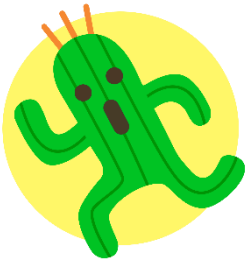
Estadística sin espinas

(1) Estadística Descriptiva



v0.3 24_02_21

Aprende sin espinas
con @carlos_cactus

Sócrates se equivocaba. El conocimiento no es lo único que crece al compartirse: La alegría también.

A la inspiración del bucle_infinito,
al Cibergrupo y al tHash_A, por su amistad,
y sobre todo, a quienes dicen "pero quiero"
cuando sienten "no puedo".

¡Un saludo sin espinas!
@carlos_cactus :D 

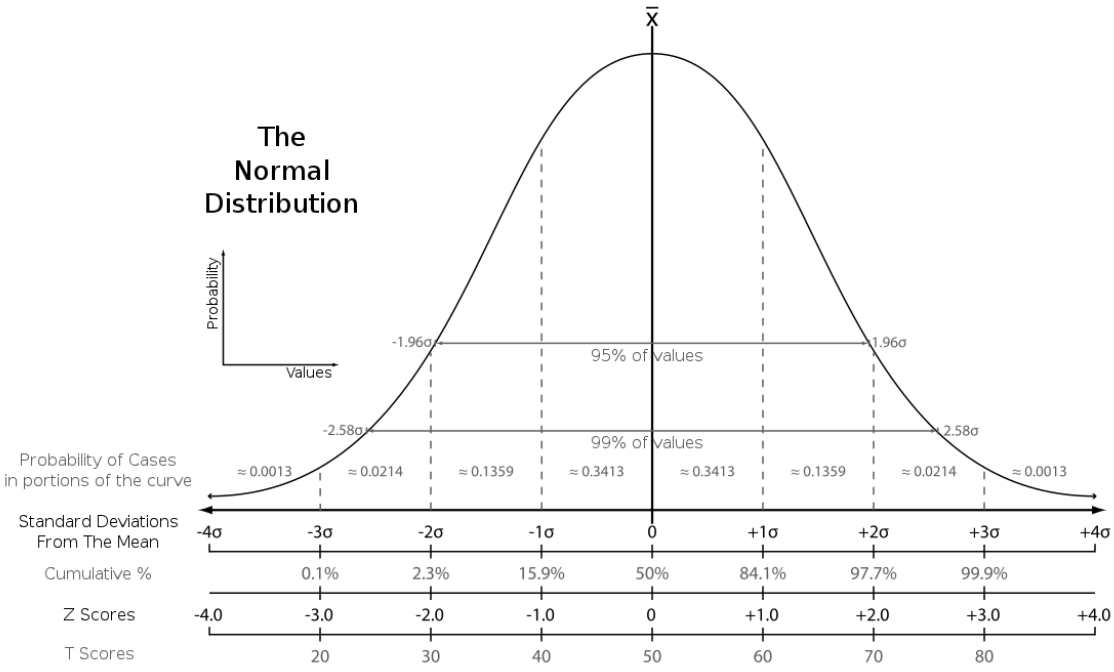
Y si quieres saber más:
¡Encuéntrame en Telegram como @carlos_cactus o habla con Espinito, el bot Sin Espinas, en @GestionSinEspinasBot.

Únete a la comunidad de Telegram Sin Espinas y no te pierdas nada!

Deja de preocuparte por aprobar y ¡Aprende sin Espinas!

ESTO ESTÁ POR TERMINAR ¡TODO ES UNA OBRA EN CONSTRUCCIÓN!
Falta:

- Introducción a R
- Depurar ejemplos de R



Índice

I. ESTADÍSTICA DESCRIPTIVA

1.1. Tipos de datos

1.1.1. POBLACIÓN

Conjunto de individuos objeto de estudio.
Se entiende por individuo cualquier tipo de entidad, animada o no.

1.1.2. MUESTRA

Una muestra es un subconjunto de una población.
Se extraen de la población mediante técnicas de muestreo.

- Una muestra se compone de OBSERVACIONES.
- Cada observación adopta un VALOR.

EJEMPLO

Se tira un dado de 6 caras 7 veces y se anotan los resultados:
1, 5, 5, 4, 3, 4, 2.
La muestra obtenida tiene 7 OBSERVACIONES.
La primera observación tiene por VALOR 1.
La segunda observación tiene por VALOR 5.
Hay 5 VALORES distintos (1, 5, 4, 3, 2) en las 7 OBSERVACIONES.

1.1.3. VARIABLES

Una variable es una característica de los individuos objetos de estudio.
Según el número de variables estudiadas a la vez, se distingue:

- Análisis univariante.
- Análisis multivariante.

El análisis de las variables pretende describir:

- Su distribución Qué valores toma y cómo los toma.
- Medidas de centro y de dispersión
- Representación gráfica de lo anterior

A partir de esa descripción, se persigue:

- Extraer conclusiones sobre la muestra.
- Extraer conclusiones sobre la población (INFERENCIA).
- Comparar poblaciones o establecer si sus diferencias son significativas
- Explorar relaciones entre variables (REGRESIÓN)

1.1.4.TIPOS DE VARIABLES

Según cómo es expresan, se distingue:

- Cualitativas (o CATEGÓRICAS)

  Representan variedades o categorías
  No se expresan numéricamente.
  Se pueden codificar las categorías mediante números.

- Cuantitativas Se expresan numéricamente

  - Cuantitativas DISCRETAS
    No adopta todos los valores del conjunto numeral.
    Exige la existencia de valores que NO PUEDE ADOPTAR entre otros 2 valores consecutivos que sí puede adoptar.

  - Cuantitativas CONTINUAS
    Pueden adoptar cualquier valor del conjunto numeral.
    Exige la existencia de valores que SÍ PUEDE ADOPTAR entre otros 2 valores consecutivos que sí puede adoptar.

- Unidimensionales

Representan una única propiedad de un individuo (sea o no una persona).

- Bidimensionales (x,y) en que x mide una propiedad e y otra.

EJEMPLOS

El número de hermanos de una persona es CUANTITATIVA DISCRETA.

El fruto seco predilecto de una persona es CUALITATIVA. Se puede codificar numéricamente mediante 1 = cacahuete, 2 = almendra…

La longitud de un cacahuete es CUANTITATIVA CONTINUA.

Un queso se fabrica en 3 formatos, distinguidos por su peso: 0,3 kg, 1 kg y 3 kg. Este tipo de variables cuantitativas DISCRETAS con POCOS valores se puede tratar como CATEGÓRICA.

1.2. Representaciones de datos

- Tabla de frecuencias (ver 1.6)

- Gráficos:

1.2.1.PICTOGRAMAS

Se asigna a cada símbolo un valor arbitrario estratégico para representar la situación.

EJEMPLO

Se desea representar en forma de pictograma la cantidad de "tickets sin espinas" que 5 personas han adquirido para el refuerzo de Estadística.

Pablo	
Pamela	
Pedro	
Portos	
Pumba	

¡Pamela sabe lo que hace! :D ¡Aprenderlo TODO y conseguir un Excelente va a ser PAN COMIDO!

1.2.2. DIAGRAMA DE TALLO Y HOJAS

Se tiene una muestra de valores:
54, 59, 35, 41, 46, 25, 47, 60, 54, 46, 49, 46, 41, 34, 22

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

Se identifica el rango:
22 a 60

Luego, se establecen 5 niveles para el TALLO del diagrama, en este caso, uno por cada DECENA.

2	Nivel 1	Decena 20
3	Nivel 2	Decena 30
4	Nivel 3	Decena 40
5	Nivel 4	Decena 50
6	Nivel 5	Decena 60

Se rellena cada nivel del tallo con los valores que son las HOJAS.

Se obtiene:

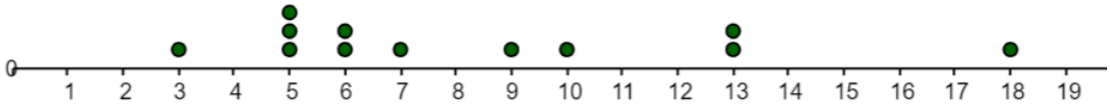
DIAGRAMA DE TALLO Y HOJAS		Se lee como
2	25	22, 25
3	45	34, 35
4	1166679	41, 41, 46, 46, 46, 47, 49
5	449	54, 54, 59
6	0	60

1.2.3. DIAGRAMA DE PUNTOS

Para una muestra de 12 observaciones cuyos valores son:

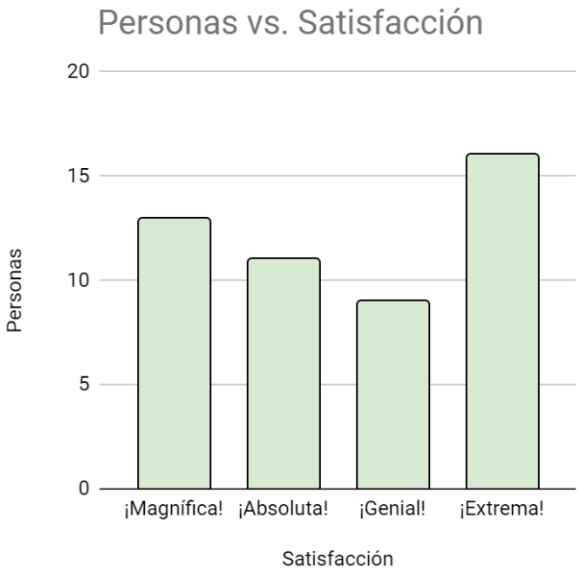
5, 5, 10, 3, 6, 6, 7, 9, 5, 13, 13, 18

Se obtiene:



Nótese el valor 18 como ATÍPICO, ANÓMALO, EXTREMO o INSÓLITO.

