% / 2 = 0,1
[2,6)	80%	80% / 4 = 0,2

El histograma resultante es el que vemos a continuación (este histograma y el anterior están dibujados con la misma escala en el eje de las x para facilitar su comparación):



Interpretación de los histogramas de densidad

El diagrama de densidad del ejemplo nos informa del hecho de que a cada una de las horas de la clase [0,2) le corresponde un 0,1 = 10% de la población, mientras que a cada una de las horas que van de la 2 a la 6 le corresponde un 80% / 4 = 20% de la población.

Ahora la superficie total es:

$$2 \times 0{,}1 + 4 \times 0{,}2 = 1.$$

El área del primer rectángulo es $2 \times 0{,}1 = 0{,}2 = 20\%$, que es precisamente la frecuencia relativa de la primera clase, y el área del segundo rectángulo es $4 \times 20\% = 0{,}8 = 80\%$, que es la frecuencia relativa de la segunda clase.

Los histogramas de densidad dan una idea de cómo se distribuyen las observaciones dentro de cada clase.

4.1. Histograma de frecuencias relativas acumuladas

La función de distribución acumulada

Al hablar de probabilidad y de variables aleatorias, utilizaremos la función de distribución acumulada. En determinados casos la representación de esta función es un histograma de frecuencias relativas acumuladas.

Para representar de manera gráfica las frecuencias relativas acumuladas, se suele utilizar un histograma. En el histograma de frecuencias relativas acumuladas la base del rectángulo es la clase y la altura, la frecuencia relativa acumulada de la clase. En este histograma la altura de los rectángulos crece hasta llegar a la altura de la última clase, que es evidentemente 1.

4.2. Ejemplos de construcción de histogramas

Como podéis observar, la descripción de cómo se calcula la distribución de frecuencias para estas variables y de cómo se hace un histograma son muy generales y dependen de los intervalos que seleccionemos. No existe una norma universal que explique cuántas ni qué clases debemos considerar; cada vez tendremos que decidirlo según el tipo y la forma de los datos, sin perder de vista el objetivo final.

La finalidad de la construcción de histogramas es obtener una representación gráfica que resuma la distribución de los datos de una manera entendedora, fácil de asimilar y que muestre aspectos relevantes de los datos.

Histogramas por ordenador

La mayoría de los programas de ordenador permiten escoger el número y la forma de los intervalos: podemos hacer pruebas hasta encontrar una "buena" representación.

Muchas veces tendremos que hacer varias pruebas hasta conseguir la representación que mejor ilustre la forma de la distribución o que argumente mejor nuestras opiniones. Si tenemos que utilizar un histograma, es conveniente tener en cuenta los aspectos siguientes:

a) Siempre que sea posible, se cogerán clases del mismo ancho.

b) El número de clases debe ser aproximadamente igual a $\sqrt{N}$, y mejor si está entre 7 y 15 (N es, como siempre, el número de observaciones).

La importancia del número de clases

Si el número de clases es demasiado pequeño, tendremos histogramas con pocas barras y bastante altas; si hay demasiadas clases, parecerá un conjunto de palos de alturas parecidas y con agujeros.

c) Para determinar dónde debe comenzar y acabar cada clase, podemos guiarnos por las consideraciones siguientes:

- Debe quedar muy claro a qué clase deben pertenecer los puntos de cambio de intervalo; generalmente utilizaremos intervalos de la forma [x, y), de manera que contienen el extremo izquierdo, pero no el derecho.

- Cuando los datos toman valores enteros como, por ejemplo, 11, 13, 15, etc., es aconsejable considerar clases de manera que su punto medio coincida con los valores enteros; en este caso definiríamos las clases [10,5-11,5), [11,5-12,5), etc.

Los resultados de un test de estadística

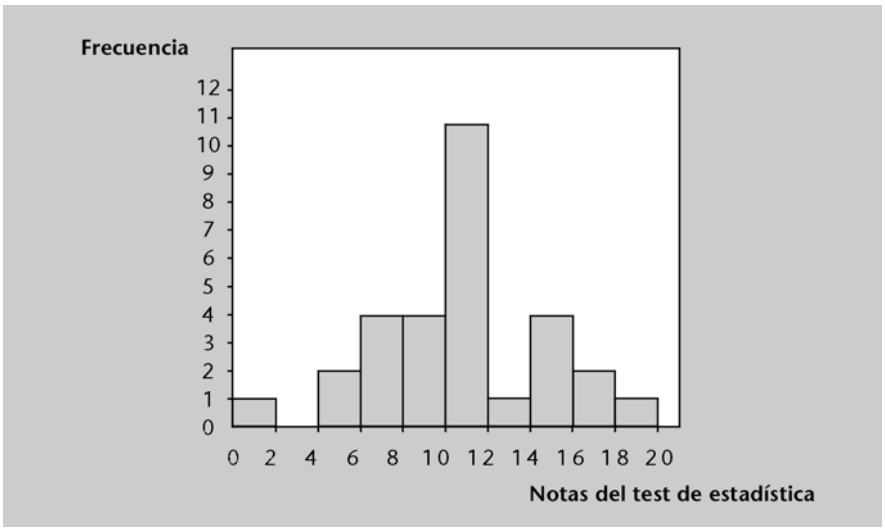
A continuación mostramos las notas de un test de estadística pasado a treinta estudiantes de un curso de la UOC durante el semestre de otoño. La puntuación posible del test va de 0 a 20 puntos.

11,5	7	5,25	19	8,5	11	10	6	15,5	11,75
9,75	15	16	16	7,5	8	11,5	0	10	11
13	10,25	10	8	14	6	14,5	10	5,6	10

Primero calculamos el valor máximo (máx = 19), mínimo (mín = 0) y el rango (máx – mín = = 19). En el paso siguiente se agrupan los datos en intervalos de manera que cubran todo el rango. En este caso podríamos agrupar los datos en intervalos de longitud 1, comenzando por el 0 y acabando por el 20. Puesto que salen demasiadas clases, optamos por hacer clases de longitud 2, de manera que obtenemos una tabla como ésta, en la que también se han calculado las frecuencias acumuladas.

		Frecuencia absoluta	Frecuencia relativa	Frecuencia absoluta acumulada	Frecuencia relativa acumulada
Intervalo	Marca de clase	n_i	f_i	N_i	F^i
[0,2)	1 = (0 + 2) / 2	1	1/30 = 0,03	1	0,03
[2,4)	3 = (2 + 4) / 2	0	0 = 0	1	0,03
[4,6)	5	2	2/30 = 0,07	3	0,10
[6,8)	7	4	4/30 = 0,13	7	0,23
[8,10)	9	4	4/30 = 0,13	11	0,37
[10,12)	11	11	11/30 = 0,37	22	0,73
[12,14)	13	1	1/30 = 0,03	23	0,77
[14,16)	15	4	4/30 = 0,13	27	0,90
[16,18)	17	2	2/30 = 0,07	29	0,97
[18,20)	19 = (18 + 20) / 2	1	1/30 = 0,03	30	1,00
Totales		30	1		

Con estos datos podemos dibujar el histograma siguiente:



Donde se aprecia una cima destacada que corresponde a la clase [10,12) y una cierta simetría alrededor de ésta. También se observa una clase vacía –la [2,4) tiene frecuencia 0.

