Estadística Descriptiva

Taller Estadística para el Periodismo de Datos

Alejandro Zappala

2020-01-14

Índice general

Acerca de estos apuntes

Estos apuntes corrresponden a la trascripción del Primer Tema de unos viejos apuntes de estadística elaborados por el equipo de la **Unidad Docente de Matemáticas del departamento de Ingeniería Topográfica y Cartografía** de la EU de Ingeniería Técnica Topográfica de la Universidad Politécnica de Madrid

Fueron impresos en el Departamento de Reprografía de la misma escuela en octubre de 1997

Los ejercicios han sido actualizados procurando respetar el original.

Acerca del Editor de este PDF

Alejandro Zappala Delgado es Ingeniero Técnico en Topografía, Cartografía y Geodesia. Obtuvo su **Máster en Cartografía y Geodesia** en la Universidad Politécnica de Madrid.

Como usuario activo de software libre ha impartido numerosos talleres y seminarios acerca de tecnologías relacionadas con el mundo cartográfico, el análisis de redes y la visualización de datos en general, así como mentorizado eventos relacionados con la cultura de datos.

A veces pasea un rato por Twitter, y puedes encontrar algunos de sus proyectos en Github.

Capítulo 1

Estadística descriptiva

1.1. POBLACIÓN Y MUESTRA

Población es un conjunto de elementos de los cuales nos interesa estudiar alguna característica común. El estudio que se haga servirá para conocer y describir a esa población.

Muestra es una parte de la población. Los resultados que se obtienen de su estudio se tratan de extrapolar para toda la población. Esta parte de la Estadística, la inferencia estadística, se estudiará más adelante.

La característica que queremos estudiar de la población presentará diversas modalidades que son los posibles valores que puede tomar.

Los caracteres pueden ser:

- **cualitativos**: las diversas modalidades no son valores numéricos. **Ejemplo**: "el color del pelo de un grupo de personas".
- **cuantitativos**: las diversas modalidades son números reales. **Ejemplo**: "el número de miembros de las familias que viven en Madrid".

Estos números son los diferentes valores que toma una variable estadística.

1.2. VARIABLES ESTADÍSTICAS

Variable estadística es una aplicación que asigna a cada elemento de la población un número real, que es el valor de la característica cuantitativa que estamos estudiando.

$$E = poblaciónE \to R$$

Una variable estadística es **discreta** si sus valores posibles pertenecen a un conjunto numerable. El caso más frecuente de variables discretas es aquel en que los valores posibles son números enteros. Así:

- El número de hijos en una familia.

- El número (o la proporción) de piezas defectuosas de un lote de 1000 piezas.

Una variable estadística es continua si sus valores posibles pertenecen a un conjunto no numerable

- El diámetro de una pieza.
- La temperatura de un cuerpo.

En general, todas las magnitudes, relacionadas con el espacio, con el tiempo, con la masa o bien las combinaciones de estos elementos son variables estadísticas continuas.

1.3. FRECUENCIAS

Sea una población de N elementos, de la cual estudiamos el carácter A que presenta las modalidades $A_1, A_2, ..., A_k$. Para cada modalidad se define:

- **Frecuencia absoluta**, n_i, es el número de elementos que poseen la modalidad A_1. Se tiene que $\sum_{i=1}^{j} n_i = N$
- **Frecuencia relativa**, f_i, es el cociente entre la frecuencia absoluta n_i y el número total de elementos N, es decir, $f_i = \frac{n_i}{N}$. Se tiene que $\sum_{i=1}^{k} f_i = 1$
- **Frecuencia absoluta acumulada**, N_i , es el número de elementos que poseen la modalidad A_i . o alguna de las anteriores (para lo cual tienen que estar ordenadas previamente), es decir, $N_i = \sum_{j=1}^{i} n_j$. Se tiene que $N_k = N$
- **Frecuencia relativa acumulada**, F_i , es el cociente entre la frecuencia absoluta acumulada y el numero total de elementos N, es decir, $F_i = \frac{N_i}{N}$. Se tiene que $F_k = 1$

1.4. DISTRIBUCIONES

Distribución de frecuencias: es el conjunto de modalidades con sus respectivas frecuencias. Según sean éstas (absolutas, relativas,. . .) así lo será la distribución correspondiente.

Las distribuciones de frecuencias se representan mediante tablas estadísticas.

Se clasifican en dos tipos:

- **Sin agrupar**: aparecen los datos individualizados con sus respectivas frecuencias. Se utiliza Cuando la variable torna pocos valores diferentes.
- **Agrupados en intervalos**: se divide el campo de la variable en intervalos llamados de clase, que tendrán como frecuencia el número de elementos que estén en el intervalo. Se utiliza cuando la variable toma muchos valores distintos entre sí.

La agrupación en intervalos tiene la ventaja de la simplicidad de los cálculos, y el inconveniente de la pérdida de información.

Los intervalos serán todos de la misma amplitud procurando que los datos se distribuyan más o menos homogéneamente a lo largo de todo el recorrido, de forma que no haya ninguna clase con muchos elementos (más del 30 %) ni varias clases con pocos o ningún elemento (menos del 5 %).

El nº de intervalos que se toma dependerá del número de datos y de la dispersión de los mismos

Algunos criterios a seguir es tomar k como el entero mas próximo a:

a) $1 + 3{,}3 \log_{10}(x)$
b) $2\sqrt{N}$

A los extremos del intervalo se les llama **límites de clase** (superior e inferior). Estos se deben tomar de forma que se solapen los intervalos, es decir, que el extremo superior de uno sea el inferior del siguiente.