1.2.4. DIAGRAMAS DE BARRAS

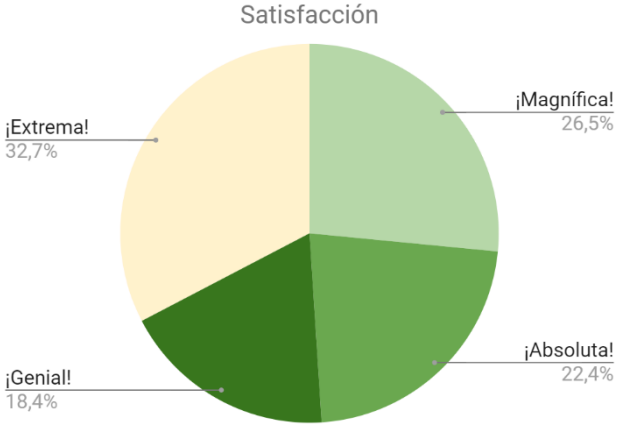


EJEMPLO

Se ha preguntado a 49 personas sobre su grado de satisfacción con el refuerzo semanal Sin Espinas y esta ha sido el resultado en forma de gráfico de barras.

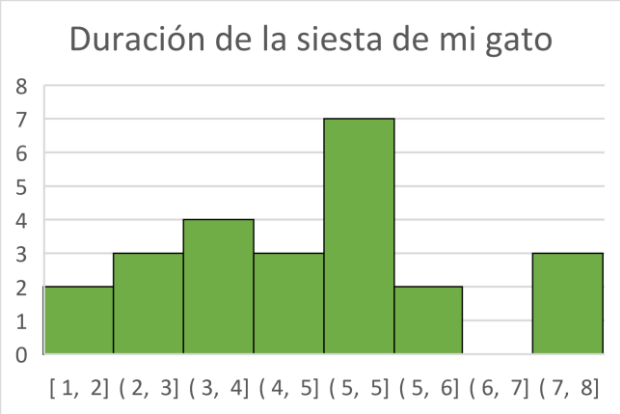
¡Con tu opinión ya serán 50!

1.2.5. DIAGRAMA DE SECTORES



Se presentan los mismos resultados del ejemplo anterior en un DIAGRAMA DE SECTORES.

1.2.6. HISTOGRAMAS DE FRECUENCIA

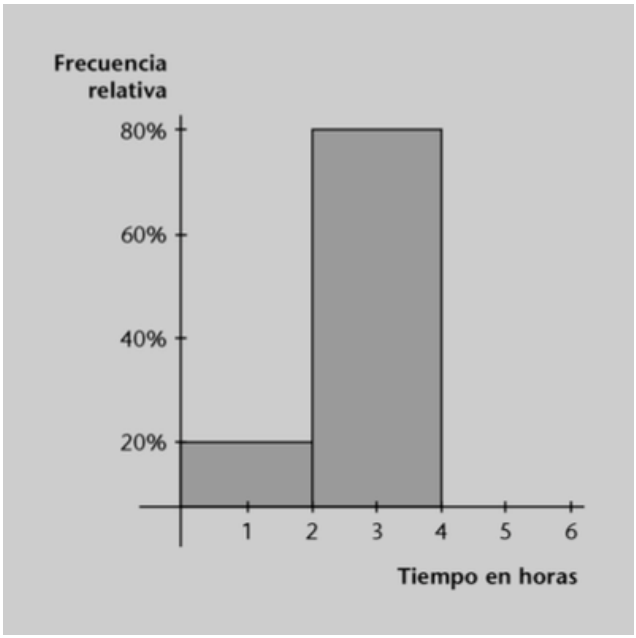


Nótese que un histograma:

Se aplica preferentemente a:

- Variables discretas con MUCHOS valores distintos (en que las otras representaciones son obsoletas).
- Variables continuas.

NO ES UN DIAGRAMA DE BARRAS, sino de ÁREAS (rectángulos).
Cada rectángulo adopta la misma base (mismo ancho, correspondiente a la LONGITUD DE CLASE) y su altura equivale a su frecuencia, de modo que la superficie de cada rectángulo respecto la superficie total representada corresponde con la FRECUENCIA RELATIVA de esa clase.



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

Exige el cálculo de la DISTRIBUCIÓN DE FRECUENCIAS. Para ello:

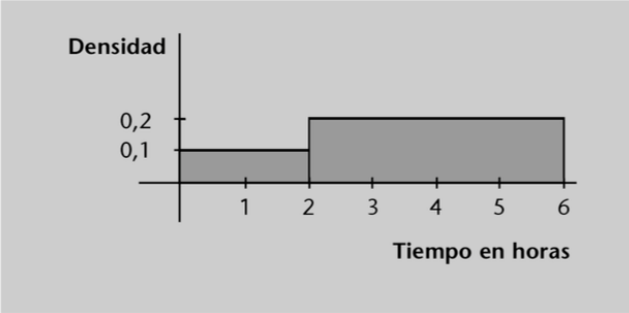
1. Agrupación de los valores en CLASES:
   - Las clases son intervalos de valores del MISMO ANCHO (misma LONGITUD).
   - Se recomienda elegir un número de clases en torno a $\sqrt{N}$.
   - Las clases con ASIMÉTRICAS en sus extremos: [cerrado , abierto)
   - Las clases recorren TODO EL RANGO de valores, sin dejar agujeros.
2. El valor promedio de la clase (suma de extremos entre 2) es el REPRESENTANTE o MARCA DE CLASE de todos los valores que pertenezcan al intervalo que define la clase:

   La clase [a,b) tiene por representante r:

$$marca\ de\ clase = \frac{a+b}{2}$$

3. De cada clase, se puede calcular y representar:
   - Frecuencia absoluta n_i
   - Frecuencia relativa r_i
   - Frecuencia absoluta acumulada N_i
   - Frecuencia relativa acumulada R_i

1.2.7. HISTOGRAMA DE DENSIDAD



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

Para calcular la DENSIDAD (altura del rectángulo) de una clase i:

$$DENSIDAD_i = altura_i = r_i / rango_i$$

Cada hora de la clase [0,2) representa el 10% de la población.
Cada hora de la clase [26) representa el 20% de la población.

1.3. Interpretación de histogramas

En la interpretación de un histograma se consideran 4 aspectos:
- Simetría
- Picos
- Colas
- Datos extremos y clases vacías

a) Simetría

Se considera que hay simetría en un histograma si se identifica un eje vertical que espeje el histograma en dos mitades simétricas.

Según los valores de la media, la moda y la mediana, se define un coeficiente de asimetría (ca), de modo que se definen 3 casos:

- Asimetría hacia la izquierda:
  Media < Mediana < Moda ca < 0
  En el histograma, se verá COLA HACIA LA IZQUIERDA

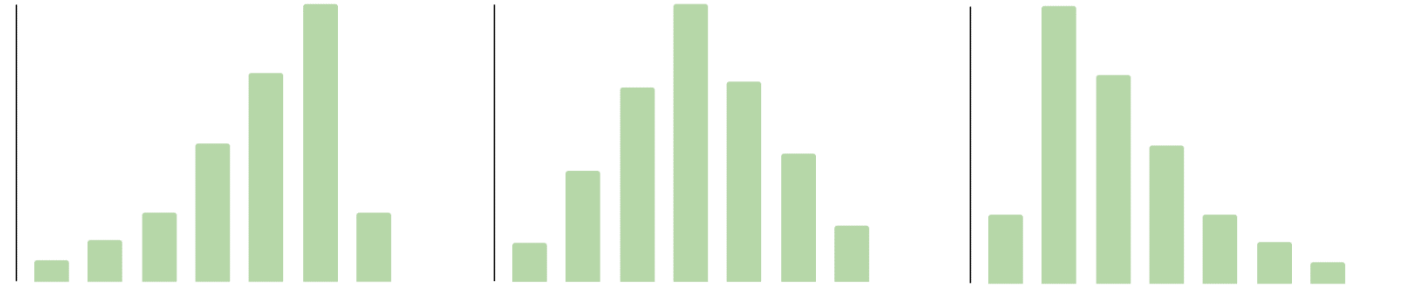
- Simetría central:
  Media = Mediana = Moda ca = 0
  En el histograma, se verá DISTRIBUCIÓN SIMÉTRICA

- Asimetría hacia la derecha:
  Moda < mediana < media ca > 0
  En el histograma, se verá COLA HACIA LA DERECHA

b) Colas

Si no hay simetría, puede haber extensiones laterales del histograma.

