Resumen y gráficas de datos

Universidad de Guayaquil

junio 04, 2019

Características importantes de los datos (I)

1. **Centro:** Valor promedio o representativo que indica la localización de la mitad del conjunto de los datos.
2. **Variación:** Medida de la cantidad en que los valores de los datos varían entre sí.
3. **Distribución:** La naturaleza o forma de la distribución de los datos (como en forma de campana, uniforme o sesgada).
4. **Valores extremos:** Valores muestrales que están muy alejados de la vasta mayoría de los demás valores de la muestra.
5. **Tiempo:** Características cambiantes de los datos a través del tiempo.

Distribución de frecuencias (I)

Definición

Una distribución de frecuencias (o tabla de frecuencias) lista valores de los datos (ya sea de manera individual o por grupos de intervalos), junto con sus frecuencias (o conteos) correspondientes.

Distribución de frecuencias: Edades de las mejores actrices

Edad de las actrices	Frecuencias
21–30	28
31–40	30
41–50	12
51–60	2
61–70	2
71–80	2

Distribución de frecuencias (II)

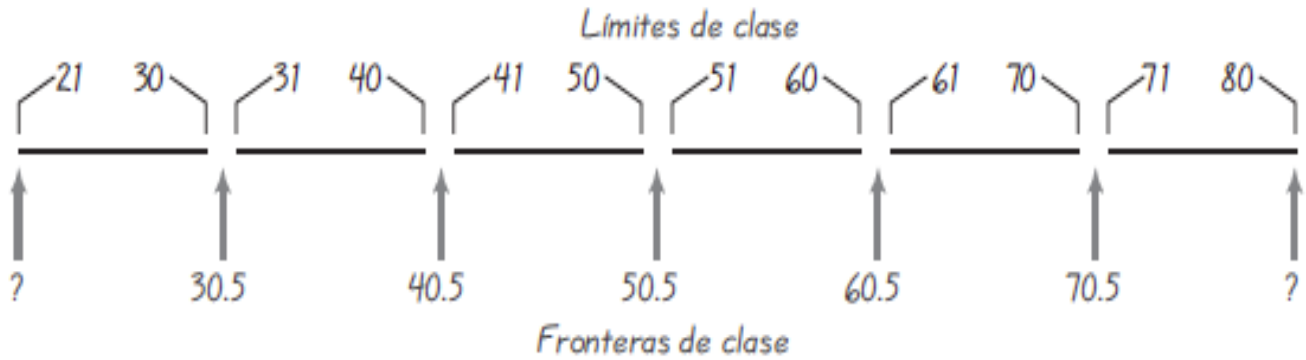
Definiciones

- Los **límites de clase inferiores** son las cifras más pequeñas que pueden pertenecer a las diferentes clases.
- Los **límites de clase superiores** son las cifras más grandes que pueden pertenecer a las diferentes clases.
- Las **fronteras de clase** son las cifras que se utilizan para separar las clases, pero sin los espacios creados por los límites de clase.
- Las **fronteras de clase** son muy útiles en la siguiente cuando elaboremos una gráfica llamada histograma.
- Las **marcas de clase** son los puntos medios de las clases. Las marcas de clase se calculan sumando el límite de clase inferior con el límite de clase superior, y dividiendo la suma entre 2.
- La **anchura de clase** es la diferencia entre dos límites de clase inferiores consecutivos o dos fronteras de clase inferiores consecutivas.

Distribución de frecuencias (III)

Distribución de frecuencias: Edades de las mejores actrices

Edad de las actrices	Frecuencias
21–30	28
31–40	30
41–50	12
51–60	2
61–70	2
71–80	2



Procedimiento para construir una distribución de frecuencias (I)

1. Es posible resumir conjuntos grandes de datos
2. Se logra cierta comprensión sobre la naturaleza de los datos
3. Se tiene una base para construir gráficas importantes, como por ejemplo los histogramas

EJEMPLO Edades de las mejores actrices

Use las edades de las mejores actrices de la tabla para construir la distribución de frecuencias. Suponga que desea incluir 6 clases.

SOLUCIÓN (I)

Paso 1: Comience seleccionando 6 clases.
Paso 2: Calcule la anchura de clase.

Procedimiento para construir una distribución de frecuencias (II)

$$Anchura \approx \frac{(valor\ más\ alto)\ -(valor\ más\ bajo)}{número\ de\ clases} = \frac{80-21}{6} = 9{,}833 \approx 10 \quad (1)$$

Paso 3: Elegimos un punto de partida de 21, que es el valor más bajo de la lista y un número conveniente

Paso 4: Sume la anchura de clase 10 al punto de partida 21 para determinar que el segundo límite inferior de clase es igual a 31. Continúe y sume la anchura de clase 10 para obtener los límites inferiores de clase restantes de 41, 51, 61 y 71.

Paso 5: Liste los límites de clase inferiores de forma vertical, como se muestra al margen. Con esta lista podemos identificar con facilidad los límites de clases superiores correspondientes, que son 30, 40, 50, 60, 70 y 80.

Paso 6: Después de identificar los límites inferiores y superiores de cada clase, proceda a trabajar con el conjunto de datos asignando una marca a cada valor. Una vez completadas las marcas, súmelas para obtener las frecuencias.