Para evitar la ambigüedad que suponen los valores de la variable que coincidan con algún extremo, se pueden seguir dos criterios:

- Incluir siempre el extremo superior, pero no el inferior de cada clase, salvo en la primera, que se incluyen los dos. Es decir, tomar intervalos de la forma (a,b].
- Tomar los extremos de los intervalos con un décima más que los dados, de forma que no se pueda coincidir con ninguno de ellos.

Se llama **marca de clase** (x_i) al punto medio del intervalo de clase $e_{i-1} - e_{i-1}$. En todos los cálculos se opera como si la marca de clase tuviera la frecuencia absoluta de todo su intervalo.

La marca de clase se obtiene sumando los límites superior e inferior de clase y dividiendo por 2, es decir, $x_i = \frac{e_{i-1}+e_i}{2}$.

Tamaño de clase o amplitud de clase "**a**" es la diferencia entre los límites de clase $a = \frac{e_k - e_0}{k}$. La distribución de frecuencias quedaría así:

Intervalo	Marca de Clase x_i	Frecuencia absoluta n_i	Frecuencia relativa f_i	Frecuencia relativa acumulada F_i	Frecuencia absoluta acumulada N_i
$[e_0 - e_1]$	x_1	n_1	f_1	F_1	N_1
$(e_1 - e_2]$	x_2	n_2	f_2	F_2	N_2
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
$(e_{i-1} - e_i]$	x_i	n_i	f_i	F_i	N_i
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
$(e_{k-1} - e_k]$	x_k	n_k	f_k	F_k	N_k

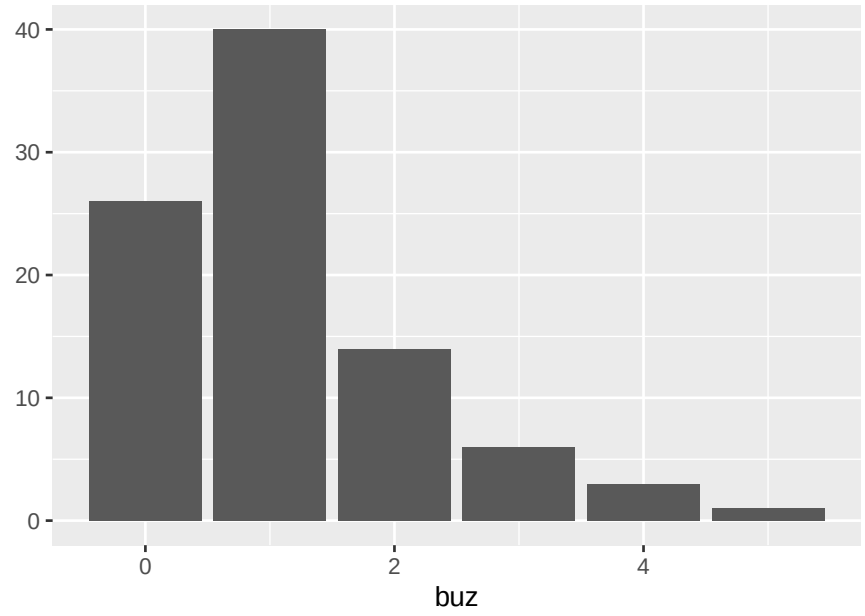
1.5. REPRESENTACIÓN GRÁFICA

De caracteres cuantitativos

a) Para variables discretas sin agrupar:

Diagrama de barras

Esta representación es válida para las frecuencias de una variable discreta, sin agrupar. Se colocan sobre el eje de las abscisas los distintos valores de la variable y sobre cada uno de ellos se levanta una línea o barra perpendicular, cuya altura es la frecuencia (absoluta, relativa) de dicho valor.



Polígono de frecuencias

Es una línea que se obtiene uniendo los extremos superiores de las barras en el diagrama de barras

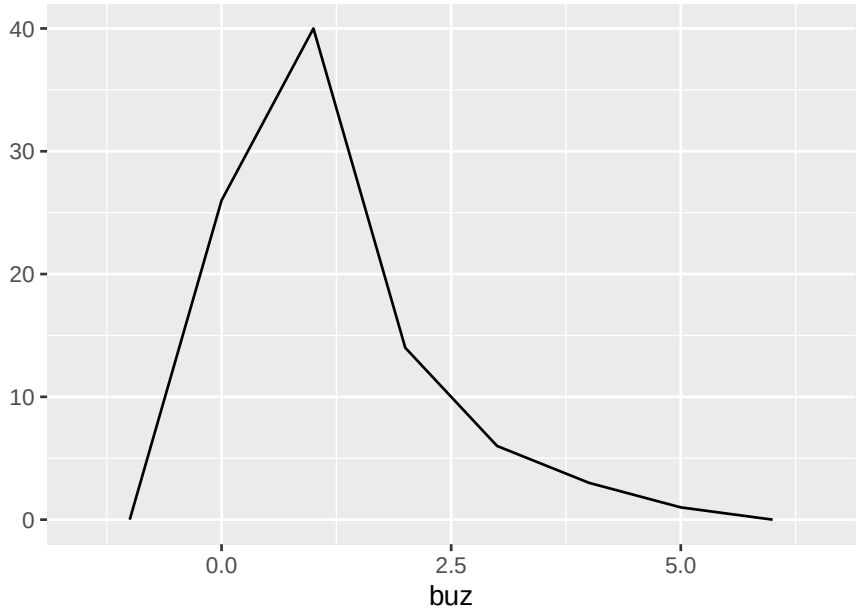
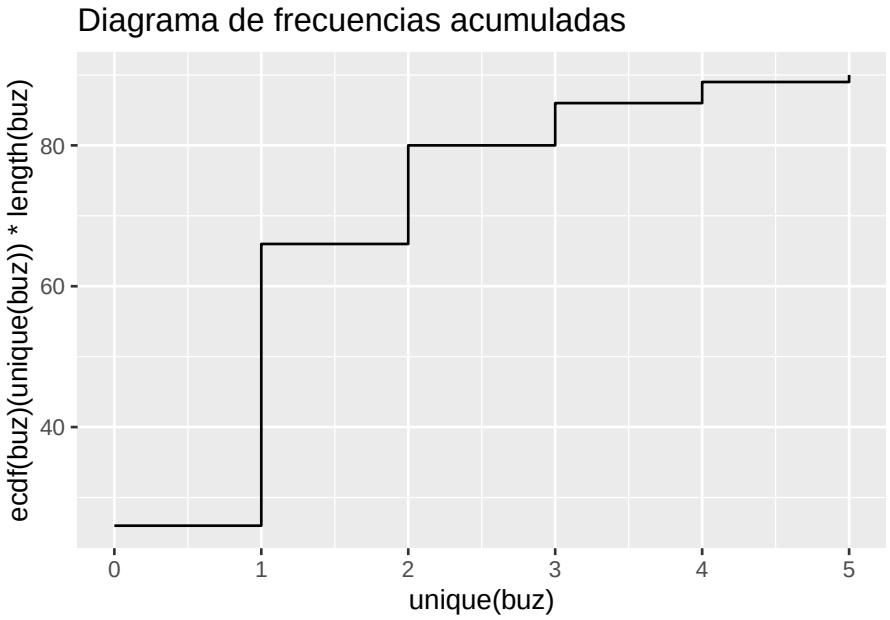