- Se habla de COLAS LARGAS si la prolongación es ACENTUADA.
  Cola LARGA hacia la izquierda/derecha
- Se habla de ASIMETRÍAS (hacia la izquierda o hacia la derecha) si la prolongación es MODERADA: Cola hacia la izquierda/ derecha



Distribución ASIMÉTRICA Cola hacia la IZQUIERDA

Distribución SIMÉTRICA

Distribución ASIMÉTRICA Cola hacia la DERECHA

c) Picos

Los picos son valores en los que las observaciones tienden a concentrarse (intervalos de alta frecuencia).

- Es un caso UNIMODAL si hay un único pico (una sola moda).
- Es un caso BIMODAL si hay MÁS de un único pico (más de una moda).

d) Datos EXTREMOS y clases VACÍAS, que puedan separar la población en grupos.

1.4. Estadísticos

1.4.1.CONCEPTO Y CLASIFICACIÓN

Los estadísticos son valores calculables que proporcionan información sobre los datos estudiados.

Se distinguen 3 grandes grupos:
- Medidas de centralización.
- Medidas de dispersión.
- Medidas de posicionamiento.

1.4.2. TAMAÑO DE LA MUESTRA (N)

Se define como n = número de individuos de la muestra.
Coincide con el número de OBSERVACIONES realizadas.

1.4.3. MEDIDAS DE CENTRALIZACIÓN

Informan del lugar central de los datos.

1.4.3.1. Moda (Mo)

Valor que más se repite en una muestra, es decir, de MÁXIMA FRECUENCIA ABSOLUTA.
Cuando más de un valor comparte la máxima frecuencia absoluta con otros, se habla de distribuciones MULTIMODALES.

1.4.3.2. Mediana (Me)

Devuelve el valor central de la muestra una vez ORDENADOS todos sus valores de menor a mayor.
Es aquel valor que deja tantas observaciones a su izquierda como a su derecha. Coincide con el segundo cuartil Q_2.

Se calcula a partir de una serie de valores ORDENADOS:

$$\{x_1, x_2, x_3, x_4, x_5, x_6, x_7\}$$

Entonces, se puede escribir:

Inferiores a Me (Q_2)	x_1	x_2	x_3				
Mediana (Q_2)				x_4			
Superiores a Me (Q_2)					x_5	x_6	x_7

Se distinguen 2 casos, aunque el concepto es el mismo. Para calcular Me:

1. ORDENAR las observaciones de forma ASCENDENTE.
2. Localizar la POSICIÓN que deja EXACTAMENTE tantas observaciones a su izquierda como a su derecha.

3. Calcular/Observar qué VALOR ocupa esa posición:
   - Si el tamaño de la muestra es IMPAR, Me es la observación central.
   - Si el tamaño de la muestra es PAR, Me es la MEDIA de las 2 observaciones en torno a las cuales estaría la central.

   CASO A)
   Si se tiene una colección IMPAR de observaciones, Me corresponde al valor central de la serie ORDENADA:

   Se tienen 5 observaciones:
   5,7,4,6,2

   Se ordenan:
   2,4,5,6,7

   Se OBSERVA un valor central:

$$\underbrace{2,4}_{2\ obs}\ ,5,\ \underbrace{6,7}_{2\ obs}$$

   La mediana es 5 (deja 2 observaciones a cada lado).

   CASO B)
   Si se tiene una colección PAR de observaciones, Me corresponde a la media aritmética de los 2 valores centrales en la serie ORDENADA:

   Se tienen 6 observaciones:
   5,7,4,6,2,9

   Se ordenan:
   2,4,5,6,7,9

   Se CALCULA un valor central:

$$\underbrace{2,4,5,}_{3\ obs}\ 5.5,\ \underbrace{6,7,9}_{3\ obs}$$

   La mediana es 5.5 (media de los valores 5 y 6, de modo que se dejan 3 observaciones a cada lado).

Rigurosamente, coincide con aquel valor que deja a su izquierda el 50% de las observaciones y a su derecha el otro 50%. Es decir, la mediana Me cumple necesariamente 2 condiciones:

- Al menos el 50% de las observaciones deben ser menores o iguales a Me.
- Al menos el 50% de las observaciones deben ser mayores o iguales a Me.

Esto implica que Me corresponde el segundo cuartil Q2.

1.4.3.3. Media aritmética ($\bar{x}$)

Devuelve el valor promedio de una distribución de datos.
Se define como:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_N}{N}$$

O alternativamente, es más práctico:

- Si se dispone de una tabla de frecuencias ABSOLUTAS:

$$\bar{x} = \frac{\sum_{i=1}^{N} n_i \cdot x_i}{N} = \frac{\sum frecuencia_i \cdot valor_i}{tamaño}$$

- Si se dispone de una tabla de frecuencias RELATIVAS:

$$\bar{x} = \frac{\sum_{i=1}^{N} n_i \cdot x_i}{N} = \sum_{i=1}^{N} \frac{n_i}{N} \cdot x_i = \sum_{i=1}^{N} r_i \cdot x_i$$

Donde:

x_i = VALOR que adopta la variable
n_i = frecuencia ABSOLUTA de cada VALOR (apariciones en la muestra)
r_i = frecuencia RELATIVA de cada VALOR (proporción de aparición)
N = tamaño de la muestra

a) Propiedades de la media

Nótese que la media cumple 4 propiedades:

Antes de enunciarlas, se recuerda que la suma de todas las observaciones resulta en el tamaño de la muestra:

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

1. El producto de la media por el tamaño de la muestra equivale a la suma de los valores de todas las observaciones:

$$N\bar{x} = x_1 + x_2 + \cdots + x_N = \sum_{i=1}^{N} x_i$$

2. La suma de las desviaciones de la media es nula:

El término $(x_i - \bar{x})$ se considera como la DESVIACIÓN RESPECTO LA MEDIA que presenta una observación i.

Entonces, se cumple, para una muestra de tamaño N:

$$(x_1 - \bar{x}) + (x_2 - \bar{x}) + \cdots + (x_N - \bar{x}) = 0$$

EJEMPLO

Se tiene la muestra:
3,4,3,5,3,6

Se ordena:
3,3,3,4,5,6

Se observa: $N = 6$

La suma de las observaciones es:

$$\sum_{i=1}^{N} x_i = 3 + 3 + 3 + 4 + 5 + 6 = 24$$

La media es:

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i \cdot n_i}{N} = \frac{3 \cdot 3 + 1 \cdot 4 + 1 \cdot 5 + 1 \cdot 6}{6} = \frac{24}{6} = 4$$

Se verifica la nulidad de la suma de las desviaciones:

$$(3 - 4) + (3 - 4) + (3 - 4) + (4 - 4) + (5 - 4) + (6 - 4) = 0$$

Lo anterior se demuestra del siguiente modo. Se tiene:

$$(3 - 4) + (3 - 4) + (3 - 4) + (4 - 4) + (5 - 4) + (6 - 4)$$

Es decir:

$$3 - 4 + 3 - 4 + 3 - 4 + 4 - 4 + 5 - 4 + 6 - 4$$

O lo que es lo mismo:

$$\underbrace{3 + 3 + 3 + 4 + 5 + 6}_{\sum_{i=1}^{N} x_i} - \underbrace{6 \cdot 4}_{N\bar{x}}$$

Pero se observa la propiedad (2):

$$N\bar{x} = \sum_{i=1}^{N} x_i$$

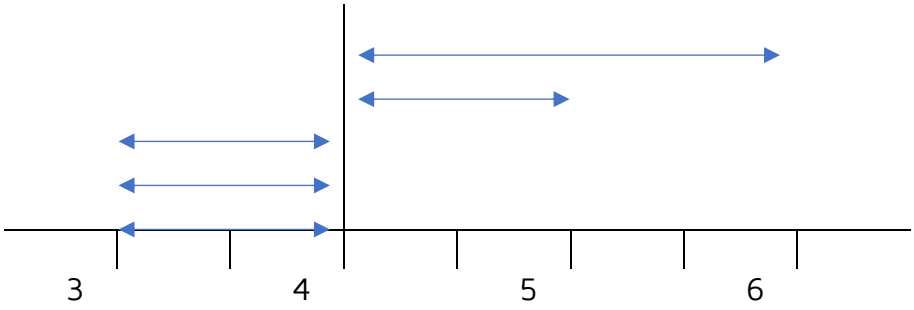
En este caso:

$$\sum_{i=1}^{N} x_i = 24 = \underbrace{6}_{N} \cdot \underbrace{4}_{\bar{x}} = N\bar{x}$$

De modo que:

$$\underbrace{3 + 3 + 3 + 4 + 5 + 6}_{N\bar{x}} - \underbrace{6 \cdot 4}_{N\bar{x}} = 0$$

Es decir, las DESVIACIONES A LADO Y LADO DE LA MEDIA SE EQUILIBRAN:



La suma de las longitudes de las desviaciones por DEBAJO de la media EQUIVALE a la suma de las longitudes de las desviaciones por ENCIMA.