Distribución de frecuencias relativas

Una distribución de frecuencias relativas incluye los mismos límites de clase que una distribución de frecuencias, pero utiliza las frecuencias relativas en vez de las frecuencias reales.

$$frecuencia\ relativa\ = \frac{frecuencia\ de\ clase}{suma\ de\ todas\ las\ frecuencias} \tag{2}$$

Distribución de las frecuencias relativas de las edades de las mejores actrices

Edad de la actriz	Frecuencia relativa
21–30	37%
31–40	39%
41–50	16%
51–60	3%
61–70	3%
71–80	3%

Distribución de frecuencias acumulativas

La frecuencia acumulativa de una clase es la suma de las frecuencias para esa clase y todas las clases anteriores.

Distribución de las frecuencias relativas de las edades de las mejores actrices

Edad de la actriz	Frecuencia relativa
21–30	37%
31–40	39%
41–50	16%
51–60	3%
61–70	3%
71–80	3%

Distribución normal

Una característica fundamental de una distribución normal es que, cuando se grafica, el resultado tiene forma de "campana"; y al inicio las frecuencias son bajas, luego se incrementan hasta un punto máximo y luego disminuyen.

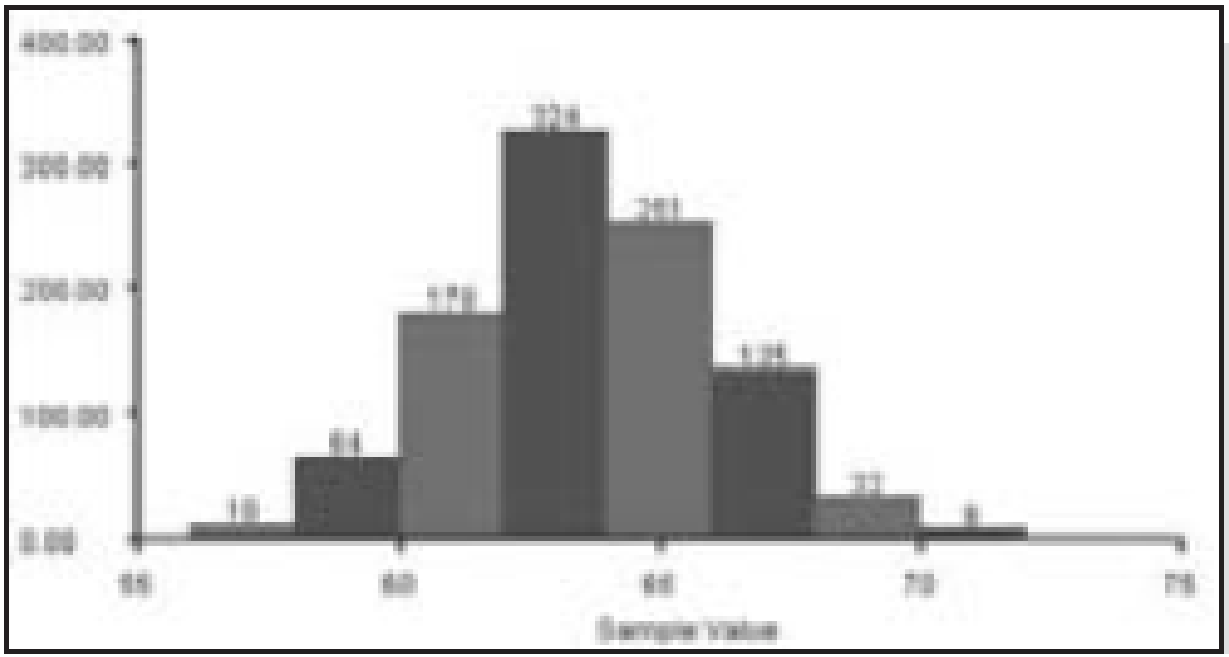
Distribución normal: Al inicio las frecuencias son bajas, luego alcanzan un nivel máximo y disminuyen nuevamente.

Estatura (pulgadas)	Frecuencia	Distribución normal:
56.0–57.9	10	← Al inicio las frecuencias son bajas, . . .
58.0–59.9	64	
60.0–61.9	178	
62.0–63.9	324	← aumentan hasta un punto máximo, . . .
64.0–65.9	251	
66.0–67.9	135	
68.0–69.9	32	
70.0–71.9	6	← disminuyen nuevamente.

Histogramas Normal

Una característica fundamental de una distribución normal es que, cuando se grafica en un histograma, el resultado es una curva en forma de "campana", como en el histograma de la figura. Las principales características de la curva en forma de campana son

1. el aumento de las frecuencias, las cuales alcanzan un punto máximo y luego disminuyen;
2. la simetría, donde la mitad izquierda de la gráfica es casi una imagen en espejo de la mitad derecha.



Distribución de frecuencias

Tabla 2-7
Monedas de un centavo elegidas al azar

Pesos de monedas de un centavo (gramos)	Frecuencia
2.40–2.49	18
2.50–2.59	19
2.60–2.69	0
2.70–2.79	0
2.80–2.89	0
2.90–2.99	2
3.00–3.09	25
3.10–3.19	8