Diagrama de frecuencias acumuladas o diagrama de barras acumulativo

Representamos en el eje de abscisas los distintos valores de la variable estadística. Levantamos sobre cada uno de ellos una perpendicular cuya longitud será la frecuencia (absoluta o relativa) acumulada correspondiente a ese valor. De esta forma aparece un diagrama de barras creciente. Trazando segmentos horizontales de cada extremo de barra a cortar la barra situada a su derecha se obtiene el **diagrama de frecuencias acumuladas**

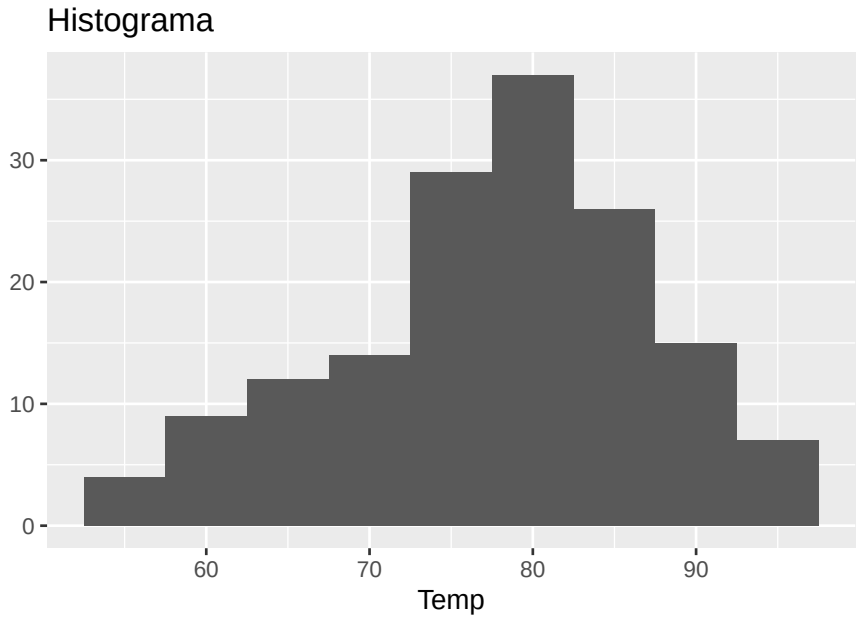


b) Para variables estadísticas agrupadas en intervalos de clase

Histograma

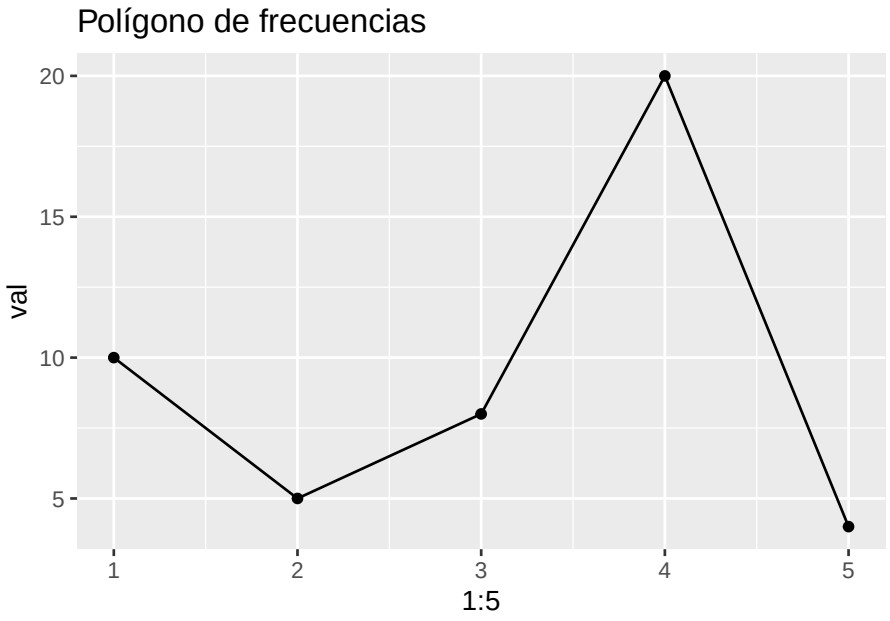
Es la representación gráfica más frecuente para datos agrupados.