3. La media es POCO ROBUSTA

Se ve afectada INTENSAMENTE por los valores EXTREMOS, es decir, tiende a desplazarse hacia los valores alejados.

EJEMPLO

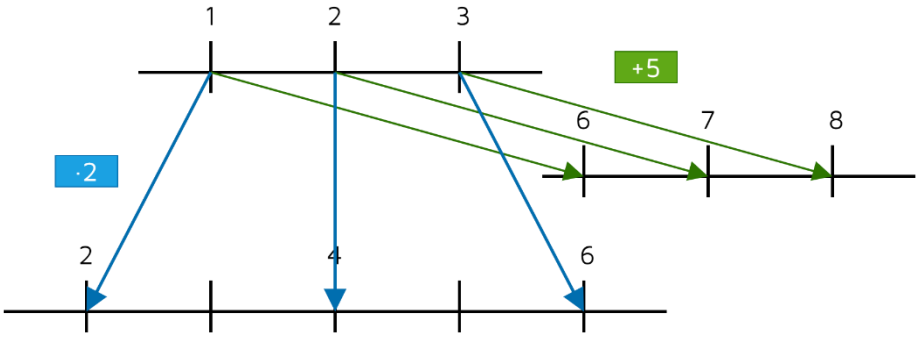
Tómese la muestra anterior:
3,3,3,4,5,6
Cuya media es 4.
Ahora, la modificación de 1 solo valor afecta fuertemente la nueva media.
Si se toma la muestra:
3,3,3,4,5,12
La nueva media es 5.

La intensidad de este efecto es inversamente proporcional al tamaño de la muestra.

1.4.3.4. Efecto de las TRANSFORMACIONES LINEALES sobre la media:

Cambio de origen: $\bar{y} = \bar{x} + c$

Cambio de escala: $\bar{y} = \bar{x} \cdot c$



Si a todas las observaciones se suma un mismo número c, la media resultante $\bar{y}$ es la media original $\bar{x}$ más ese número c. Esto es un CAMBIO DE ORIGEN.

Si todas las observaciones se multiplican por un mismo número c, la media resultante $\bar{y}$ es la media original $\bar{x}$ multiplicada por ese número c. Esto es un CAMBIO DE ESCALA.

EJEMPLO
Se tiene:

3,3,3,4,5,6 con media 4.

Es decir:
$x_1 = 3, x_2 = 3, x_3 = 3, x_4 = 4, x_5 = 5, x_6 = 6, \bar{x} = 4$

Se toma el escalar c = 2 y se suma a cada observación:

$$y_i = x_i + C$$

Es decir:

$$y_1 = 3 + 2, y_2 = 3 + 2, y_3 = 3 + 2, y_4 = 4 + 2, y_5 = 5 + 2, y_6 = 6 + 2$$

Entonces:

$$\bar{y} = \frac{5 + 5 + 5 + 6 + 7 + 8}{6} = \frac{36}{6} = 6 = 4 + 2 = \bar{x} + c$$

Es decir:

$$\bar{y} = \frac{(x_1 + c + x_2 + \cdots + x_N + c)}{N} = \frac{Nc + \sum_{i=1}^{N} x_i}{N} = \frac{Nc}{N} + \frac{\sum_{i=1}^{N} x_i}{N} = c + \bar{x}$$

Y análogamente se demuestra para el cambio de escala:

$$\bar{y} = \frac{(c \cdot x_1 + c \cdot x_2 + \cdots + c \cdot x_N)}{N} = \frac{c \cdot \sum_{i=1}^{N} x_i}{N} = c \cdot \bar{x}$$

1.4.3.1. Comparación entre mediana Me y media $\bar{x}$

Nótese:
- La media tiende a desplazarse hacia los extremos (NO ES ROBUSTA).
- La mediana no se ve tan afectada por los extremos, pero NO LOS REPRESENTA.
- En una distribución simétrica, ambas, Me y $\bar{x}$ coinciden.

EJEMPLO

En la muestra:
1,2,3,4,5,6,7

Se tiene: $\bar{x} = Me = 4$

Pero si uno de los valores extremos se desplaza:
1,2,3,4,5,6,17

La media se desplaza:

$$\bar{x} = 5.42$$

La mediana permanece:

$$Me = 4$$

1.4.3.2. MEDIDAS DE CENTRO Y DATOS TABULADOS

Se describen 3 situaciones destacadas, según cómo se presentan los datos:

a) Tabla de frecuencias ABSOLUTAS

x_i	n_i
0	10
5	15
9	6
Total	31

$$Mo = 5$$

$$Me = 5 \leftrightarrow \frac{N}{2} = \frac{31}{2} = 15{,}5$$

En la serie ordenada de 31 observaciones, la posición central es la posición 15,5.
La observación que ocupa la posición 15 adopta como valor 5 (es el quinto 5 después de 10 observaciones con valor 0).
La observación que ocupa la posición 16 adopta como valor 5 (es el sexto 5 después de 10 observaciones con valor 0).
La mediana es el promedio entre 5 y 5, es decir, 5.

Para la media, como se dispone de las frecuencias absolutas se recurre a:

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

$$\bar{x} = \frac{\sum_{i=1}^{N} n_i \cdot x_i}{N} = \frac{\overbrace{10}^{veces} \cdot \overbrace{0}^{valor} + \overbrace{15}^{veces} \cdot \overbrace{5}^{valor} + \overbrace{6}^{veces} \cdot \overbrace{9}^{valor}}{31} = 4{,}1612$$

b) Tabla de frecuencias RELATIVAS

Se cumple:

$$\bar{x} = \frac{\sum_{i=1}^{N} n_i \cdot x_i}{N} = \sum_{i=1}^{N} \frac{n_i}{N} \cdot x_i = \sum_{i=1}^{N} r_i \cdot x_i$$

$$\bar{x} = \frac{10}{31} \cdot 0 + \frac{15}{31} \cdot 5 + \frac{6}{31} \cdot 9 = 4{,}1612$$

x_i	r_i	r_i (%)
0	$\frac{10}{31}$	32,258%
5	$\frac{15}{31}$	48,387%
9	$\frac{6}{31}$	19,354%

c) DATOS AGRUPADOS

En un histograma, el asociado a cada frecuencia o densidad (altura de columna) es el REPRESENTANTE DE CLASE, es decir, el valor promedio de esa clase. Ese valor REPRESENTANTE sustituye al valor de la observación:

Clase	representante m_i	r_i	r_i (%)
[0,1)	0,5	$\frac{10}{31}$	32,258%
[1,6)	3,5	$\frac{15}{31}$	48,387%
[5,15)	10	$\frac{6}{31}$	19,354%

$$\bar{x} = \sum_{i=1}^{N} r_i \cdot m_i$$

En este caso:

$$\bar{x} = \frac{10}{31} \cdot 0{,}5 + \frac{15}{31} \cdot 3{,}5 + \frac{6}{31} \cdot 10 = 3{,}306$$

1.4.4. MEDIDAS DE DISPERSIÓN

Informan de cómo se distribuyen los datos en torno a las posiciones de sus valores centrales.

EJEMPLO
La media de 2 muestras es 5. Las muestras son:
muestra A = {1,5,9}
muestra B = {5,5,5}
Las medidas de centralización no informan de cómo varían los datos.
Hacen falta otro tipo de indicadores.

1.4.4.1. Máximo (max) mínimo (min) y rango

Para variables numéricas:
a) El mayor valor de entre los valores que adopta la muestra es el MÁXIMO.
b) El menor valor de entre los valores que adopta la muestra es el MÍNIMO.
c) El rango es la diferencia entre MAX y MIN.
Para una muestra A:

$$Rango_A = MAX(A) - MIN(A)$$

1.4.4.2. Cuartiles (Q)

HAY MÚLTIPLES DEFINICIONES NO EQUIVALENTES PARA CALCULAR CUARTILES

Los cuartiles representan COMPARTIMENTOS que distribuyen las observaciones de la muestra en 4 grupos para representar qué volumen ocupa cada uno respecto los demás. Se definen 3 cuartiles:

PRIMER cuartil (Q1), que cumple:
- Es el VALOR superior o igual al 25% de OBSERVACIONES.
- Es el VALOR inferior o igual al 75% de OBSERVACIONES.