Cálculo del valor medio

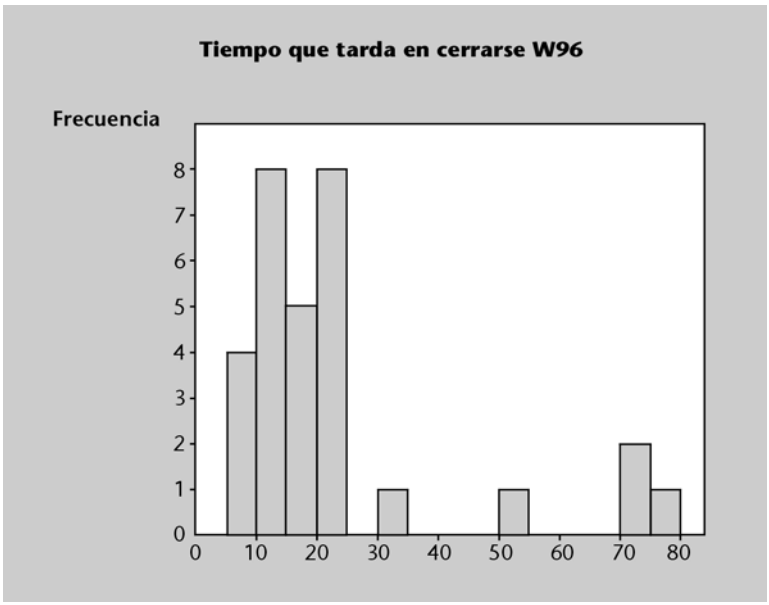
Para calcular el punto medio de un intervalo, se suman los extremos y se divide el resultado por 2.

Tiempo que tarda en cerrarse el ordenador

En este segundo ejemplo de histograma utilizaremos los datos siguientes, que corresponden al tiempo que ha tardado en cerrarse nuestro ordenador (equipado con Windows 96) desde que hemos dado la instrucción adecuada, las últimas treinta veces que lo hemos utilizado:

16	32	10	24	23	12	15	21	16	10
24	20	21	71	12	50	76	15	19	8
8	10	12	23	9	71	13	14	20	6

Con un programa estándar hemos obtenido el histograma siguiente:



Como podéis ver, aquí las clases son [0,5), [5,10) ... hasta [75,80). Hay dos picos, correspondientes a los intervalos [10,15) y [20,25), aunque el comportamiento es bastante irregular y nada simétrico. También hay datos alejados o extremos en torno a los 70 segundos y parece que haya tres agrupaciones en los datos: de 0 a 30, de 50 a 60 y de 70 a 80 segundos. Cada una de estas agrupaciones podría estar relacionada con las circunstancias en las que lo apagamos y con los programas que se han ejecutado.

4.3. Interpretación de los histogramas

En la interpretación de los histogramas es conveniente destacar los aspectos siguientes:

a) La simetría: el histograma será simétrico si es posible trazar un eje vertical de manera que la parte que hay a la derecha del eje sea aproximadamente igual a la imagen reflejada en un espejo de la parte izquierda.

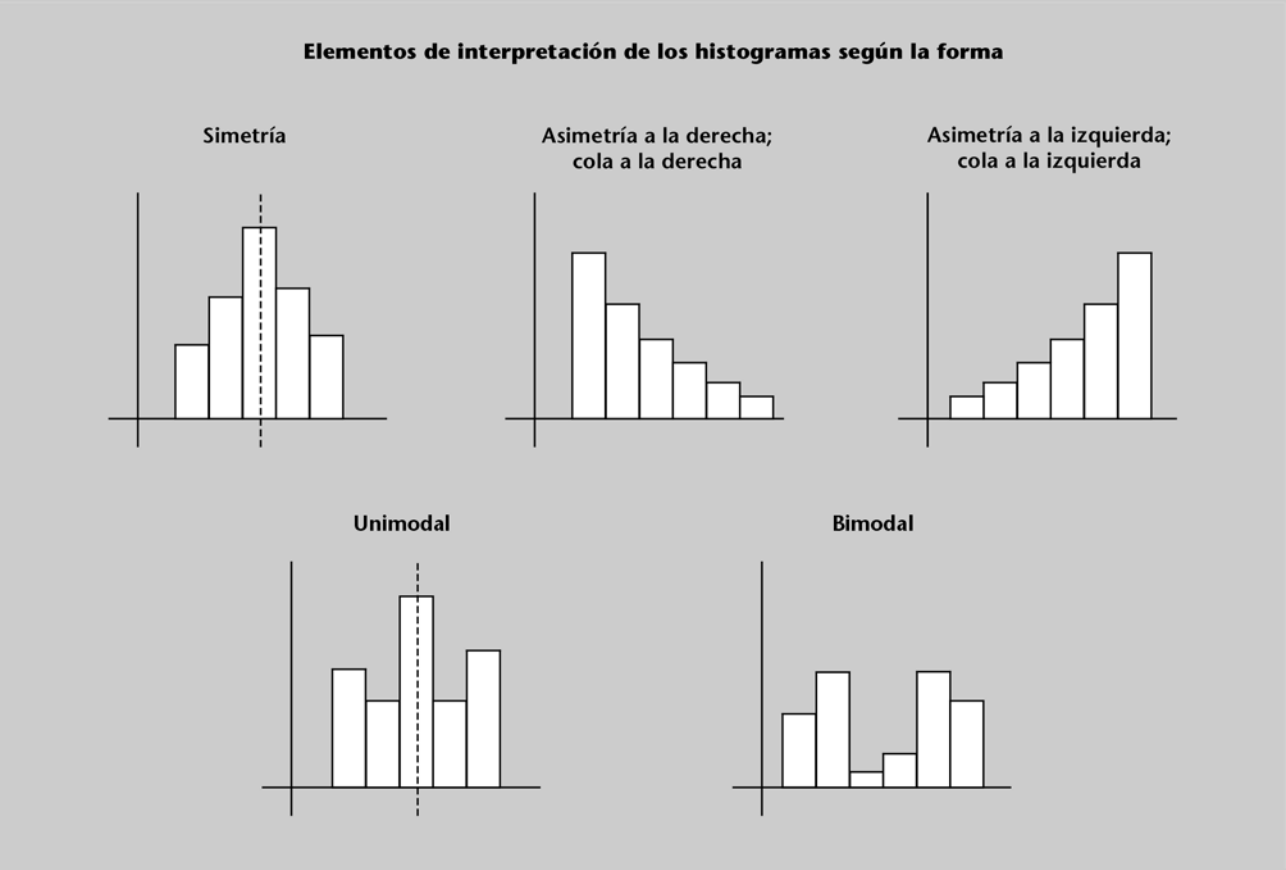
b) Los picos: corresponden a intervalos en los que se tienden a concentrar los valores de la variable (son intervalos con frecuencias grandes). Si sólo hay un pico destacado, diremos que es un caso unimodal; si tenemos dos picos de altura similar, será un caso bimodal.

c) Las colas: si no hay simetría, es posible que uno de los lados del histograma se extienda mucho más lejos que el otro. En este caso diremos que hay una

cola en el lado más extenso (llamado caso de cola larga) o que hay asimetría hacia este mismo lado.

d) Hay que detectar los datos extremos, las clases vacías y si estas clases vacías separan a la población en grupos.