En un histograma se representan las frecuencias mediante áreas. De tal forma que un histograma es un conjunto de rectángulos que tienen como base los intervalos de clase y cuya superficie son las frecuencias (absolutas o relativas). Por tanto las alturas son proporcionales a las frecuencias, y será el cociente entre la frecuencia y la amplitud del intervalo.



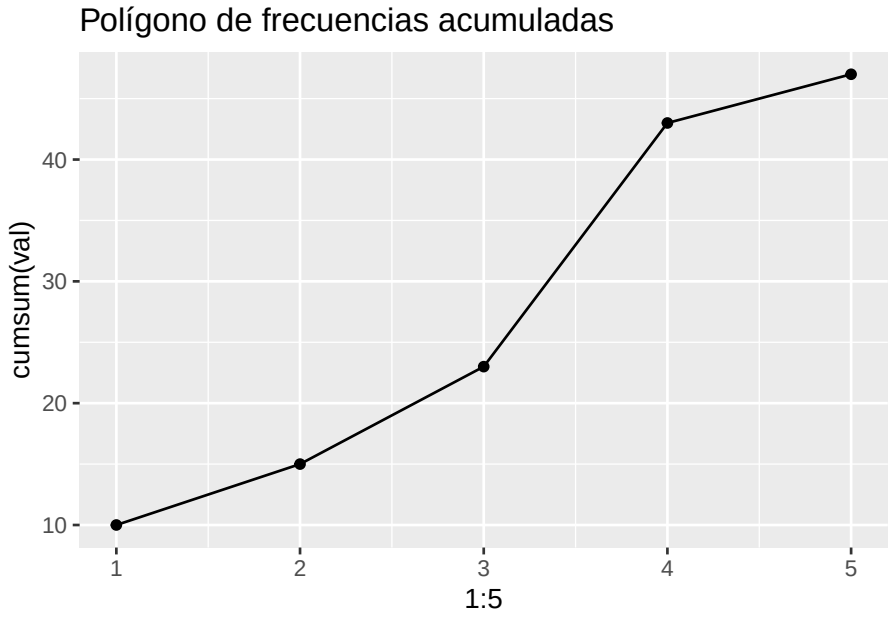
Polígono de frecuencias

El polígono de frecuencias es una línea que se obtiene uniendo los puntos medios de las bases superiores (los techos) de cada rectángulo en el histograma. De forma que empiece y acabe sobre el eje de abscisas, en el punto medio del que sería el intervalo anterior al primero y el último respectivamente.



Polígono de frecuencias acumuladas

En eje de abscisas representamos los distintos intervalos de clase que han de estar naturalmente solapados. Sobre el extremo superior de cada intervalo se levanta una línea vertical de longitud equivalente a la frecuencia (absoluta o relativa) acumulada del mismo. se obtiene así un diagrama de barras creciente , que uniendo sus extremos da lugar al polígono de frecuencias acumuladas absolutas o relativas



De caracteres cualitativos

Las representaciones gráficas más usuales son:

Diagrama de barras

Se representan en el eje de abscisas los distintos caracteres cualitativos y se levantan sobre ellos rectángulos de bases iguales que no tienen que estar solapados y cuyas alturas serán las correspondientes a la frecuencia absoluta de cada carácter.

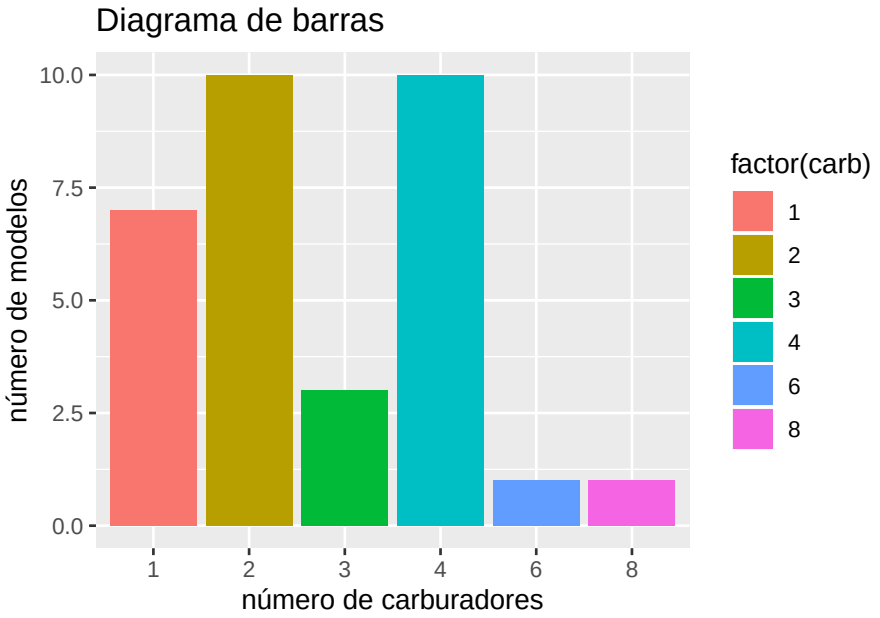
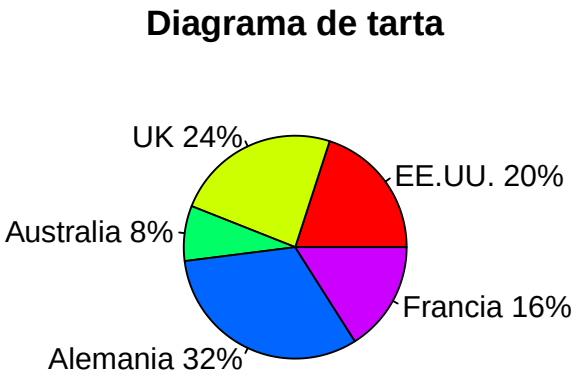


Diagrama de sectores (de tarta, de queso...)