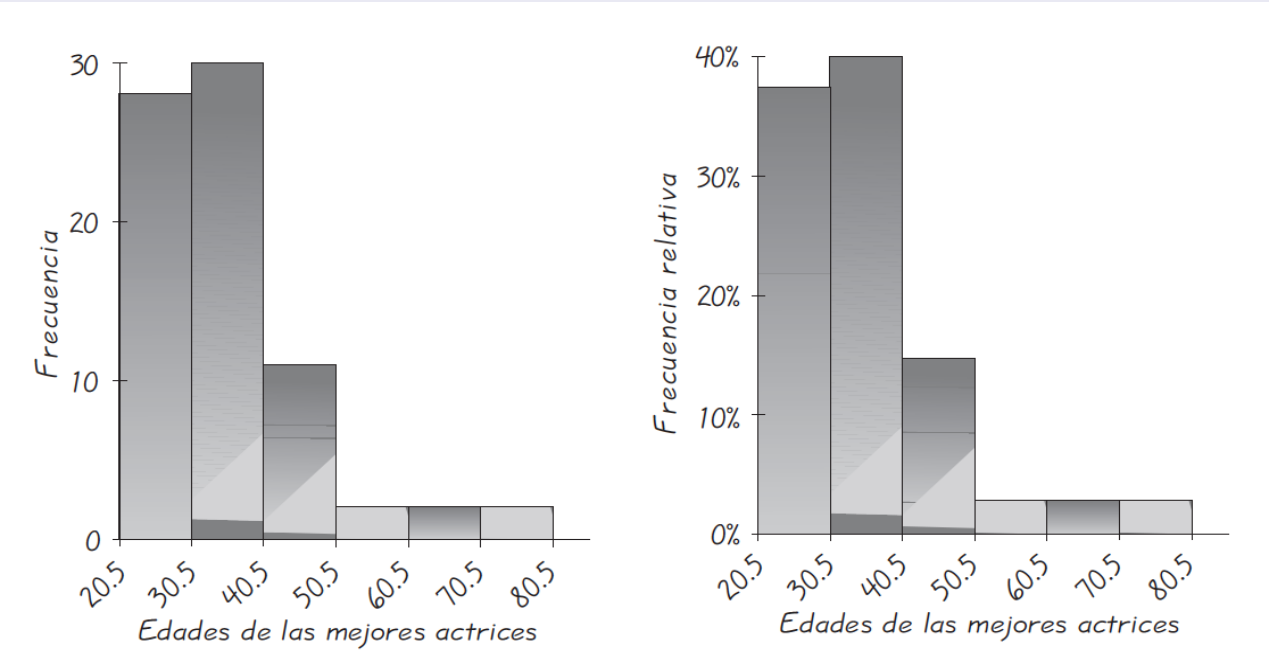
Tabla 2-8
Edades de actores y actrices ganadores del Óscar

Edad	Actrices	Actores
21–30	37%	4%
31–40	39%	33%
41–50	16%	39%
51–60	3%	18%
61–70	3%	4%
71–80	3%	1%

Histogramas

Definición

Un histograma es una gráfica de barras donde la escala horizontal representa clases de valores de datos y la escala vertical representa frecuencias. Las alturas de las barras corresponden a los valores de frecuencia; en tanto que las barras se dibujan de manera adyacente (sin huecos entre sí).



Histogramas en R (I)

```
x=c(3,2,5,1,3,1,5,6,2,2,2,1,3,5,2)# ingresar un vector cualquiera
table(x)# calcula frecuencia absoluta
```

```
## x
## 1 2 3 5 6
## 3 5 3 3 1
```

```
table(x)[4]# accedo a la cuarta entrada o cuarto nivel del vector x
```

```
## 5
## 3
```

```
sum(table(x))#sumo todos los elementos
```

```
## [1] 15
```

Histogramas en R (II)

```
x=c(3,2,5,1,3,1,5,6,2,2,2,1,3,5,2)# ingresar un vector cualquiera
prop.table(table(x))#frecuencias relativas
```

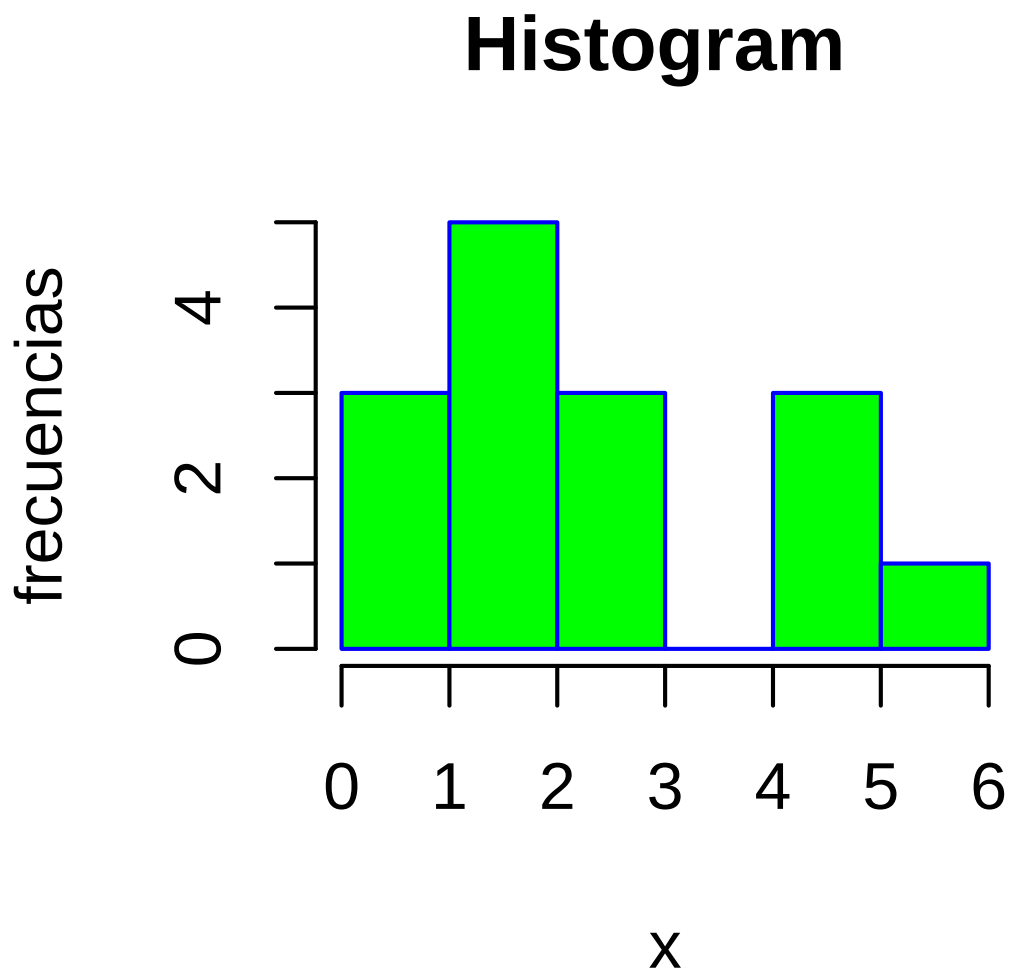
```
## x
##         1         2         3         5         6
## 0.20000000 0.33333333 0.20000000 0.20000000 0.06666667
```

```
cumsum(table(x))# suma acumulativa
```

```
##  1  2  3  5  6
##  3  8 11 14 15
```