SEGUNDO cuartil (Q2), que cumple:
- Es el VALOR superior o igual al 50% de OBSERVACIONES.
- Es el VALOR inferior o igual al 50% de OBSERVACIONES.
- Coincide con la MEDIANA Me

TERCER cuartil (Q3), que cumple:
- Es el VALOR superior o igual al 75% de OBSERVACIONES.
- Es el VALOR inferior o igual al 25% de OBSERVACIONES.

También se puede visualizar del siguiente modo:
Ante una serie de observaciones ordenadas de forma ascendente:

$$\{x_1, x_2, x_3, x_4, x_5, x_6, x_7\}$$

Se puede escribir:

Inferiores a Q_2	x_1	x_2	x_3				
Mediana				x_4			
Superiores a Q_2					x_5	x_6	x_7

Y se calcula la mediana de cada subconjunto: para $\{x_1, x_2, x_3\}$ inferior a Me:

Inferiores a Q_1	x_1		
Q_1		x_2	
Entre Q_1 y Me			x_3

Y para $\{x_5, x_6, x_7\}$ superiores a Me:

Entre Q_2 y Q_3	x_5		
Q_3		x_6	
Superiores a Q_3			x_7

Nótese que para encontrar el valor m EXACTAMENTE entre 2 valores a y b, se puede recurrir a su promedio:

$$m = \frac{a+b}{2}$$

De modo que m EQUIDISTA de a y b, independientemente de si hay un número par o impar de observaciones.

O bien, a la expresión TOTALMENTE EQUIVALENTE:

$$m = a + 0.5 \cdot (b - a)$$

EJEMPLO

Se tienen las siguientes 11 observaciones:
3,5,6,1,3,5,2,8,7,0,1

Se desea calcular los cuartiles Q1, Q2 y Q3.

1. Se ordenan las 11 observaciones de menor a mayor:
   0,1,1,2,3,3,5,5,6,7,8

2. Se calcula Q2 (Me)

   El VALOR CENTRAL estará EXACTAMENTE en la 6ª observación, APARECE A LA SERIE ORDENADA, dejando 5 observaciones a su izquierda y 5 a su derecha:

$$\underbrace{0,1,1,2,3,}_{\substack{5\ observaciones\\ A\ LA\ IZQUIERDA}}\ \underbrace{3}_{Q2=Me}\ ,\ \underbrace{5,5,6,7,8}_{\substack{5\ observaciones\\ A\ LA\ DERECHA}}$$

   Por tanto, Me = Q2 = 3.

3. Se calcula Q1 y Q3 como la mediana de cada uno de los grupos que Q2 deja a su izquierda y derecha respectivamente:

$$\underbrace{0,1,1,2,3}_{\substack{5\ observaciones\\ A\ LA\ IZQUIERDA}} \rightarrow \underbrace{0,1,}_{\substack{2\ obs\\ IZQ}}\ \underbrace{1}_{Q1}\ ,\ \underbrace{2,3,}_{\substack{2\ obs\\ DER}} \rightarrow Q1 = 1$$

$$\underbrace{5,5,6,7,8}_{\substack{5\ observaciones\\ A\ LA\ DERECHA}} \rightarrow \underbrace{5,5,}_{\substack{2\ obs\\ IZQ}}\ \underbrace{6}_{Q3}\ ,\ \underbrace{7,8}_{\substack{2\ obs\\ DER}} \rightarrow Q3 = 6$$

EJEMPLO

Se tiene la muestra ORDENADA con 7 observaciones:
{12,13,14,15,19,19,170}

De lo cual:

$$\underbrace{12,13,14,}_{\substack{3\ obs\\ IZQ}}\underbrace{15}_{Q2},\underbrace{19,19,170}_{\substack{3\ obs\\ DER}} \rightarrow Q2 = Me = 15$$

Entonces, para Q1, ESTRICTAMENTE a la izquierda de Q2:

$$\underbrace{12,\overbrace{13}^{Q1},14,}_{\substack{3\ obs\\ IZQ}} \rightarrow Q1 = 13$$

Y, para Q3, ESTRICTAMENTE a la DERECHA de Q2:

$$\underbrace{19,\overbrace{19}^{Q3},170}_{\substack{3\ obs\\ DER}} \rightarrow Q3 = 19$$

EJEMPLO

Se tienen 12 observaciones de las cuales se desea calcular Q1, Q2 y Q3: 3,5,6,1,3,5,2,8,7,0,1,8

1. Se ordenan las 12 observaciones de menor a mayor:

   0,1,1,2,3,3,5,5,6,7,8,8

2. Se calcula Q2 (Me)

   El valor central estará EXACTAMENTE entre la 6ª y la 7ª observación, NO APARECE EN LA SERIE ordenada, será la MEDIA 6ª y la 7ª observación y dejará 6 observaciones a su izquierda y 6 a su derecha:

$$\underbrace{0,1,1,2,3,3}_{\substack{6\ observaciones\\ A\ LA\ IZQUIERDA}}\ \underbrace{?}_{Q2=Me}\ ,\ \underbrace{5,5,6,7,8,8}_{\substack{6\ observaciones\\ A\ LA\ DERECHA}}$$

   Por tanto, el valor que equidista de 3 y 5 es Q2:

$$Me = Q2 = \frac{3+5}{2} = 4$$

3. Se calcula Q1 y Q3 como la mediana de cada uno de los grupos que Q2 deja a su izquierda y derecha respectivamente:

$$\underbrace{0,1,1,2,3,3}_{\substack{6\ observaciones\\ A\ LA\ IZQUIERDA}} \rightarrow \underbrace{0,1,1,}_{\substack{3\ obs\\ IZQ}}\ \underbrace{?}_{Q1}\ ,\underbrace{2,3,3}_{\substack{3\ obs\\ DER}} \rightarrow Q1 = \frac{1+2}{2} = 1.5$$

$$\underbrace{5,5,6,7,8,8}_{\substack{6\ observaciones\\ A\ LA\ DERECHA}} \rightarrow \underbrace{5,5,6,}_{\substack{3\ obs\\ IZQ}}\ \underbrace{?}_{Q3}\ ,\underbrace{7,8,8}_{\substack{3\ obs\\ DER}} \rightarrow Q3 = \frac{6+7}{2} = 6.5$$

EJEMPLO

De la muestra ORDENADA con 6 observaciones se desea Q1, Q2 y Q3: {12,13,14,15,16,17}
De lo cual:

$$\underbrace{12,13,14,}_{\substack{3\ obs\\ IZQ}}\ \underbrace{?}_{Q2}\ ,\underbrace{15,16,17}_{\substack{3\ obs\\ DER}} \rightarrow Q2 = Me = \frac{14+15}{2} = 14.5$$

Para Q1, ESTRICTAMENTE a la izquierda de Q2: $\underbrace{12,\overbrace{13}^{Q1},14,}_{\substack{3\ obs\\ IZQ}} \rightarrow Q1 = 13$

Para Q3, ESTRICTAMENTE a la DERECHA de Q2: $\underbrace{15,\overbrace{16}^{Q3},17}_{\substack{3\ obs\\ DER}} \rightarrow Q3 = 16$

EJEMPLO

Se tiene la muestra ORDENADA con 9 observaciones:

8,12,13,14,15,19,109,170,180

De lo cual:

$$\underbrace{8,12,13,14,}_{\substack{4\ obs\\ IZQ}}\ \underbrace{15}_{Q2}\ ,\ \underbrace{19,109,170,180}_{\substack{4\ obs\\ DER}} \rightarrow Q2 = Me = 15$$

Entonces, para Q1, ESTRICTAMENTE a la izquierda de Q2:

$$\underbrace{8,12,13,14,}_{\substack{4\ obs\\ IZQ}} \rightarrow \underbrace{8,12,}_{\substack{2\ obs\\ IZQ}}\ \underbrace{?}_{Q1}\ ,\ \underbrace{13,14}_{\substack{2\ obs\\ DER}} \rightarrow Q1 = \frac{12+13}{2} = 12.5$$

Y, para Q3, ESTRICTAMENTE a la DERECHA de Q2:

$$\underbrace{19,109,170,180}_{\substack{4\ obs\\ DER}} \rightarrow \underbrace{19,109,}_{\substack{2\ obs\\ IZQ}}\ \underbrace{?}_{Q3}\ ,\ \underbrace{170,190}_{\substack{2\ obs\\ DER}} \rightarrow Q3 = \frac{109+170}{2} = 139.5$$

Además de los cuartiles, se pueden establecer otras agrupaciones:

- Deciles D_1, D_2, … D_9
- Percentiles P_1, P_2, … P_{99}

1.4.4.3. Rango intercuartílico

Se define el rango intercuartílico como:

$$Rango\ intercuartílico = Q3 - Q1$$

Es una medida de la dispersión de la muestra, ya que entre los valores Q1 y Q3 se encuentran el 50% de las observaciones. Por tanto:

- Un rango intercuartílico AMPLIO es signo de ALTA dispersión.
- Un rango intercuartílico ESTRECHO es signo de BAJA dispersión.