En un círculo se asigna un sector circular a cada uno de los caracteres cualitativos, siendo la amplitud del sector proporcional a la frecuencia del carácter. No son muy recomendables, ya que es más fácil discernir con precisión tamaños entre longitudes que entre áreas.



Pictogramas

Cada modalidad se representa por un dibujo de tamaño proporcional a la frecuencia de la misma. También es frecuente tomar un dibujo estándar y repetirlo un número de veces proporcional a la frecuencia

Cartograma o Mapa Temático

Es la representación sobre mapas del carácter estudiado. Usualmente las distintas modalidades que adopta este carácter se representan con colores de distinta intensidad o distintas tramas.

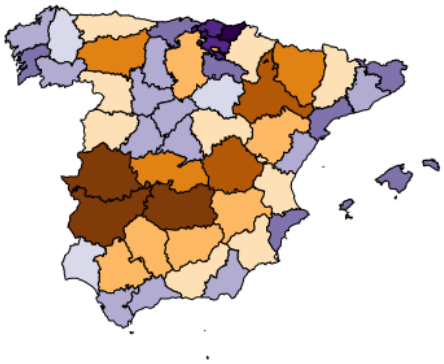


Figura 1.1: Cartograma

EJEMPLO 1.1. Los precios de 95 ordenadores portátiles en diciembre de 2019, dados en euros, son los siguientes:

3000	1200	740	1750	3409	580	840	1700	2300	715
3910	545	1380	815	565	3000	890	1580	800	3650
2240	1975	1745	3030	2350	3700	735	990	800	930
915	1100	1280	1163	1410	2050	3600	1260	1600	735
4260	1500	1000	1000	1600	1900	2150	2495	3200	850
540	2900	4500	3600	1035	1520	2495	1357	750	715
2775	2540	1470	395	3900	995	2200	900	1500	1500
1995	2650	1335	885	360	2100	2400	1200	1335	3310
600	755	500	990	765	1020	630	1555	640	950
630	1500	2300	3500	1825					

Se pide: a) Obtener una distribución de frecuencias agrupadas. b) Dibujar el histograma y polígono de frecuencias. c) Construir el polígono de frecuencias acumuladas

Solución

a) Obtener una distribución de frecuencias agrupadas.

El más caro es 4500 y el más barato 360, luego el recorrido es $4500-360 = 4140$

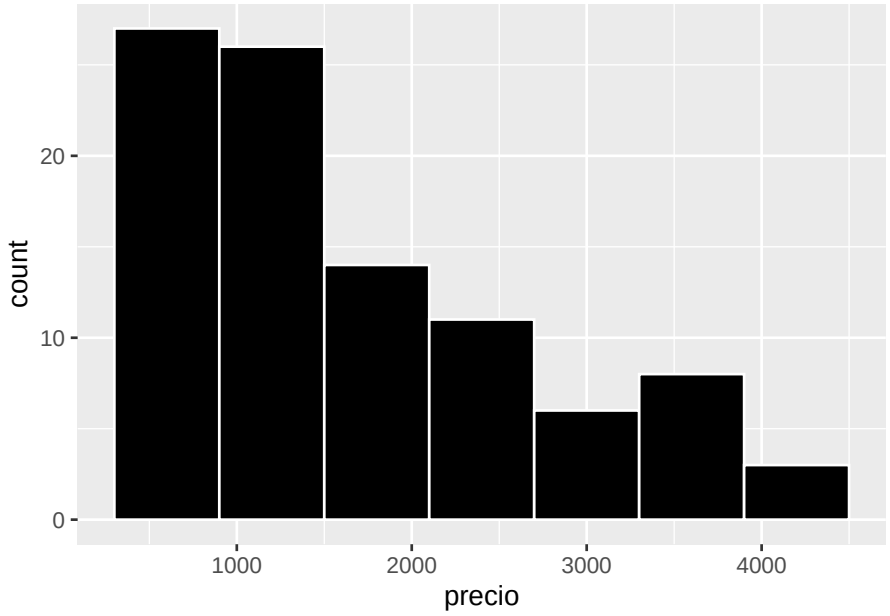
Elegimos intervalos de la misma amplitud de modo que, los datos se distribuyan de forma relativamente homogénea a lo largo del recorrido

Con el criterio , k igual al entero más próximo a $1+3{,}3\log_{10}(95) \approx 7{,}6$, elegimos 7 intervalos de amplitud 600y semiabiertos $(a, b]$