A continuación representamos de forma gráfica algunos de estos casos e indicamos el eje de simetría en los que hay:



5. Resumen

En esta sesión hemos introducido el vocabulario básico de la estadística (población/muestra, individuo, variable, frecuencia, distribución de la variable, etc.). Según el tipo de variable considerada, se han presentado varias formas de representar su distribución (diagrama de barras y de sectores para variables cualitativas y numéricas discretas con pocos valores y diagramas de puntos, de tallo y hojas e histograma para las cuantitativas).

Ejercicios

1. Considerad los cuatro programas más vistos durante una semana de una cadena de TV determinada un día concreto. El número de espectadores (en miles) que ven cada uno de los programas es el que se muestra en la tabla siguiente:

Programa	Número de espectadores
"Viva la estadística"	875
"El hermano"	925
"Matanza sangrienta"	742
"Informativo del día"	682

a) Representad gráficamente los datos de la manera que creáis más oportuna.

b) Si la semana siguiente el número de espectadores de estos mismos programas es el que refleja la tabla que hay a continuación, representad los datos de manera que se pueda comparar el número de espectadores en las dos semanas.

Programa	Número de espectadores
"Viva la estadística"	100
"El hermano"	200
"Matanza sangrienta"	100
"Informativo del día"	682

2. Una empresa utiliza como procesador de textos el famoso programa Macrohard Phrase. Se ha pedido a los usuarios de este programa que anoten cuántas veces se les ha "colgado" el ordenador durante un mes mientras trabajaban con éste. Se han obtenido los resultados siguientes:

47	23	28	0	11	12	35	36	40	30	37	14

Representad gráficamente los datos de la manera que creáis más adecuada.

3. Dado que el comportamiento del programa Macrohard parece harto preocupante, hemos decidido recoger datos sobre el número de veces que se ha "colgado" en un determinado mes en cincuenta ordenadores diferentes:

Ejemplo: el histograma de Macrohard.

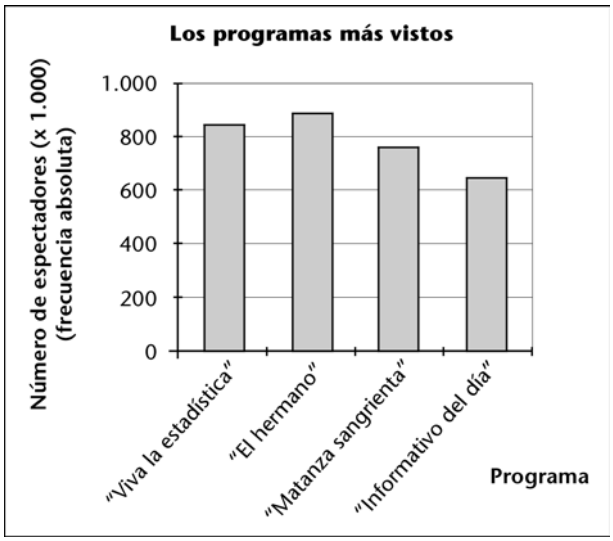
0	9	12	14	19
2	10	12	14	20
4	10	12	14	20
5	10	12	15	21
6	11	12	15	22
6	11	12	17	29
6	11	12	17	29
7	11	13	18	32
8	11	13	18	39
9	12	14	19	39

Encontrad la distribución de frecuencias de esta variable, representadla en forma de histograma y comentad sus características. Representad el histograma de frecuencias relativas acumuladas.

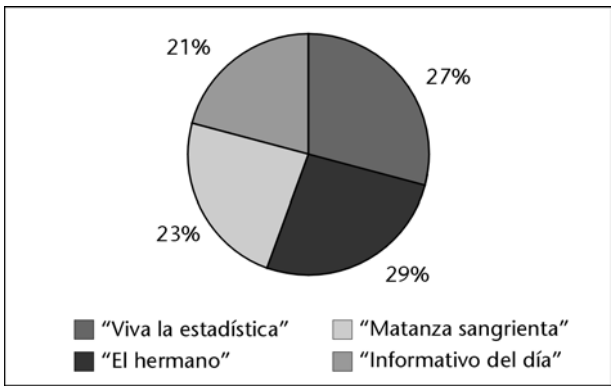
Solucionario

1.

a) Podemos hacer un gráfico de barras como éste (por ejemplo, con Excel):



También podemos representar un diagrama de sectores con las frecuencias relativas:



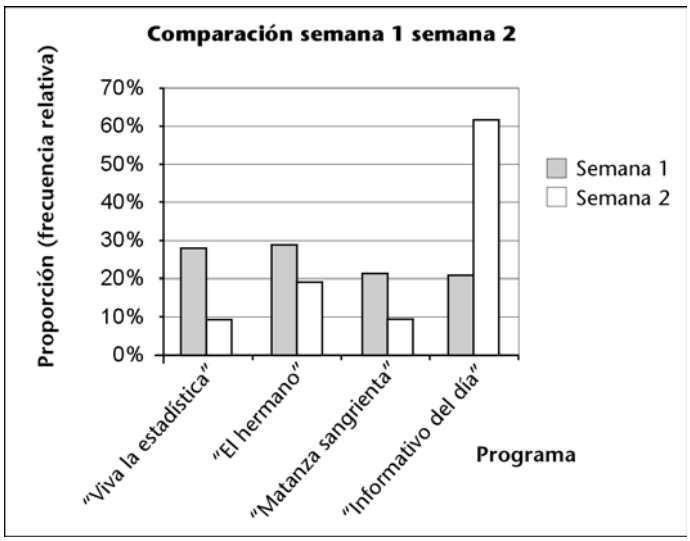
en el que podemos ver que las audiencias están muy igualadas.

b) Para poder comparar las dos semanas, es mejor calcular primero las frecuencias relativas de cada programa en cada semana:

	Frecuencia absoluta		Frecuencia relativa	
Programa	Semana 1	Semana 2	Semana 1	Semana 2
Viva la estadística	875	100	27,14%	9,24%

	Frecuencia absoluta		Frecuencia relativa	
Programa	Semana 1	Semana 2	Semana 1	Semana 2
El hermano	925	200	28,69%	18,48%
Matanza sangrienta	742	100	23,01%	9,24%
Informativo del día	682	682	21,15%	63,03%
Total	3.224	1.082	100% = 1	1

A partir de estos datos podemos crear el gráfico que se muestra a continuación, en el que para cada programa se representan dos barras, una para cada semana:



En el gráfico se ve claramente que el "Informativo del día" ha aumentado en gran medida su cuota de pantalla, de modo que supera ampliamente los otros programas, aunque el número de espectadores se ha mantenido constante.