Histogramas en R (III)

```
hist(x, main="Histogram ",xlab="x",
     ylab="frecuencias",border="blue",
     col="green",breaks = c(0:6))
```



Histogramas en R (IV)

```
#cargar datos a partir de una hoja de excel
library(readxl)
calificacion=read_excel("/Users/Gustavo/AnacondaProjects/R_projects/
table(calificacion$notas)#frecuencias absolutas
```

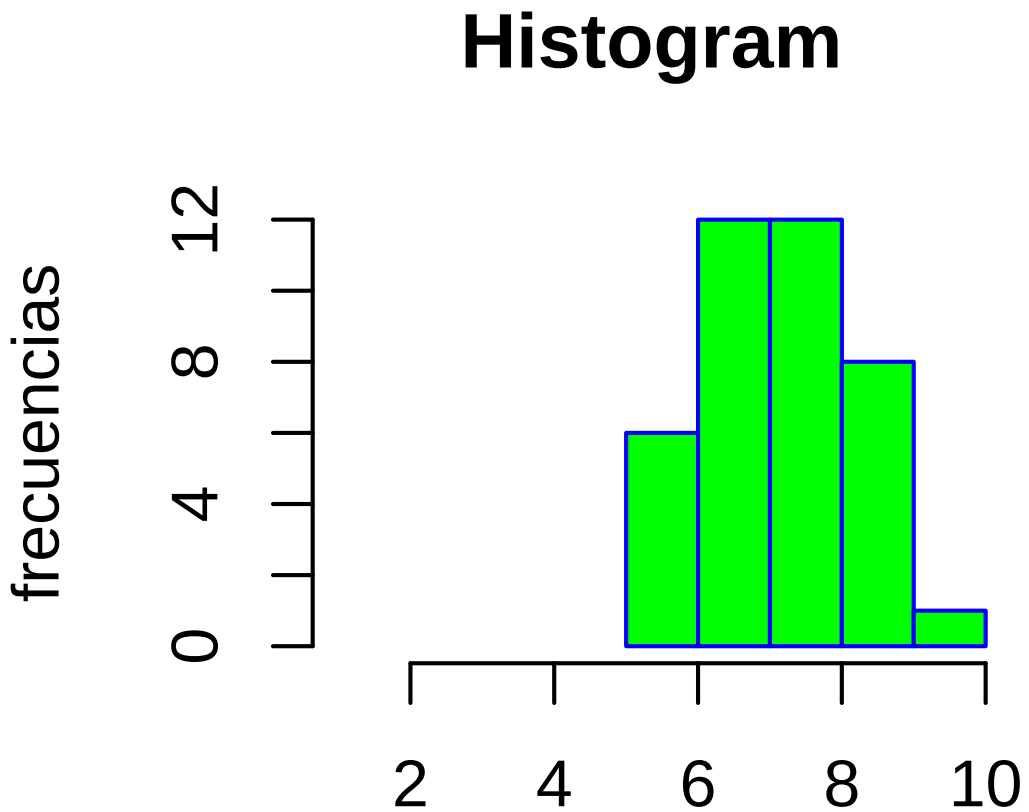
```
##
##  5  6  7  8  9 10
##  2  4 12 12  8  1
```

```
prop.table(table(calificacion$notas))#frecuencias relativas
```

```
##
##          5          6          7          8          9          10
## 0.05128205 0.10256410 0.30769231 0.30769231 0.20512821 0.02564103
```