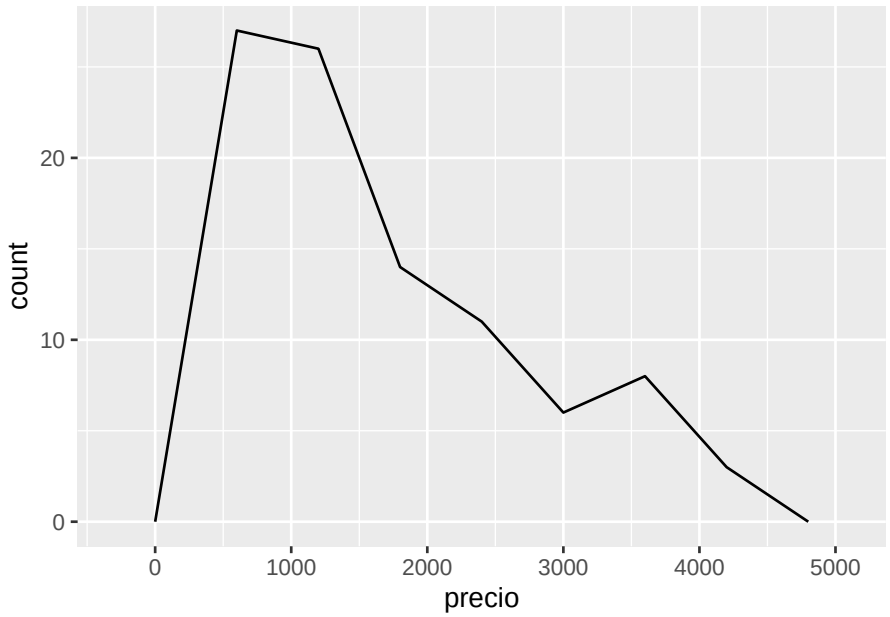
Distribución de frecuencias:

INTERVALOS	marca de clase	absoluta	absoluta acumulada	relativa	relativa acumulada
$e_{i-1} - e_i$	x_i	n_i	N_i	f_i	F_i
[300-900]	600	27	27	0.2842	0.284
(900-1500]	1200	26	53	0.2737	0.558
(1500-2100]	1800	14	67	0.1474	0.705
(2100-2700]	2400	11	78	0.1158	0.821
(2700-3300]	3000	6	84	0.0632	0.884
(3300-3900]	3600	8	92	0.0842	0.968
(3300-4500]	4200	3	95	0.0316	1.000

Histograma:



Polígono de frecuencias



1.6. MEDIDAS DE CENTRALIZACIÓN

Nos dan una idea de los valores de la variable alrededor de los que se agrupa la distribución.

Media aritmética (mean)

La **media** de una variable estadística es la suma ponderada de los valores posibles por sus respectivas frecuencias

$$\overline{X} = \sum_{i=1}^{k} f_i x_i = \sum_{i=1}^{k} \frac{n_i}{N} x_i = \frac{1}{N} \sum_{i=1}^{k} n_i x_i$$

- x_i = valores que toma la variable o marca de clase
- f_i = frecuencias relativas ($f_i = \frac{n_i}{N}$)
- n_i = frecuencias absolutas
- N = número total de la población o muestra

Propiedades

1) La media de las diferencias a la media es nula, es decir, la suma de las desviaciones de los valores de la variable estadística respecto de su media es cero.

$$\sum_{i=1}^{k}(x_i - \overline{X})f_i = \sum_{i=1}^{k} f_i x_i - \sum_{i=1}^{k} \overline{X} f_i = \overline{X} - \overline{X} \sum_{i=1}^{k} f_i = \overline{X} - \overline{X} = 0$$

2) La suma de los cuadrados de las desviaciones de los valores de la variable respecto de cualquier número α es mínimo para $\alpha = \overline{X}$

Mediana (median)

La **mediana** es el valor de la variable que ocupa el lugar central, es decir, que la mita de la población es menor y la otra mitad mayor que él.

Por tanto, supuestos los valores de la variable ordenados en forma creciente, deja igual número de observaciones inferiores que superiores a ella.

La mediana es un valor M tal que $F(M) = 1/2$, se define así como la raíz de una ecuación.

Forma de calcularla:

a) En distribuciones sin agrupar

En las distribuciones sin agrupar, en general, no tiene solución, puesto que la función $F(x)$ varía por saltos:

1) Si ningún coincide

Si ningún valor posible x_i corresponde a $F(x_i) = 1/2$ se conviene considerar como mediana el valor x_i tal que:

$$F(x_{i-1}) < \frac{1}{2} < F(x_i)$$

o lo que es igual:

$$f_1 + f_2 + \cdots + f_{i-1} < \frac{1}{2} + f_2 + \cdots + f_{i-1} + f_i$$

y con frecuencias absolutas:

$$n_1 + n_2 + \cdots + n_{n-1} < \frac{N}{2} < n_1 + n_2 + \cdots + n_{i-1} + n_i$$

EJEMPLO 1.1. : Sea la variable estadística $X = (5, \underline{1}, 5, 2, 4, 2, 3, 6, 5)$, ordenando los valores y tachando desde los extremos se tiene $X = (\underline{1}, \underline{2}, \underline{2}, \underline{3}, 4, \underline{5}, \underline{5}, \underline{5}, \underline{6})$, resultando el termino central $M = 4$

2) Si uno de los valores x_i corresponde a $F(x_i) = \frac{1}{2}$ (lo que ocurre solamente si el total N de la población es par) la mediana está indeterminada entre los valores x_i y x_{i+1}. El intervalo (x_i, x_{i+1}) se denomina mediano, o bien llamamos mediana al punto medio de dicho intervalo.

EJEMPLO 1.2. : Sea la variable estadística $X = (4, 1, 5, 2, 2, 3, 4, 5)$, entonces $X = (\underline{1}, \underline{2}, \underline{2}, 3, 4, \underline{4}, \underline{5}, \underline{5})$, resultando el intervalo mediano $[3, 4]$, o bien \$M=3.5.