Terminología

En el caso específico de las audiencias de televisión, la frecuencia relativa se denomina cuota de pantalla.

2. Dado que tenemos pocos datos, podemos optar por hacer un diagrama de tallo y hojas:

0	0
1	124
2	38
3	0567
4	07

Como podemos ver, se trata de un diagrama bastante simétrico, distribuido en torno a los valores entre 20 y 30, con cierta tendencia a desplazarse hacia los valores grandes (30 en adelante). Realmente, el programa no parece demasiado fiable.

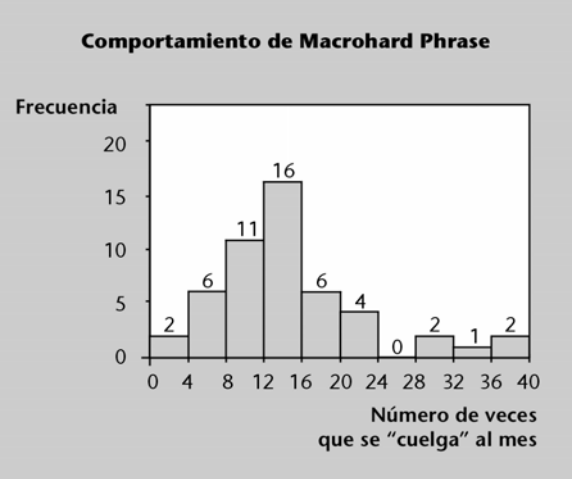
3. El valor mínimo que toma la variable es 0, el máximo es 39 y hay 50 observaciones. Puesto que el rango es aproximadamente 40, podemos hacer 10 clases de ancho 4, comenzando en 0 y acabando en 40. Probamos esta opción. De entrada, tabulamos los datos para obtener las frecuencias de cada clase:

Comenzar en 0

Es razonable comenzar en 0, ya que no puede ser que un ordenador se "cuelgue" menos de cero veces.

		Frecuencia absoluta	Frecuencia relativa	Frecuencia absoluta acumulada	Frecuencia relativa acumulada
Intervalo	Marca de clase	n_i	f_i	N_i	F_i
[0,4)	2	2	4%	2	0,04
[4,8)	6	6	12%	8	0,16
[8,12)	10	11	22%	19	0,38
[12,16)	14	16	32%	35	0,7
[16,20)	18	6	12%	41	0,82
[20,24)	22	4	8%	45	0,9
[24,28)	26	0	0%	45	0,9
[28,32)	30	2	4%	47	0,94
[32,36)	34	1	2%	48	0,96
[36,40)	38	2	4%	50	1
Totales		50	1		

Y después dibujamos el histograma, en el que, por comodidad, hemos escrito las frecuencias de cada clase.



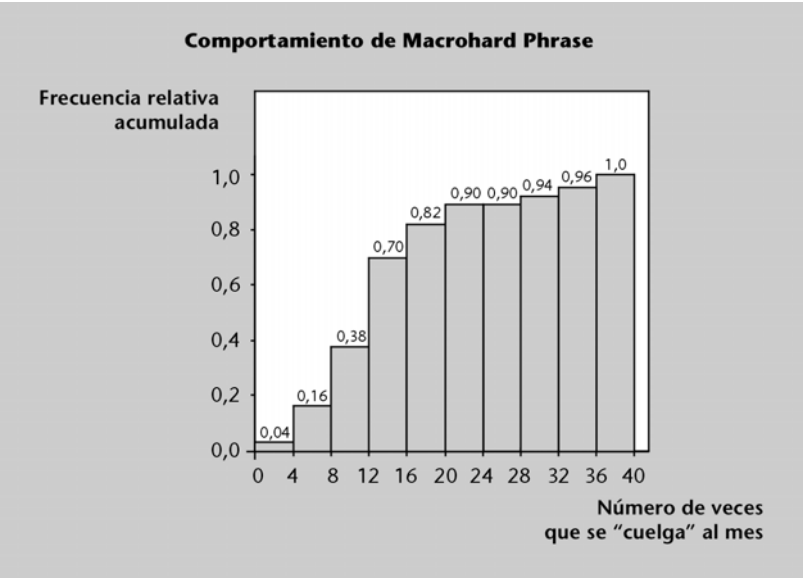
Caso particular

En caso de que alguna observación fuese exactamente 40, podríamos optar por considerar el último intervalo de la forma [36,40].

Interpretación

Dos ordenadores se han "colgado" entre 0 y 4 veces, 6 entre 16 y 20 veces, etc.

El histograma de frecuencias relativas acumuladas resulta de la manera siguiente:



Interpretación

El 70% de los ordenadores se ha "colgado" menos de 16 veces; ell 90%, menos de 28 veces, etc

La frecuencia absoluta es la medida estadística que nos permite reconocer la cantidad de veces que se repite un dato. Principalmente, es usada en estadística descriptiva, y es útil para obtener información sobre las características de una muestra o población.

Además, se puede ocupar tanto con variables cualitativas como cuantitativas, siempre y cuando se puedan ordenar. En cuanto a estas, pueden ser discretas (ordenadas de menor a mayor) y continuas (de menor a mayor, agrupadas por intervalos).

Sumado a esto, permite calcular la frecuencia relativa, pero abordaremos este tema en el próximo apartado.

Pero entonces, ¿qué es la frecuencia absoluta? Es la cantidad de veces que se repite un dato, y la suma de estos debe ser igual al número total de datos de la muestra o población. A este número se lo denomina frecuencia absoluta acumulada

Retomemos el ejemplo de las clases virtuales y veamos cómo se calculan ambas frecuencias.

Frecuencia relativa

Para calcular la frecuencia relativa, primeramente, necesitamos conocer la absoluta. Luego, se aplica una fórmula que consiste en dividir cada frecuencia absoluta por la frecuencia absoluta acumulada.

En el ejemplo, dividiremos cada frecuencia absoluta por 10. Como resultado, siempre obtendremos números comprendidos entre 0 y 1, debido a que los valores siempre serán de menor tamaño que la muestra.

Como consecuencia de esto, la frecuencia relativa acumulada, es decir, la suma de todas las frecuencias relativas, debe ser igual 1. Si los valores se multiplican por 100, se obtendrá un porcentaje.

Entonces, este cálculo nos da información sobre la proporción o el peso que tienen los valores de la muestra. La utilidad de esto, es que nos va a permitir hacer comparaciones entre muestras de distintos tamaños.