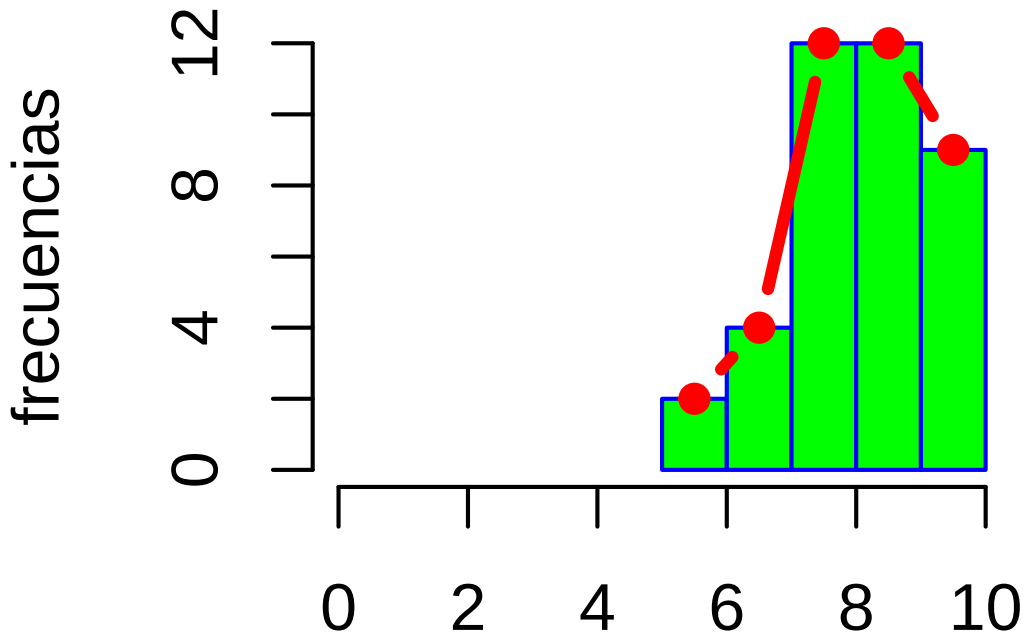
Histogramas en R (IV)

```
x=calificacion$notas
hist(x, main="Histogram ",xlab="x",
     ylab="frecuencias",border="blue",
     xlim= c(1,10),
     col="green")
```

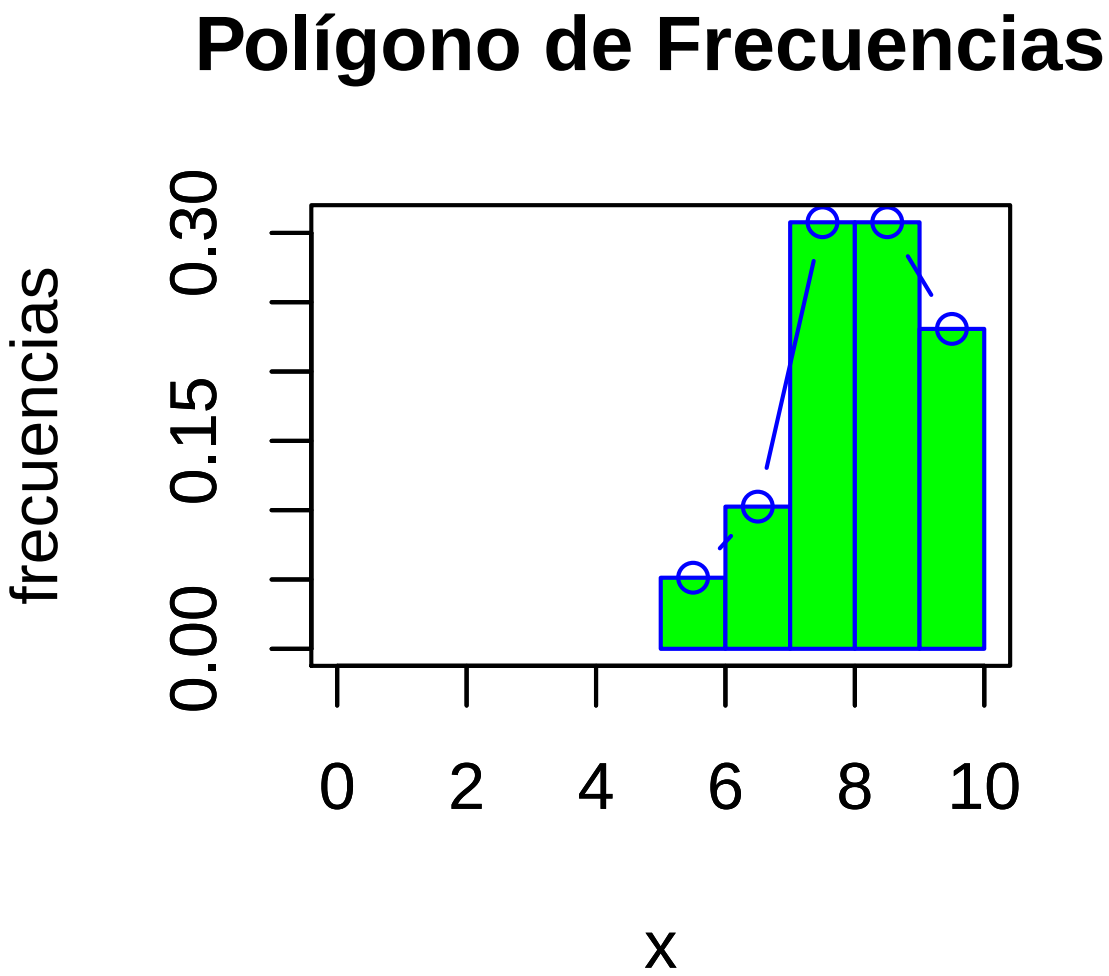


Polígono de frecuencias absolutas

```
x=calificacion$notas
h=hist(x, main=" ",xlab="x",
     ylab="frecuencias",border="blue",
     xlim= c(0,10),right = F,
     col="green")
par(new=TRUE)
lines(h$mids, h$counts, type = "b", pch = 20, col = "red", lwd = 3)
```