1.4.4.4. Medidas de forma

Parámetros que informan sobre la forma de distribución de los datos.

a) Números resumen

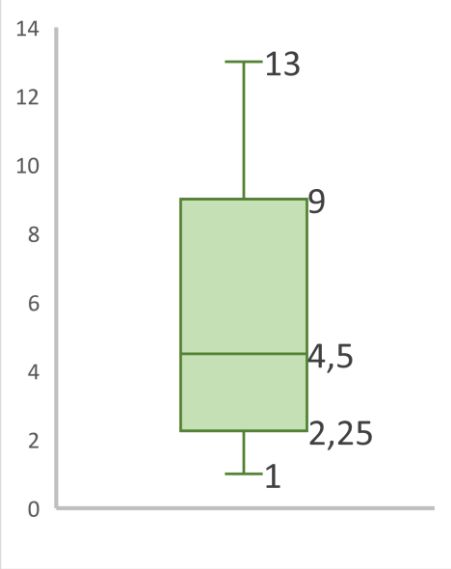
Se definen los 5 números resumen de una muestra como:

- Cuartiles Q1, Q2 (Me) y Q3.
- Mínimo y máximo.

b) BOX-PLOT (Diagrama de caja)

Los 5 números resumen suelen representarse mediante un diagrama de caja o *box-plot*:

El diagrama está dividido en 4 secciones:



- "bigote" inferior De mínimo a Q1
- "caja" inferior De Q1 a Q2
- "caja" superior De Q2 a Q3
- "bigote" superior De Q3 a máximo

Los datos del ejemplo son: 1,1,2,2,2,3,3,4,4,4,5,5,7,7,9,9,9,11,12,13

De modo que:
Q1 Separa bigote inferior de caja inferior
Q2 Separa caja inferior de caja superior
Q3 Separa caja superior de bigote superior

- Los puntos aislados fuera de las cajas corresponden con valores atípicos.

- Cada sección aglutina APROXIMADAMENTE el mismo número de observaciones.

- La LONGITUD de cada sección es proporcional al RANGO de las observaciones que aglutina, y NO REPRESENTA LA CANTIDAD DE OBSERVACIONES, SINO CUÁN DISPERSAS SON, aproximadamente.

- Se usan para COMPARAR la misma variable de muestras o poblaciones distintas.

- Su principal limitación es el efecto que tienen en los bigotes los valores extremos: un solo valor tira de ellos irremediablemente.

1.4.4.5. varianza poblacional σ^2

→ **NO CONFUNDIR con la CUASIVARIANZA MUESTRAL, denotada por $\hat{s}^2$.**

a) Concepto de varianza σ^2 (sigma cuadrada)

Se podría describir la distribución de los datos en torno a la media calculando la DESVIACIÓN PROMEDIA respecto la media.
Pero como la suma de las desviaciones es nula, ese promedio daría 0.
Para sortear esa anulación, se recurre al CUADRADO DE LAS DESVIACIONES.

El promedio de los cuadrados de las desviaciones respecto la media es la VARIANZA POBLACIONAL σ^2.

b) Definición de VARIANZA POBLACIONAL σ^2

- La varianza POBLACIONAL σ^2 informa sobre la dispersión de los datos en torno a la media aritmética.
- Es la suma del cuadrado de las desviaciones de cada observación respecto la media dividida entre el número N de observaciones, es decir, es la desviación promedio respecto la media.
- Es el cuadrado de la desviación típica σ.
- Siempre es positiva

Se define como:

$$\sigma^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_N - \bar{x})^2}{N} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

No se aconseja usar esta expresión, ya que es MUY TEDIOSA de introducir en la calculadora de mano. En su lugar, **ES ESTRATÉGICO APLICAR**:

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

Se denota σ_x^2 la VARIANZA POBLACIONAL de la VARIABLE X.

EJEMPLO

Se desea conocer la VARIANZA POBLACIONAL σ^2 de los siguientes datos: 3,2,7,3,7,7,6

1. Se observa el tamaño de la muestra N = 7.

2. Se calcula la media $\bar{x} = \frac{2+2\cdot3+6+3\cdot7}{7} = 5$

3. Se calcula $x_i^2 \cdot n_i$, es decir, el CUADRADO DE CADA VALOR por su frecuencia ABSOLUTA.

x_i	n_i	$x_i^2 \cdot n_i$
2	1	2^2
3	2	$3^2 \cdot 2$
6	1	6^2
7	3	$7^2 \cdot 3$

4. Se aplica:

$$\sigma^2 = \frac{\sum_{i=1}^{N} x_i^2 \cdot n_i}{N} - (\bar{x})^2 = \frac{2^2 + 2 \cdot 3^2 + 6^2 + 3 \cdot 7^2}{7} - 5^2 = 4.28571$$

Nótese que se RECOMIENDA INTENSAMENTE aplicar la expresión empleada arriba y se DESACONSEJA usar su equivalente:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 \rightarrow DESACONSEJADA$$

Esta otra es más CLARA en cuanto al SENTIDO de la varianza σ pero es MUY TEDIOSO introducir las desviaciones de la media $(x_i - \bar{x})^2$ en la calculadora de mano, mientras que la tecla x^2 facilita MUCHO la aplicación de la expresión que no explicita esas desviaciones.

c) Calcular varianzas con CALCULADORA

En la calculadora CASIO fx-991SP X II Iberia, pulsar:

1. MENU > 6 > 1

2. Introducir datos en la tabla de frecuencias. Se usa = a modo de "intro".

3. OPTN > 3

4. SE CONSULTA la varianza POBLACIONAL σ^2
(NO CONFUNDIR CON LA CUASIVARIANZA MUESTRAL s^2)

1.4.4.6. Desviación típica POBLACIONAL σ

Se define la DESVIACIÓN TÍPICA POBLACIONAL σ como la raíz cuadrada positiva de la varianza poblacional:

$$\sigma = \sqrt{\sigma^2}$$

Para conocer la DESVIACIÓN TÍPICA POBLACIONAL σ basta con calcular la raíz cuadrada de la varianza.

1.4.4.7. Cuasivarianza MUESTRAL $\hat{s}^2$

NO CONFUNDIR con la varianza POBLACIONAL, denotada por σ^2.

Formalmente, la CUASIvarianza MUESTRAL se denota por $\hat{s}^2$ o bien $\hat{s}^2_{n-1}$ para distinguirla de la VARIANZA MUESTRAL s^2.

De ahora en adelante, $\hat{s}^2$ denota CUASIVARIANZA MUESTRAL, o sea, corregida, es decir, dividiendo entre n – 1.

Se define la CUASIVARIANZA MUESTRAL $\hat{s}^2$ como:

$$\hat{s}^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_N - \bar{x})^2}{n-1} = \frac{1}{n-1}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

Alternativamente, se recomienda:

$$\hat{s}^2 = \frac{(\sum_{i=1}^{N} x_i^2) - n(\bar{x})^2}{n-1}$$

El término n – 1 que caracteriza la expresión de la CUASIVARIANZA MUESTRAL $\hat{s}^2$ CORREGIDA responde a la denominada CORRECCIÓN DE BESSEL, y permite amortiguar el efecto de un SESGO que provoca la aplicación de la expresión de la varianza poblacional a una muestra de la población.

1.4.4.8. Cuasidesviación típica MUESTRAL $\hat{s}$

La cuasidesviación típica MUESTRAL se define como:

$$\hat{s} = \sqrt{\hat{s}^2}$$

Estos 2 parámetros MUESTRALES se aplican en INFERENCIA, mientras que no son protagonistas en estadística DESCRIPTIVA.