La frecuencia absoluta y relativa son dos conceptos importantes en estadística que se utilizan para analizar y visualizar datos. En este artículo, explicaremos qué son la frecuencia absoluta y relativa, cómo calcularlas y las diferencias entre ellas.
¿Qué es la frecuencia absoluta?
La frecuencia absoluta es el número de veces que aparece un valor en un conjunto de datos. Se puede calcular utilizando la siguiente fórmula:
Frecuencia absoluta = número de veces que aparece el valor
Por ejemplo, si tenemos los siguientes datos: 1, 2, 3, 2, 4, 2, 5, 1, 6, 1, la frecuencia absoluta de 2 es 3 porque aparece tres veces en el conjunto de datos.
¿Qué es la frecuencia relativa?
La frecuencia relativa es el porcentaje de veces que aparece un valor en un conjunto de datos. Se puede calcular utilizando la siguiente fórmula:
Frecuencia relativa = (frecuencia absoluta / tamaño del conjunto de datos) x 100%
Por ejemplo, si tenemos los mismos datos que en el ejemplo anterior, la frecuencia relativa de 2 es 27.27% porque aparece tres veces y el tamaño del conjunto de datos es 11.
Diferencias entre frecuencia absoluta y relativa
La principal diferencia entre la frecuencia absoluta y la relativa es que la primera es un número entero que indica cuántas veces aparece un valor en un conjunto de datos, mientras que la segunda es un porcentaje que indica la proporción de veces que aparece ese valor en relación con el tamaño del conjunto de datos.
Otra diferencia importante es que la frecuencia absoluta se utiliza para analizar datos brutos, mientras que la frecuencia relativa se utiliza para analizar datos en términos de su proporción en el conjunto de datos.

Ejemplos prácticos
Supongamos que tenemos un conjunto de datos que representa las edades de un grupo de personas:
16, 19, 21, 22, 25, 19, 20, 18, 22, 23, 21, 24, 19, 18, 20, 22, 19, 23, 25, 20
Podemos calcular la frecuencia absoluta y relativa de cada edad de la siguiente manera:
Edad | Frecuencia absoluta | Frecuencia relativa
— | — | —
16 | 1 | 5%
18 | 2 | 10%
19 | 4 | 20%
20 | 3 | 15%
21 | 2 | 10%
22 | 3 | 15%
23 | 2 | 10%
24 | 1 | 5%
25 | 2 | 10%

De esta tabla, podemos ver que la edad más común en el conjunto de datos es 19, que aparece cuatro veces y representa el 20% del total de edades. También podemos ver que las edades 16 y 24 son las menos comunes, ya que aparecen solo una vez cada una.
Conclusión
La frecuencia absoluta y relativa son dos conceptos importantes en estadística que se utilizan para analizar y visualizar datos. La frecuencia absoluta indica cuántas veces aparece un valor en un conjunto de datos, mientras que la frecuencia relativa indica la proporción de veces que aparece ese valor en relación con el tamaño del conjunto de datos. Al comprender estos conceptos y cómo calcularlos, podemos obtener una mejor comprensión de los datos y utilizarlos de manera más efectiva.

1. Definición de frecuencia absoluta y relativa
La frecuencia absoluta se refiere al número de veces que se repite un valor en un conjunto de datos. Por otro lado, la frecuencia relativa se refiere al porcentaje de veces que se repite un valor en relación al total de datos.
2. Cálculo de la frecuencia absoluta
Para calcular la frecuencia absoluta, sigue estos pasos:
– Cuenta el número de veces que se repite cada valor en el conjunto de datos.
– Registra el número de veces que se repite cada valor en una tabla o gráfico.
– Suma todas las frecuencias absolutas para obtener el total de datos.
Por ejemplo, si tenemos los siguientes datos: 2, 4, 6, 2, 8, 2, 4, la frecuencia absoluta de 2 es 3, la de 4 es 2, la de 6 es 1 y la de 8 es 1.
3. Cálculo de la frecuencia relativa
Para calcular la frecuencia relativa, sigue estos pasos:
– Divide la frecuencia absoluta de cada valor entre el total de datos.
– Multiplica el resultado por 100 para obtener el porcentaje.

Por ejemplo, si tenemos los mismos datos del ejemplo anterior, el total de datos es 7. Para calcular la frecuencia relativa de 2, se divide 3 entre 7 y se multiplica por 100, lo que da un resultado de 42,86%. De esta forma, se calculan las frecuencias relativas de todos los valores.
4. Diferencias entre frecuencia absoluta y relativa
La principal diferencia entre ambas es que la frecuencia absoluta se refiere al número de veces que se repite un valor, mientras que la frecuencia relativa se refiere al porcentaje de veces que se repite un valor en relación al total de datos.
Además, la frecuencia absoluta se utiliza para hacer comparaciones entre conjuntos de datos, mientras que la frecuencia relativa se utiliza para analizar la distribución de los datos y obtener una idea general de su comportamiento.
En resumen, aprender a calcular la frecuencia absoluta y relativa es esencial para cualquier análisis estadístico. Sigue estos sencillos pasos y podrás realizar tus propios cálculos de forma rápida y efectiva.

En estadística, la frecuencia absoluta se refiere al número de veces que ocurre un determinado valor o evento en un conjunto de datos. Por otro lado, la frecuencia relativa se refiere a la proporción o porcentaje que representa la frecuencia absoluta en relación con el tamaño total del conjunto de datos.
Para calcular la frecuencia relativa, se sigue la siguiente fórmula:
Frecuencia relativa = Frecuencia absoluta / Tamaño total del conjunto de datos
Por ejemplo, si tenemos un conjunto de datos de 100 personas y 20 de ellas son hombres, la frecuencia absoluta de hombres sería 20 y la frecuencia relativa sería 20/100 = 0.2 o 20%.
Es importante destacar que la frecuencia relativa siempre estará comprendida entre 0 y 1, o entre 0% y 100%, ya que representa una proporción del total de datos.
Algunas diferencias importantes entre la frecuencia absoluta y la frecuencia relativa son:
– La frecuencia absoluta se refiere al número de veces que ocurre un evento, mientras que la frecuencia relativa se refiere a la proporción que representa ese evento en relación con el total de datos.
– La frecuencia absoluta puede ser un número entero o decimal, mientras que la frecuencia relativa siempre será un número decimal o porcentaje.
– La suma de todas las frecuencias absolutas siempre será igual al tamaño total del conjunto de datos, mientras que la suma de todas las frecuencias relativas siempre será igual a 1 o 100%.
En resumen, la frecuencia relativa es una herramienta útil en estadística para comprender la proporción que representa un evento en relación con el total de datos. Su cálculo es sencillo y se puede utilizar junto con la frecuencia absoluta para obtener una visión más completa de los datos.

En estadística, la frecuencia absoluta es una medida que indica la cantidad de veces que un valor específico aparece en un conjunto de datos. Es una herramienta importante para entender la distribución de los datos y para realizar análisis posteriores.

Las **medidas de posición** son parámetros estadísticos que permiten definir un conjunto de datos. Es decir, las medidas de posición ayudan a saber cómo es un conjunto de datos.

En estadística, existen dos tipos de medidas de posición: las **medidas de posición central**, que permiten determinar los valores centrales de un conjunto de datos, y las **medidas de posición no central**, que sirven para dividir los datos en intervalos iguales.

¿Cuáles son las medidas de posición?