Polígono de frecuencias relativas



Ojiva

```
x=calificacion$notas
plot(h$breaks,cumsum(table(x)), main="Ojiva",xlab="x",type="b",
     ylab="",xlim= c(0,10),ylim=c(0,max(cumsum(table(x)))),
     col="red")
```

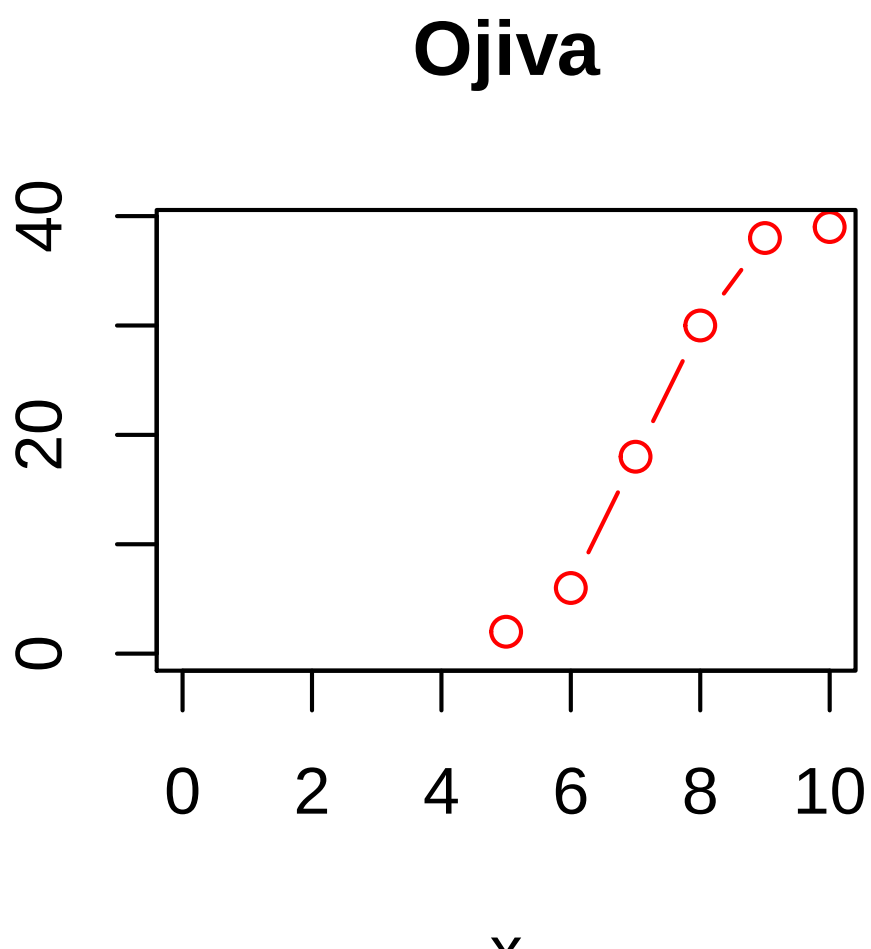


Gráfico circulares

```
slices = c(10, 12,4, 16, 8)
lbls = c("US", "UK", "Australia", "Alemania", "Francia")
pct = round(slices/sum(slices)*100)
lbls = paste(lbls, pct) # anadir porcentaje
lbls = paste(lbls,"%",sep="") # ad % to labels
pie(slices,labels = lbls, col=rainbow(length(lbls)),
    main="Gráfico circulares")
```

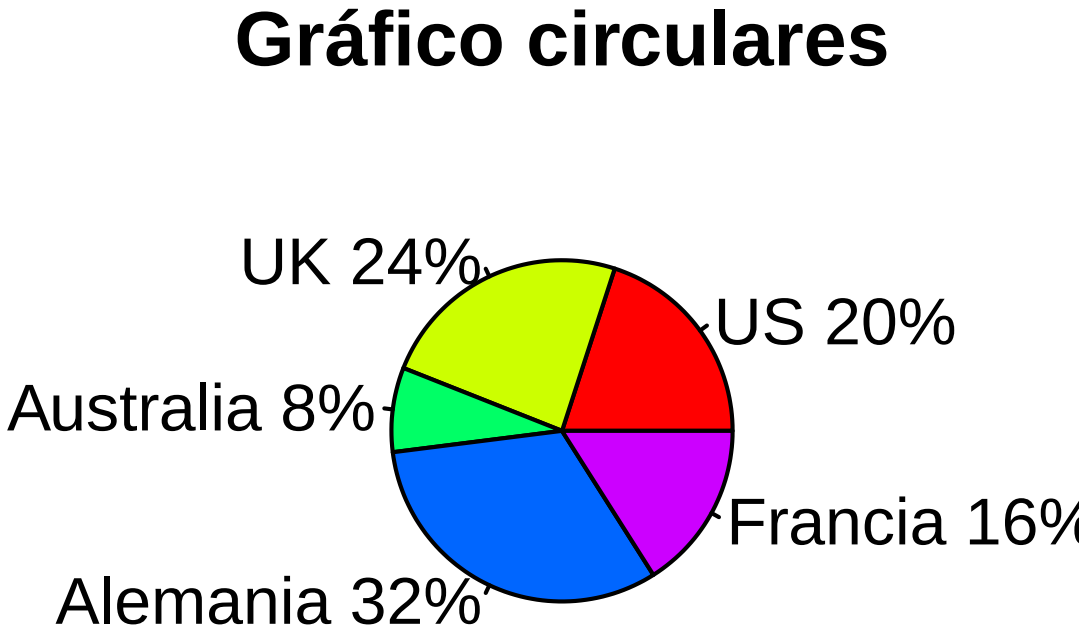
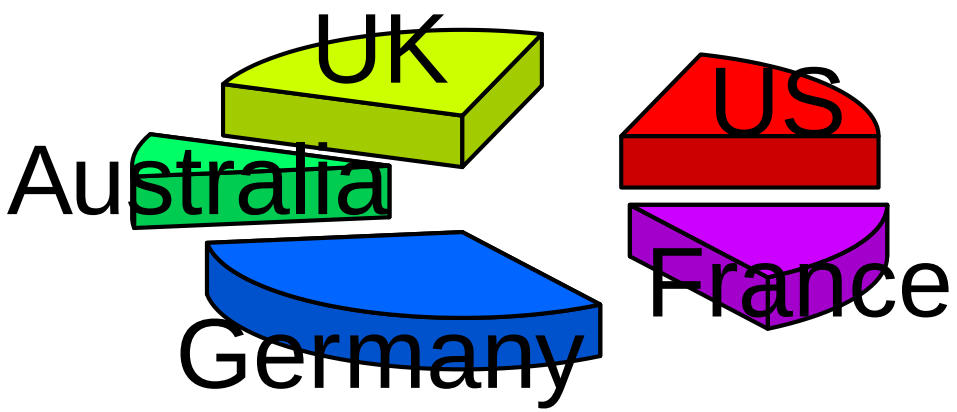