1.4.4.9. Relación entre cuasivarianza muestral $\hat{s}^2$ y varianza poblacional σ^2:

Se verifica:

$$\hat{s}^2 = \frac{N}{N-1}\sigma^2$$

1.4.4.10. Relación entre cuasidesviación muestral $\hat{s}$ y desviación típica poblacional σ:

Se verifica:

$$\hat{s} = \sqrt{\frac{N}{N-1}}\sigma$$

1.4.4.11. Síntesis de VARIANZA Y DESVIACIÓN POBLACIONAL Y MUESTRAL

Parámetros NO CORREGIDOS	Poblacional	**Muestral**
Varianza	$\sigma^2 = \frac{\sum x_i^2}{N} - \bar{x}^2$	$s^2 = \frac{\sum x_i^2}{n} - \bar{x}^2$
Desviación	$\sigma = \sqrt{\frac{\sum x_i^2}{N} - \bar{x}^2}$	$s = \sqrt{\frac{\sum x_i^2}{n} - \bar{x}^2}$
	SOLO SE USA EN ESTADÍSTICA DESCRIPTIVA	**NO SE USA EN ESTE CURSO**

Parámetros SÍ CORREGIDOS	**Poblacional**	Muestral
Cuasivarianza	$\hat{\sigma}^2 = \frac{(\sum_{i=1}^{N} x_i^2) - n(\bar{x})^2}{N-1}$	$\hat{s}^2 = \frac{\sum x_i^2 - n(\bar{x})^2}{n-1}$
Cuasidesviación	$\hat{\sigma} = \sqrt{\frac{(\sum_{i=1}^{N} x_i^2) - n(\bar{x})^2}{N-1}}$	$\hat{s} = \sqrt{\frac{\sum x_i^2 - n(\bar{x})^2}{n-1}}$
	NO SE USA EN ESTE CURSO	SOLO SE USA EN ESTADÍSTICA INFERENCIAL

1.4.4.12. Expresiones alternativas de la varianza POBLACIONAL

La expresión más PRÁCTICA para el cálculo a partir de la serie de DATOS o de la TABLA DE FRECUENCIAS es:

$$\sigma^2 = \frac{\sum_{i=1}^{N} n_i \cdot \mathrm{x_i^2}}{N} - \bar{x}^2$$

La expresión más indicativa del significado de la varianza es:

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

En función de las frecuencias ABSOLUTAS:

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

En función de las frecuencias RELATIVAS:

$$\sigma^2 = \sum_{i=1}^{N} f_i(x_i - \bar{x})^2$$

En caso que se disponga de clases de DATOS AGRUPADOS, el representante m de la marca de clase reemplaza al valor concreto de x, es decir:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N} n_i(m_i - \bar{x})^2$$

O bien:

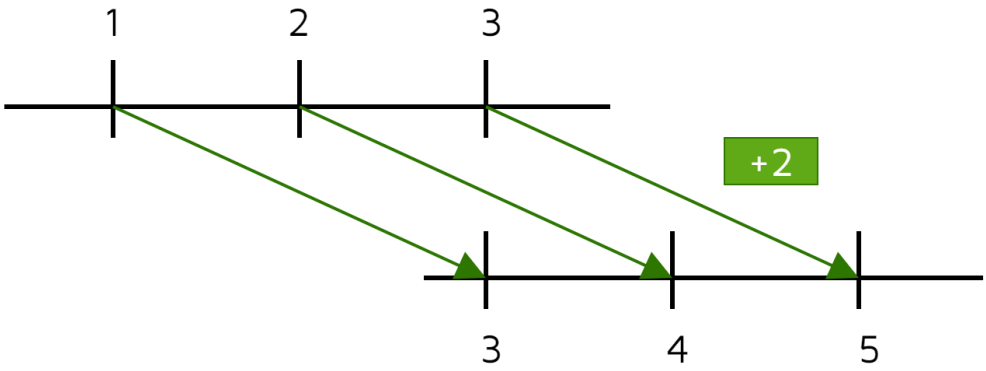
$$\sigma^2 = \sum_{i=1}^{N} f_i(m_i - \bar{x})^2$$

1.4.4.13. Efecto de las TRANSFORMACIONES LINEALES sobre la varianza σ^2 y sobre la desviación típica σ

Nótese que, si todas las observaciones adoptan el mismo valor, se cumple:

$$\sigma^2 = 0$$

a) Cambio de origen:



- Se cumple:

$$\sigma_y^2 = \sigma_x^2$$
$$\sigma_y = \sigma_x$$

- Tanto la varianza como la desviación típica son INMUNES a los cambios de origen.

- Es decir, si a todas las observaciones se suma un mismo número K, la varianza resultante σ_y^2 es la MISMA que la original σ_x^2. Así igual con la desviación típica.

Esto se demuestra:
Se escoge un escalar K.
Se definen las nuevas observaciones y a partir de las observaciones previas x:

$$y_i = x_i + K$$

Entonces, se cumple que la nueva media es:

$$\bar{y} = \bar{x} + K$$

De lo cual, la nueva varianza es:

$$\sigma_y^2 = \frac{(y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + \cdots + (y_N - \bar{y})^2}{N}$$

Es decir:

$$\sigma_y^2 = \frac{[x_1 + K - (\bar{x} + K)]^2 + [x_2 + K - (\bar{x} + K)]^2 + \cdots + [x_N + K - (\bar{x} + K)]^2}{N}$$

De modo que todos los términos K se cancelan y se alcanza:

$$\sigma_y^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_N - \bar{x})^2}{N} = \sigma_x^2$$

Con lo cual, se verifica que la varianza es INMUNE a cambios de ORIGEN (sumar K a todas las observaciones).

Para la desviación típica, por tanto, se cumple respecto el cambio de ORIGEN:

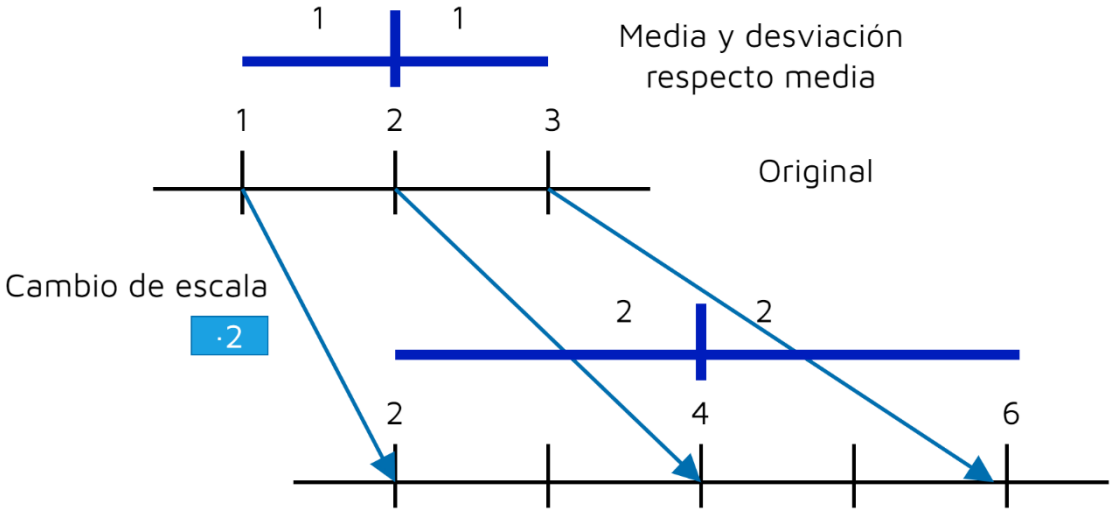
$$\sigma_y = \sigma_x$$

b) Cambio de escala:

- Se cumple:

$$\sigma_y^2 = K^2 \cdot \sigma_x^2$$
$$\sigma_y = |K| \cdot \sigma_x$$

- Nótese que K escala en un solo sentido la varianza: independientemente del signo del escalar K (ya que se encuentra al cuadrado).
- Si todas las observaciones se multiplican por un mismo número K, la varianza resultante σ_y^2 es la varianza original σ_x^2 multiplicada por EL CUADRADO de ese número K.
- Si todas las observaciones se multiplican por un mismo número K, la nueva DESVIACIÓN TÍPICA resultante σ_y es la desviación original σ_x multiplicada por el VALOR ABSOLUTO de ese número K.



Esto se demuestra:

Se escoge un escalar K.
Se define el conjunto de nuevas observaciones y a partir de las observaciones previas x:

$$y_i = x_i \cdot K$$

Entonces, se cumple que la nueva media es:

$$\bar{y} = \bar{x} \cdot K$$

De lo cual, la nueva varianza es:

$$\sigma_y^2 = \frac{(y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + \cdots + (y_N - \bar{y})^2}{N}$$

Es decir:

$$\sigma_y^2 = \frac{[Kx_1 - (K\bar{x})]^2 + [Kx_2 - (K\bar{x})]^2 + \cdots + [Kx_N - (K\bar{x})]^2}{N}$$

De modo que:

$$\sigma_y^2 = K^2 \frac{[(x_1 - \bar{x})]^2 + [(x_2 - \bar{x})]^2 + \cdots + [(x_N - \bar{x})]^2}{N} = K^2\sigma_x^2$$

Con lo cual, multiplicar todas las observaciones por K provoca en la varianza un escalado en un factor K^2.

Para la desviación típica, por tanto, se cumple con el cambio de ESCALA:

$$\sqrt{\sigma_y^2} = \sqrt{K^2\sigma_x^2} \rightarrow \sigma_y = |K| \cdot \sigma_x$$

Es decir:

$$\sigma_y = |K| \cdot \sigma_x$$

El efecto que tiene el cambio de ESCALA en la DESVIACIÓN TÍPICA es proporcional a K (no a K^2): nótese el VALOR ABSOLUTO de K.