En estadística, las medidas de posición son:

- **Medidas de posición central**: indican los valores centrales de una distribución.
    - **Media**: es el promedio de todos los datos de la muestra.
    - **Mediana**: es el valor del medio de todos los datos ordenados de menor a mayor.
    - **Moda**: es el valor que más se repite del conjunto de datos.
- **Medidas de posición no central**: dividen el conjunto de datos en partes iguales.
    - **Cuartiles**: dividen la muestra de datos en cuatro partes idénticas.
    - **Quintiles**: separan los datos en cinco partes iguales.
    - **Deciles**: parten el conjunto de datos en diez intervalos de la misma amplitud.
    - **Percentiles**: dividen los datos en cien partes equivalentes. A continuación se explica

cada tipo de medida de posición más detalladamente.

Medidas de posición central

Las **medidas de posición central** indican el valor central de una distribución, es decir, sirven para encontrar un valor representativo del centro de un conjunto de datos. Principalmente, existen tres métricas de posición central: la media, la mediana y la moda.

Media

Para calcular la **media** se deben sumar todos los valores y luego dividirlos entre el número total de observaciones. Por lo tanto, la fórmula de la media es la siguiente:

La media también se conoce como **media aritmética** o **promedio**. Además, la media de una distribución estadística es equivalente a su esperanza matemática.

Mediana

La **mediana** es el valor del medio de todos los datos ordenados de menor a mayor. Es decir, la mediana divide todo el conjunto de datos ordenados en dos partes iguales.

El cálculo de la mediana depende de si el número total de datos es par o impar:

- Si el número total de datos es **impar**, la mediana será el valor que está justo en el medio de los datos. Es decir, el valor que está en la posición $(n+1)/2$ de los datos ordenados.

- Si el número total de datos es **par**, la mediana será la media de los dos datos que están en el centro. Esto es, la media aritmética de los valores que están en la posiciones $n/2$ y $n/2+1$ de los datos ordenados.

Donde es el número total de datos de la muestra y *Me* es la mediana.

Moda

En estadística, la **moda** es el valor del conjunto de datos que tiene una mayor frecuencia absoluta, es decir, la moda es el valor que más se repite de un conjunto de datos.

Por lo tanto, para calcular la moda de un conjunto de datos estadísticos basta con contar el número de veces que aparece cada dato en la muestra, y el dato más repetido será la moda.

La moda también se puede decir **moda estadística** o **valor modal**.

Se pueden distinguir tres tipos de modas según el número de valores que están más repetidos:

- **Moda unimodal**: solo hay un valor con el máximo número de repeticiones. Por ejemplo, [1, 4, 2, 4, 5, 3].
- **Moda bimodal**: el máximo número de repeticiones se produce en dos valores diferentes y ambos valores se repiten el mismo número de veces. Por ejemplo, [2, 6, 7, 2, 3, 6, 9].
- **Moda multimodal**: tres o más valores tienen el mismo número máximo de repeticiones. Por ejemplo, [3, 3, 4, 1, 3, 4, 2, 1, 4, 5, 2, 1].

Medidas de posición no central

Las **medidas de posición no central** sirven para dividir el conjunto de datos estadísticos en intervalos iguales. Principalmente, se distinguen cuatro tipos de medidas de posición no central: los cuartiles, los quintiles, los deciles y los percentiles.

Cuartiles

En estadística, los **cuartiles** son los tres valores que dividen a un conjunto de datos ordenados en cuatro partes iguales. Por lo tanto, el primer, segundo y tercer cuartil representan respectivamente el 25%, 50% y 75% del conjunto de datos estadísticos.

Los cuartiles se representan mediante una Q mayúscula y el subíndice del cuartil, así pues, el primer cuartil es Q_1, el segundo cuartil es Q_2, y el tercer cuartil es Q_3.

Quintiles

Los **quintiles** son cuatro valores que dividen a un conjunto de datos ordenados en cinco partes iguales. De manera que el primer, segundo, tercer y cuarto quintil representan respectivamente al 20%, 40%, 60% y 80% de los datos de la muestra.

Por ejemplo, el tercer quintil es más grande que el 60% de todos los datos recopilados, pero es más pequeño que el resto de los datos.

El símbolo de los quintiles es la letra K mayúscula junto con el subíndice del quintil, esto es, el primer quintil es K_1, el segundo quintil es K_2, el tercer quintil es K_3, y el cuarto quintil es K_4. Aunque también se puede representar mediante la letra Q (no recomendable ya que genera confusión con los cuartiles).

Deciles

Los **deciles** son nueve valores que dividen a un conjunto de datos ordenados en diez partes iguales. De modo que el primer, segundo, tercer,... decil representa el 10%, 20%, 30%,... de la muestra o población.

Por ejemplo, el valor del cuarto decil es más grande que el 40% los datos, pero más pequeño que el resto de los datos.

En general, los deciles se representan mediante la letra D mayúscula y el subíndice del decil, es decir, el primer decil es D_1, el segundo decil es D_2, el tercer decil es D_3, etc.

Percentiles

Los **percentiles** son los valores que dividen a un conjunto de datos ordenados en cien partes iguales. De manera que un percentil indica el valor por debajo del cual se encuentra un porcentaje del conjunto de datos.

A modo de ejemplo, el valor del percentil 35 es más grande que el 35% de los datos observados, pero es más pequeño que el resto de datos.

Los percentiles se representan mediante la letra P mayúscula y el subíndice del percentil, es decir, el primer percentil es P_1, el percentil 40 es P_{40}, el percentil 79 es P_{79}, etc.

Los números índices simples son aquellos que se refieren únicamente a una variable o fenómeno. Así, al ser estar referidos a variables unidimensionales son meras relaciones porcentuales entre los valores del fenómeno entre los momentos del tiempo a comparar.

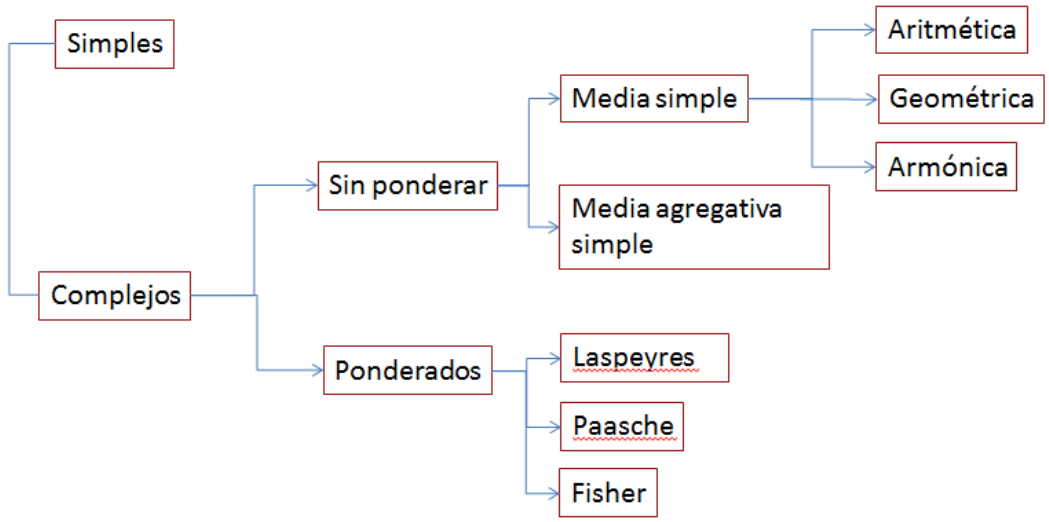
Se trata de una metodología que permite estudiar las variaciones de los distintos fenómenos, permitiendo comparar situaciones de distintos momentos del tiempo y/o el espacio. Podemos decir por tanto que las variaciones de un número índice nos indican los cambios de una magnitud que no es susceptible a una medición exacta, ni a una evaluación directa.