En la tabla estadística, la mediana se determina a partir de la columna que da las frecuencias (o las frecuencias absolutas) acumuladas.

b) En distribuciones agrupadas

En las agrupadas pueden darse dos casos:

INTERVALO	x_i	n_i	Ni
$e_0 - e_1$	x_1	n_1	N_1
$e_1 - e_2$	x_2	n_2	N_2
$\cdots$	$\cdots$	$\cdots$	$\cdots$
$e_{j-1} - e_j$	x_j	n_j	N_j
$\cdots$	$\cdots$	$\cdots$	$\cdots$
$e_{k-1} - e_k$	x_k	n_k	N

1) $\frac{N}{2}$ coincide con uno de los recogidos en la columna de frecuencias acumuladas, por ejemplo N_j, en este caso la mediana es e_j.

2) $\frac{N}{2}$ está entre N_{j-1} y N_j. La mediana se encontrará en el intervalo $(e_{j-1} -$

e_j). La mediana será $M = e_{j-1} + h$ y por interpolación lineal (regla de tres) se obtiene h.

Amplitud del intervalo $a = e_j - e_{j+1}$

Comparación entre media y mediana

Ambas os quieren dar una idea de sobre qué valores está centrada la distribución. Son de cálculo sencillo..

La mediana no está influenciada por la existencia de algún valor muy extremo (outlier) y la media sí. Por ejemplo, si se consideran las rentas de los individuos de una país, se obtiene que hay un pequeño número con grandes rentas y un número muy grande con pequeñas rentas, entonces la media no refleja la situación real y, en cambio, la mediana caracteriza mejor el valor central.

La media tiene la ventaja de ser susceptible a operaciones algebraicas y la mediana no.

Moda

Moda es el valor de la variable que se presenta con más frecuencia dentro de la distribución.

En las distribuciones sin agrupar se observa directamente el valor de mayor frecuencia.

En las agrupadas, definimos la clase **modal** como la que tiene más frecuencia.

NOTA: Algunas distribuciones pueden presentar varias modas. Cada moda corresponde a un máximo absoluto del diagrama de barras o histograma.

Veamos la diferencia entre media, moda y mediana en un ejemplo:

EJEMPLO 1.3.

Consideremos los salarios de los empleados en una fábrica, el director gana 25800 €/mes, el subdirector 10700 €/mes, seis jefes 2700 €/mes cada uno, los cinco capataces 2150 y los diez operarios 1075 cada uno. ¿Cuál es el salario medio? ¿Cuál el salario mediano? Y ¿Cuál el salario modal?

Solución:

x_i	1075	2150	2700	10700	25800
n_i	10	5	6	1	1
N_i	10	15	21	22	23

$\overline{X} = \sum_{i=1}^{k} \frac{n_i x_i}{N} = \frac{74200}{23} = 3226{,}087$ €/mes de media aritmética

$\frac{N}{2} = \frac{23}{2} = 11{,}5 \to N_1 = 10 < 11{,}5 < 15 = N_2$ luego la mediana es $x_4$2150 y la moda corresponde a $n_1 = 10$ que es 1075 €/mes

¿Cuál de los valores anteriores describe mejor los sueldos percibidos por los empleados de esta fábrica?

Cuantiles

Cuantil de orden α es un valor de la variable estadística que deja a su izquierda una parte α de la población y a la derecha una parte $1-\alpha$ de la población.

El Cuantil de orden $\alpha(0 \leq \alpha \leq 1)$ es $x_\alpha \mid F(x_\alpha) = \alpha$

Los más utilizados son los **cuartiles** Q_1, Q_2 y Q_3 que dejan a su izquierda $\frac{1}{4}$, $\frac{1}{2}$ y $\frac{3}{4}$ de la población respectivamente.

Obsérvese que $Q_2 = M$ (Mediana)

Los **deciles** D_1, D_2, $\ldots$, D_9 dejan a su izquierda $\frac{1}{10}, \frac{2}{10}, \ldots, \frac{9}{10}$ de la población respectivamente.

Los **percentiles** P_1, P_2, $\ldots$, P_{9p} dejan a su izquierda $\frac{1}{100}, \frac{2}{100}, \ldots, \frac{9}{100}$ de la población respectivamente.

El cálculo de los mismos es similar al cálculo de la mediana.

1.7. MEDIDAS DE DISPERSIÓN

Las medidas de dispersión nos dan una idea de la mayor o menor concentración de los valores de la variable alrededor de algún valor.

EJEMPLO: Si consideramos 8 alumnos con calificación de 10 y 8 alumnos con un cero; la media aritmética será 5. Si los 16 alumnos tienen un 5, la media aritmética también será cinco, sin embargo, las dos situaciones son claramente distintas y la media es más representativa en el segundo caso, al estar los valores concentrados en un único valor. La diferencia entre uno y otro caso se pone de manifiesto con las medidas de dispersión.

Rango o recorrido

Es la diferencia entre el mayor y el menor valor de la variable estadística.

EJEMPLO: Si las observaciones son: $8, 3, 5, 7, 1, 1, 8$, el recorrido es $8 - 1 = 7$

Recorrido semi-intercuartílico

El **recorrido semi-intercuantílico** o desviación cuartílica es la diferencia entre los cuartiles Q_1 y Q_3 dividida entre dos: $\frac{Q_3 - Q_1}{2}$.

Varianza

Es la media de los cuadrados de las desviaciones a la media

$\sigma^2 = \frac{\sum_{i=1}^{k}(x_i - \bar{X})^2 n_i}{N}$

$x_1 =$ valores de la variable o marca de clase.

Resulta igual a la media de los cuadrados menos el cuadrado de la media

$\sigma^2 = \frac{\sum_{i}^{j} x_i \check{s}}{N} - (\bar{X})^2$

Varianza muestral o cuasivarianza

Si los datos que estamos analizando son todos los elementos de la población, la varianza nos es útil para analizar la dispersión de dichos datos respecto de la media. Por el contrario, si los datos son una muestra de la población, la varianza nos sirve para analizar la dispersión de dicha muestra pero, ¿nos sirve para analizar la dispersión de a población?