Gráfico circulares 3D

```
library(plotrix)
slices <- c(10, 12, 4, 16, 8)
lbls <- c("US", "UK", "Australia", "Germany", "France")
pie3D(slices,labels=lbls,explode=0.5,
    main="Gráfico circulares 3D")
```



Características importantes de los datos

1. **Centro:** Valor promedio o representativo que indica la localización de la mitad del conjunto de los datos.
2. **Variación:** Medida de la cantidad en que los valores de los datos varían entre sí.
3. **Distribución:** La naturaleza o forma de la distribución de los datos (como en forma de campana, uniforme o sesgada).
4. **Valores extremos:** Valores muestrales que están muy alejados de la vasta mayoría de los demás valores de la muestra.
5. **Tiempo:** Características cambiantes de los datos a través del tiempo.

Medidas de tendencia central

Medidas de Tendencia Central

Encontrar un valor numérico que ayude a sintetizar todos los datos.

Media

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

Mediana

Punto medio

$$\tilde{X} = \begin{cases} X_{(\frac{n+1}{2})}, & \text{si } n \text{ es impar} \\ \frac{1}{2}(X_{(\frac{n}{2})} + X_{(\frac{n}{2}+1)}), & \text{si } n \text{ es par} \end{cases}$$

Moda

Punto de mayor frecuencia

Media Ponderada

$$\overline{M}_W = \sum w_i x_i$$

Condicion :

$$\sum w_i = 1$$

$$0 \leq w_i \leq 1$$

EJEMPLO Calcule

- A continuación se presenta una lista de cantidades de plomo (medidas en en el aire. Calcule la mediana de esta muestra. 5.40 1.10 0.42 0.73 0.48 1.10
- Las modas de los siguientes conjuntos de datos:

1. 5.40 1.10 0.42 0.73 0.48 1.10
2. 27 27 27 55 55 55 88 88 99
3. 1 2 3 6 7 8 9 10

- Identifique una razón importante por la que, la media y la mediana no son estadísticos que puedan servir de manera precisa y efectiva como medidas de tendencia central.