Un ejemplo de este tipo de situaciones es la variación de los *precios a lo largo del tiempo*.

La mayoría de los números índices que se emplean en la práctica están asociados a variables relacionadas con precios, salarios, producción, etc...

Clasificación de los números índices

Antes de continuar definiendo los números índices simples, presentaremos un esquema que muestra una clasificación de los números índices, simples y compuestos, cuyas definiciones iremos desarrollando a lo largo de este documento.



Números índices simples

Los números índices simples son aquellos que se refieren únicamente a una variable o fenómeno. Al estar referidos a variables

unidimensionales son meras relaciones porcentuales entre los valores del fenómeno entre los momentos del tiempo a comparar.

Definición.- Números índices simples

Si consideramos una variable a lo largo del tiempo X_t*, obtendremos sus índices dividiendo cada valor de la variable (correspondiente a un momento del tiempo) entre el valor de dicha variable en el instante que tomaremos como referencia y denominaremos periodo base (*X_0*).*

$$I_{t/0}(X) = \frac{X_t}{X_0}$$

Es habitual definir el índice en términos porcentuales, para ello solo debemos multiplicar por 100 la expresión anterior, es decir,

$$I_{t/0}(X) = \frac{X_t}{X_0} \times 100$$

Al calcular un número índice estamos haciendo en realidad un cambio de variable, convirtiendo la magnitud original X_t, en una magnitud $I(X)$. De este modo todos los estadísticos que hayamos obtenido para X_t podrán obtenerse para $I(X)$.

Nos interesará el incremento del índice, que medirá en cada caso la variación porcentual que se ha producido en la magnitud X desde el instante 0 hasta el instante t, y se expresará mediante:

$$\Delta I_{t/0}(X) = \frac{X_t - X_0}{X_0} \times 100 = \frac{X_t}{X_0} \times 100 - 100 = I_{t/0}(X) - 100.$$

Si $\Delta I_{t/0}(X) = 30$ podemos decir que la magnitud X se ha incrementado un 30% desde el instante 0 hasta el instante t.

Propiedades de los números índices simples.

1. Identidad

Si el periodo base y el de comparación coinciden el índice toma valor 100.

2. Inversión

Si el periodo base y el de comparación se invierten el índice toma el valor inverso al que tomaba originalmente, es decir,

$$I_{t/0}(X) = \frac{1}{I_{0/t}(X)}$$

3. Propiedad Circular

Si tomamos 3 instantes de tiempo que cumplan la siguiente relación (0<t'<t), para cualquier magnitud X se cumplirá que:

$$I_{t/0}(X) = I_{t'/0}(X) \times I_{t/t'}(X)$$

4. Propiedad de encadenamiento

Considerando los instantes de tiempo t=0,1,2,...t, siendo 0<1<2...<t, para cualquier magnitud X se cumplirá que:

$$I_{t/0}(X) = I_{1/0}(X) \times I_{2/1}(X) \times \ldots \times I_{t-1/t-2}(X) \times I_{t/t-1}(X)$$

5. Existencia

El índice tomará valores reales y finitos para cualquier valor de la variable observada.

6. Propiedad del producto

Si consideramos una magnitud compleja Z obtenida como producto entre dos magnitudes simples X e Y, se verifica que:

$$I_{t/0}(Z) = I_{t/0}(X) \times I_{t/0}(Y)$$

7. Propiedad del Cociente

Si consideramos una magnitud compleja Z obtenida como cociente entre dos magnitudes simples X e Y, se verifica que:

$$I_{t/0}(Z) = \frac{I_{t/0}(X)}{I_{t/0}(Y)}$$

8. Proporcionalidad

Si se produce un cambio de escala en la variable el índice será proporcional al original.

Ejemplo de cálculo de números índices simples

La siguiente tabla incluye los valores del precio del aceite en euros por tonelada a lo largo del tiempo. Obtenga el índice simple asociado a esta magnitud.

T	Precio (€/T)
0	3600
1	3750
2	3650
3	3550
4	3700
5	3800
6	3850

Números índices complejos. Definición.

Definición.- Números índices complejos

Los números índices complejos hacen referencia a varios conceptos a la vez y su evolución en el espacio y/o. Podemos decir que utilizan magnitudes complejas o variables n-dimensionales.

Un ejemplo de número índice complejo sería el IPC que nos aporta información sobre la variación de precios de los productos incluidos en la lista de la compra de las familias españolas.

Si todos los conceptos considerados tienen la misma importancia, se calculan para ellos Índices complejos sin ponderar. Cuando cada concepto tiene distinta importancia debemos emplear Índices complejos ponderados.

Números índices complejos sin ponderar.

Definición. - Números índices complejos sin ponderar

Consideraremos la magnitud compleja H formada por k magnitudes simples {$H_1, H_2,...,H_k$}. Para analizar la evolución de H debemos tener

en cuenta la evolución de todas las magnitudes simples que la componen.

Obtendremos por tanto el índice de H en función de los índices de{H_1, $H_{2,...,}H_k$}. Podremos hacerlo mediante sus medias simples (usando la media aritmética, la media geométrica o la media armónica) o bien mediante la media agregativa simple, es decir comparando simplemente las sumas de los distintos valores con el periodo de referencia.

Definición.- Índice de la media aritmética simple

El índice complejo de la media aritmética simple se obtiene mediante la media aritmética de los índices simples, es decir,

$$I_{t/0}(H) = \frac{1}{k}\sum_{i=1}^{k} I_{t/0}(H_i)$$

Los números índices complejos sin ponderar, presentan las siguientes limitaciones:

1. La unidades de medida de las distintas magnitudes simples pueden ser diferentes, lo que nos impedirá hacer comparaciones entre distintos índices.
2. Dan la misma importancia a cada magnitud simple.

Ejemplo de cálculo del número índice de la media aritmética simple

Una empresa fábrica un producto de coste H que depende de 3 componentes con la misma importancia relativa y costes H_1, H_2 y H_3. Calcule para H el índice de la media aritmética simple a partir de la información de la siguiente tabla:

t	H_1	H_2
0	3	1
1	3,5	3
2	3	3

3	2,5	2
4	3	4
5	4	5
6	4,5	7

Solución.-

El primer paso será calcular los índices simples para los precios de cada uno de los componentes:

t	H_1	H_2	H_3	$I_{t/0}(H_1)$
0	3	1	3	100
1	3,5	3	2	116,67
2	3	3	1	100
3	2,5	2	1	83,333
4	3	4	4	100
5	4	5	5	133,33
6	4,5	7	2	150

El índice complejo se obtiene calculando en cada momento la media aritmética simple de los índices simples:

t	H_1	H_2	H_3	$I_{t/0}(H_1)$	$I_{t/0}(H_2)$
0	3	1	3	100	100
1	3,5	3	2	116,67	300
2	3	3	1	100	300
3	2,5	2	1	83,333	200
4	3	4	4	100	400
5	4	5	5	133,33	500
6	4,5	7	2	150	700

Podemos decir que el coste del producto es un 61,11% mayor en el momento que en el momento 0.