Si tomamos ahora otra muestra de la misma población, obtendremos una varianza para esa segunda muestra, etc. Para cada muestra tendremos el valor de su varianza correspondiente. Sería deseable que la media de todo ese conjunto de varianzas muestrales fuese la varianza de la población. Pero esto no es así (como se demuestra en Inferencia Estadística).Sin embargo, sí hay una medida de dispersión, la cuasivarianza que cumple esa condición, es decir tal que la media de las cuasivarianzas muestrales es la varianza de la población. Es por esto por lo que a veces se calcula la cuasivarianza (varianza muestral) en lugar de la varianza.

La **varianza muestral** viene dada por:

$S^2 = \frac{N}{N-1}\sigma^2$, es decir: $S^2 = \frac{N}{N-1}\frac{\sum_{i=1}^{k}(x_i-\overline{X})^2}{N} = \frac{\sum_{i=1}^{k}(x_i-\overline{X})^2}{N-1}$

Nótese que para N suficientemente grande, la diferencia entre σ^2 y S^2 es muy pequeña.

Desviación típica

La **desviación típica** o desviación cuadrática media es la raíz cuadrada positiva de la varianza:

$\sigma = +\sqrt{\sigma^2} = +\sqrt{\sum_{i=1}^{k}(x_i - \overline{X})^2 f_i}$ o bien, $\sigma = +\sqrt{\sum_{i=1}^{k} x_i f_i - \overline{X}^2}$

Tiene la ventaja sobre la varianza de que está en las mismas unidades que la variable.

Desviación típica muestral

La **desviación típica muestral** es la raíz cuadrada positiva de la varianza muestral.

$$S = +\sqrt{\sum_{i=1}^{k}\frac{(x_i - \overline{X})^2 n_i}{(N-1)}} = \sqrt{\frac{N}{N-1}}\sigma$$

Coeficiente de variación de Pearson

Es el cociente de la desviación típica y la media: $CV = \frac{\sigma}{\overline{X}}$

Es independiente de la unidad que se utilice, pues no tiene unidades y por tanto nos permite comparar la dispersión de dos distribuciones que tengan unidades muy diferentes, o que tengan medias muy distintas.

EJEMPLO 1.4.: Con los datos del ejemplo 1.1., calcular:

a) Media aritmética, moda y mediana.
b) Varianza y desviación típica.
c) Varianza y desviación típica muestral.
d) Coeficiente de variación de Pearson.

Solución:

x_i	n_i	$x_i n_i$	Ni	x_i^2	$x_i^2 n_i$
0	26	0	26	0	0
1	40	40	66	1	40
2	14	28	80	4	56
3	6	18	86	9	54
4	3	12	89	16	48
5	1	5	90	25	25
	90	103			223

Momentos

Se llama **momento de orden r** respecto al valor "c", a la cantidad:

$\sum_{i=1}^{k}(x_i - c)^r f_i = \sum_{i=1}^{k}(x_i - c)^r \frac{n_i}{N}$, donde r es un entero positivo.

Según los valores de "c", se definen varias clases de momentos:

Momentos no centrales o respecto al origen

$c = 0 \rightarrow m_r = \sum_{i=1}^{k} x_i^r f_i = \sum_{i=1}^{k} x_i^r \frac{n_i}{N}$

Los primeros momentos no centrales son iguales a:

$m_0 = 1 : m_1 = \overline{X}$

Momentos centrales o respecto a la media

$c = \overline{X} \rightarrow \mu_0 = \sum_{i=1}^{k}(x_i - \overline{X})^r f_i = \sum_{i=1}^{k}(x_i - \overline{X})^r \frac{n_i}{N}$

Los primeros momentos centrales son iguales a:

$\mu_0 = 1; \mu_1 = 0 \ ; \ \mu_2 = \sigma^2 = m_2 - m_1^2$

1.8. CARACTERÍSTICAS DE FORMA

Además de la tendencia central y de la dispersión se puede tratar de caracterizar la forma de una distribución mediante un índice resumido que nos determina la **asimetría** o el **apuntamiento** de una distribución.

Sesgo

Si la muestra es simétrica respecto de la media, entonces $\sum_{i=1}^{k}(x_i - \overline{X})^3 = 0$ mientras que esta suma será mayor en valor absoluto cuanto más asimétricos sean los datos.

Para obtener una medida adimensional (sin unidades), se define el **coeficiente de asimetría** o **sesgo** o **coeficiente de Fisher** como:

$$g_1 = \frac{\sum_{i=1}^{k}(x_i - \overline{X})^3 f_i}{\sigma^3} = \frac{\mu_3}{\sigma^3}$$

Si la muestra no es simétrica, diremos que es **sesgada**. Será **sesgada a la derecha** si $g_1 > 0$, y sesgada a la izquierda, si $g_1 < 0$.

Curtosis

El **coeficiente de Curtosis** es el grado de apuntamiento de una distribución. Será mayor cuanto mayor sea la concentración de valores alrededor de la media.

Se mide en relación a la distribución Normal, de la misma media y desviación típica.

El coeficiente de apuntamiento de Fisher es: $g_2 = \frac{\mu_4}{\sigma^4} - 3$

De forma que es nulo para la distribución normal. Si el coeficiente es positivo la distribución está más apuntada que la distribución Normal (de la misma media y desviación típica), y se dice **leptocúrtica**. Si es menos apuntada el coeficiente es negativo y se dice **platicúrtica**. **Mesocúrtica** es cuando el coeficiente es nulo.