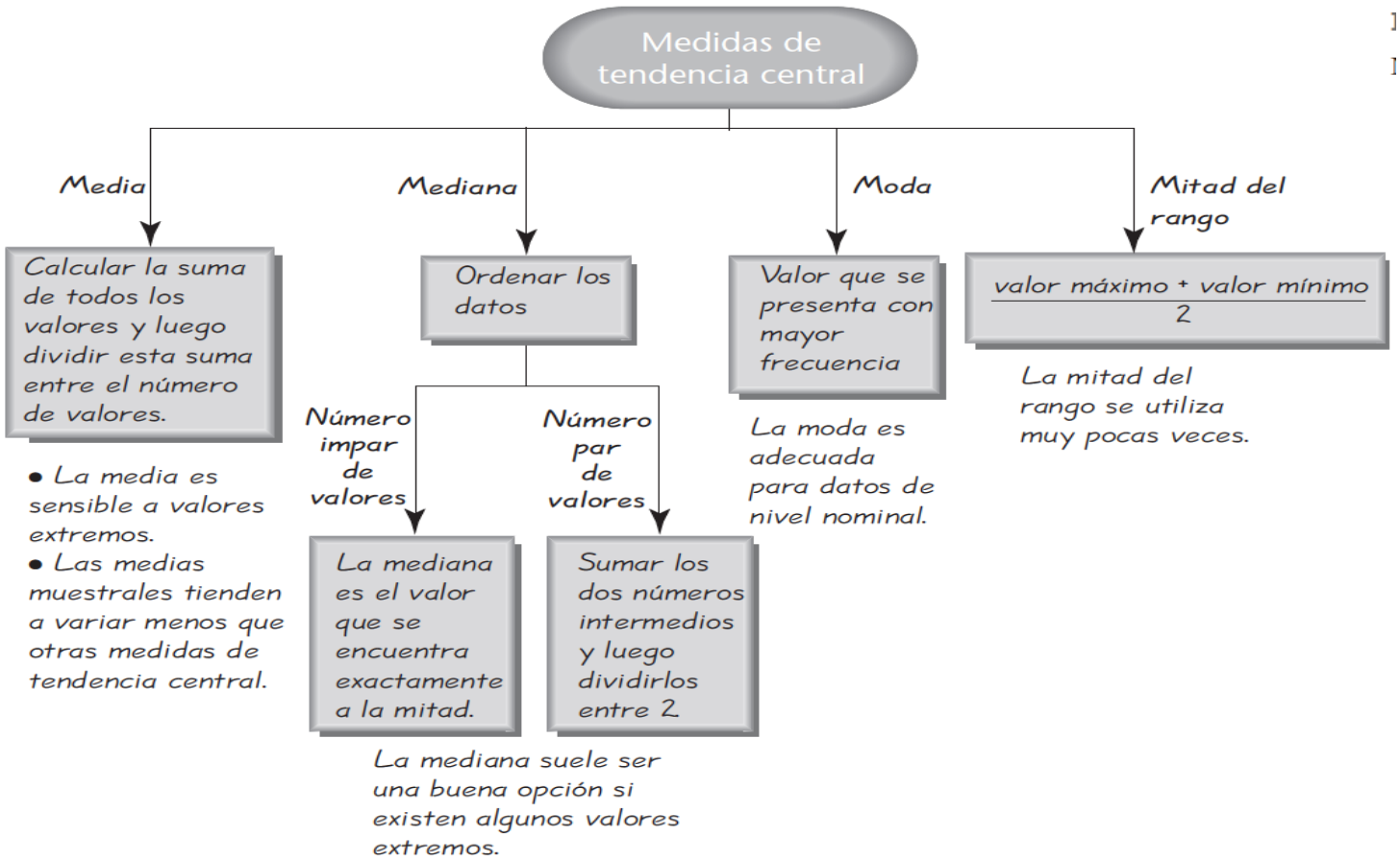
1. Códigos postales: 12601 90210 02116 76177 19102

La media de una distribución de frecuencias

Edad de las actrices	Frecuencia f	Marca de clase x	$f \cdot x$
21–30	28	25.5	714
31–40	30	35.5	1065
41–50	12	45.5	546
51–60	2	55.5	111
61–70	2	65.5	131
71–80	2	75.5	151
Totales:	$\Sigma f = 76$		$\Sigma(f \cdot x) = 2718$

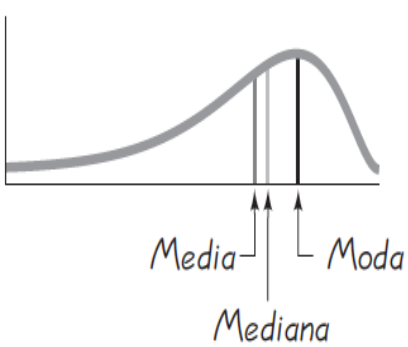
$$\bar{x} = \frac{\Sigma(f \cdot x)}{\Sigma f} = \frac{2718}{76} = 35.8$$

La mejor medida de tendencia central

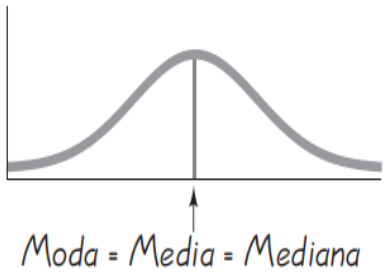


Sesgo

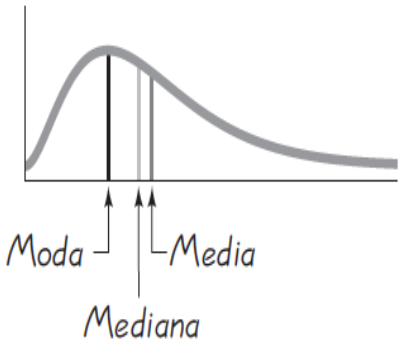
Una distribución de datos está sesgada si no es simétrica y se extiende más hacia un lado que hacia el otro. (Una distribución de datos es simétrica si la mitad izquierda de su histograma es aproximadamente una imagen en espejo de su mitad derecha).



a) Sesgada a la izquierda (sesgo negativo): la media y la mediana están a la izquierda de la moda.



b) Simétrica (sesgo cero): la media, la mediana y la moda son iguales.



c) Sesgada la derecha (sesgo positivo): la media y la mediana están a la derecha de la moda.

Ejercicios

Velocidad	Frecuencia
42–45	25
46–49	14
50–53	7
54–57	3
58–61	1

Temperatura	Frecuencia
96.5–96.8	1
96.9–97.2	8
97.3–97.6	14
97.7–98.0	22
98.1–98.4	19
98.5–98.8	32
98.9–99.2	6
99.3–99.6	4

Ejemplo

HORAS	x	f	F	x·f
55-60	57,5	5	5	287,5
60-65	62,5	18	23	1125
65-70	67,5	20	43	1350
70-75	72,5	50	93	3625
75-80	77,5	17	110	1317,5
80-85	82,5	16	126	1320
85-90	87,5	4	130	35