En el momento 2 se ha producido un incremento en el coste del producto del 44,44% del coste inicial y así sucesivamente.

Números índices complejos ponderados.

- Los números índices complejos ponderados tienen en cuenta la importancia relativa de cada una de las magnitudes simples que componen el indice complejo.
- Para ponderar estas importancias relativas debemos asignar a cada una de las magnitudes un peso que denominaremos wi, de forma que la suma de los pesos asociados a todas las magnitudes simples debe ser igual a la unidad, es decir,

$$\sum_{i=1}^{k} w_i = 1$$

- Estos pesos pueden variar también a lo largo del tiempo por lo que definiremos como wit al peso referido a la importancia relativa de la magnitud i en el instante t.

Índice de Laspeyres

Considerando de nuevo la magnitud compleja H, formada por k magnitudes simples H_i, el índice de Laspeyres se define como la media ponderada de los índices simples de las magnitudes, es decir,

$$L_{t/0}(H) = \sum_{i=1}^{k} w_{i0} I_{t/0}(H_i)$$

A la hora de definir los pesos, tendremos en cuenta únicamente la importancia relativa de la magnitud simple i en el instante 0.

Ejemplo de cálculo del número Laspeyres

Consideremos de nuevo el coste del producto H que depende de los costes de los componentes H_1, H_2 y H_3 y supongamos ahora que en el instante 0, la importancia relativa de cada componente en el coste total, viene dada por los siguientes pesos: ($w_{10} = 0,2$; $w_{20} = 0,4$, $w_{30} = 0,4$). Obtenga en esta situación el índice de Laspeyres.

t	H_1	H_2	H_3	$I_{t/0}(H_1)$	$I_{t/0}(H_2)$
0	3	1	3	100	100
1	3,5	3	2	116,67	300
2	3	3	1	100	300
3	2,5	2	1	83,333	200
4	3	4	4	100	400
5	4	5	5	133,33	500
6	4,5	7	2	150	700

Lo habitual es que las ponderaciones de los precios se realicen en base a las cantidades empleadas. En este sentido, obtendremos las ponderaciones en el año base necesarias para el calculo del índice de Laspeyres mediante:

$$w_{i0} = \frac{p_{i0}q_{i0}}{\sum_{i=1}^{k} p_{i0}q_{i0}}$$

Así consideraremos la importancia relativa de cada artículo en el año base. El índice de precios de Laspeyres será entonces:

$$L_{t/0}(H) = \sum_{i=1}^{k} w_{i0} I_{t/0}(H_i) = \sum_{i=1}^{k} \frac{p_{i0}q_{i0}}{\sum_{i=1}^{k} p_{i0}q_{i0}} \times \frac{p_{it}}{p_{i0}} = \frac{\sum_{i=1}^{k} p_{it}q_{i0}}{\sum_{i=1}^{k} p_{i0}q_{i0}}$$

Índice de Paasche

Para la magnitud compleja H, formada por k magnitudes simples H_i, el índice de Paasche se define mediante,

$$P_{t/0}(H) = \sum_{i=1}^{k} \frac{I_{t/0}(H)}{w_{it}}$$

La principal diferencia con el índice de Laspeyres está en las ponderaciones, ya que el índice de Laspeyres se refiere únicamente a las ponderaciones en el periodo base y el de Paasche toma en cuenta las ponderaciones en cada instante t.

Como ya hemos visto, lo habitual es que las ponderaciones de los precios se realicen en base a las cantidades empleadas. Ahora debemos calcular ponderaciones diferentes según el momento del tiempo, y lo haremos mediante:

$$w_{it} = \frac{p_{it}q_{it}}{\sum_{i=1}^{k} p_{it}q_{it}}$$

Así consideraremos la importancia relativa de cada artículo en el año considerado. El índice de precios de Paasche será entonces:

$$P_{t/0}(H) = \sum_{i=1}^{k} \frac{\frac{p_{it}}{p_{i0}}}{\frac{p_{it}q_{it}}{\sum_{i=1}^{k} p_{it}q_{it}}} = \sum_{i=1}^{k} \frac{p_{it}\sum_{i=1}^{k} p_{it}q_{it}}{p_{i0}p_{it}q_{it}} = \frac{\sum_{i=1}^{k} p_{it}q_{it}}{\sum_{i=1}^{k} p_{i0}q_{it}}$$

Índice de Fisher

Dada una magnitud compleja H, formada por k magnitudes simples H i , el índice de Fisher se define como la raíz cuadrada del producto entre los índices de Paasche y Laspeyres, es decir,

$$F_{t/0}(H) = \sqrt{L_{t/0}(H) \times P_{t/0}(H)}$$

Así, el índice de Fisher es en sí un promedio de los anteriores.

Ejemplo de números complejos ponderados

Dada la siguiente información sobre precios en euros y cantidades vendidas de una serie de artículos:

	productos			
años	2015		2016	
	P	Q	P	Q
Patatas	1	150	1,2	140
Judías	2	430	1,9	420
Aceite	5	300	4,8	350
Pescado	15	250	16	220

Obtenga los índices de Laspeyres, Paasche y Fisher para cada año.

Solución.- Índice de Laspeyres:

$$L_{t/0}(H) = \frac{\sum_{i=1}^{k} p_{it} q_{i0}}{\sum_{i=1}^{k} p_{i0} q_{i0}}$$

Productos Años

	2015		2016		2017
	P	Q	P	Q	P
Patatas	1	150	1,2	140	1,3
Judías	2	430	1,9	420	1,95
Aceite	5	300	4,8	350	5,1
Pescado	15	250	16	220	15,5

$L_{t/0}(H)$	100	102,83	102,85

Solución.- Solución: Índice de Paasche:

$$P_{t/0}(H) = \frac{\sum_{i=1}^{k} p_{it} q_{it}}{\sum_{i=1}^{k} p_{i0} q_{it}}$$

	Años					
Productos	2015		2016		2017	
	P	Q	P	Q	P	Q
Patatas	1	150	1,2	140	1,3	160

Judías	2	430	1,9	420	1,95	450
Aceite	5	300	4,8	350	5,1	360
Pescado	15	250	16	220	15,5	280

$P_{t/0}(H)$	100	102,25	102,85

Solución: Índice de Fisher:

$$F_{t/0}(H) = \sqrt{L_{t/0}(H) \times P_{t/0}(H)}$$

	Años	
	2015	2016
Laspeyres	100	102,83
Paasche	100	102,25
Fisher	100	**102,54**