1.4.5. REGLA DE TCHEBICHEV

Para un conjunto de datos con media $\bar{x}$ y desviación típica σ_x se cumple que, siendo m un número cualquiera, la proporción P de observaciones que incluye el intervalo $(\bar{x} - m\sigma_x, \bar{x} + m\sigma_x)$ es:

$$P(m) = 1 - \frac{1}{m^2}$$

Este es un resultado relacionado con la distribución de variables aleatorias. Nótese que para $m = \sqrt{2}$ se encuentra el intervalo en torno a la media que alberga el 50% de las observaciones:

$$\left(\bar{x} - \sqrt{2}\sigma_x, \bar{x} + \sqrt{2}\sigma_x\right) \textit{ aloja el } 50\% \textit{ de observaciones ya que}$$
$$P(\sqrt{2}) = 1 - \frac{1}{\left(\sqrt{2}\right)^2} = 0{,}5$$

Se observa:

Valor m	Intervalo	% mínimo de observaciones dentro del intervalo según Tchebichev
1	$(\bar{x} - 1\sigma_x, \bar{x} + 1\sigma_x)$	$1 - \frac{1}{1^2} = 0\%$
$\sqrt{2}$	$\left(\bar{x} - \sqrt{2}\sigma_x, \bar{x} + \sqrt{2}\sigma_x\right)$	$1 - \frac{1}{\left(\sqrt{2}\right)^2} = 50\%$
2	$(\bar{x} - 2\sigma_x, \bar{x} + 2\sigma_x)$	$1 - \frac{1}{2^2} = 75\%$
3	$(\bar{x} - 3\sigma_x, \bar{x} + 3\sigma_x)$	$1 - \frac{1}{3^2} = 88.\hat{8}\%$
4	$(\bar{x} - 4\sigma_x, \bar{x} + 4\sigma_x)$	$1 - \frac{1}{4^2} = 93.75\%$

Nótese que para m = 1 la regla no es informativa, y que para valores de m > 4 se encuentran más del 93.75% de las observaciones en el intervalo.

EJEMPLO

Se supone una muestra de 500 modelos de ordenador distintos, cuyo precio medio es de 2000€, con una desviación típica de 100€. La regla de Tchebichev permite asegurar que:

Como MÍNIMO, el precio del 75% de los ordenadores (es decir, de 350 ordenadores) se halla en el intervalo:

$$(\bar{x} - 2\sigma_x, \bar{x} + 2\sigma_x)$$

$$(2000 - 2 \cdot 100, 2000 + 2 \cdot 100) = (1800, 2200)$$

Como MÍNIMO, el precio del $88.\hat{8}\%$ de los ordenadores (es decir, de 444 ordenadores) está en el intervalo:

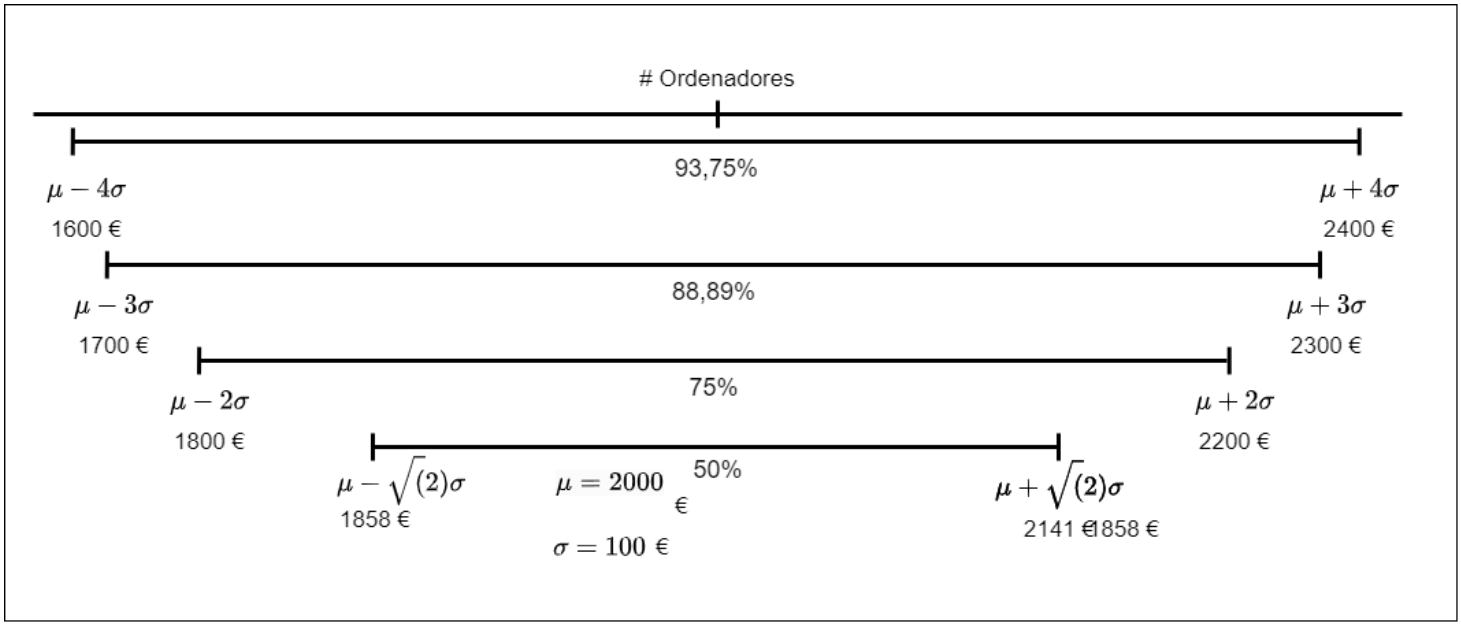
$$(\bar{x} - 3\sigma_x, \bar{x} + 3\sigma_x)$$

$$(2000 - 3 \cdot 100, 2000 + 3 \cdot 100) = (1700, 2300)$$

Como MÍNIMO, el precio del 93.75% de los ordenadores (es decir, de 468 ordenadores) está en el intervalo:

$$(\bar{x} - 4\sigma_x, \bar{x} + 4\sigma_x)$$

$$(2000 - 4 \cdot 100, 2000 + 4 \cdot 100) = (1600, 2400)$$



1.4.6. CÁLCULO DE VARIANZA A PARTIR DE DATOS TABULADOS

En la construcción de la tabla de frecuencias, se puede optar por:

x_i	n_i	$x_i \cdot n_i$	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
2	6	12		
3	15	45		
4	10	40		
5	9	45		
TOTAL	40			

1. Encontrar la media a partir de la tercera columna
2. Calcular las desviaciones
3. Luego calcular el cuadrado de las desviaciones
4. Aplicar la definición de varianza.

Es decir:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i \cdot n_i}{N} = \frac{12 + 45 + 40 + 45}{40} = \frac{142}{40} = 3.55$$

xi	ni	xi · ni	xi - $\bar{x}$	$(x_i - \bar{x})^2$
2	6	12	2 – 3.55	$(2 - 3.55)^2$
3	15	45	3 – 3.55	$(3 - 3.55)^2$
4	10	40	4 – 3.55	$(4 - 3.55)^2$
5	9	45	5 – 3.55	$(5 - 3.55)^2$
TOTAL	40			

Y aplicar la definición:

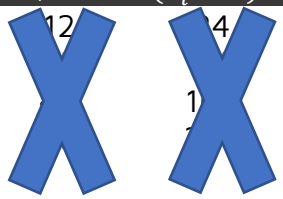
$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N} n_i (x_i - \bar{x})^2$$

Lo cual es:

$$\sigma^2 = \frac{6 \cdot (2 - 3.55)^2 + 15 \cdot (3 - 3.55)^2 + 10 \cdot (4 - 3.55)^2 + 9 \cdot (5 - 3.55)^2}{40} = 0.9975$$

Este método es MUY TEDIOSO. En lugar de calcular las desviaciones:

x_i	n_i	$x_i \cdot n_i$	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
2	6	12	12	4
3	15	45		
4	10	40		1
5	9	45		



Se recomienda, una vez calculada la media, calcular $x_i^2 \cdot n_i$ y aplicar la definición equivalente de varianza:

x_i	n_i	$x_i \cdot n_i$	$x_i^2 \cdot n_i$
2	6	12	24
3	15	45	135
4	10	40	160
5	9	45	225
TOTAL	40		

$$\sigma^2 = \frac{\sum_{i=1}^{n} x_i^2 \cdot n_i}{N} - \bar{x}^2 = \frac{24 + 135 + 160 + 225}{40} - 3.55^2 = 0.9975$$

1.4.7. COEFICIENTE DE VARIACIÓN (CV)

Informa sobre la representatividad de la media de los datos. Se calcula como:

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

Se cumple que $CV \geq 0$. Según el valor de CV se puede inferir:

$$\begin{cases} 0 < CV < 0.5 & Hay\ que\ cuestionarse\ la\ representatividad\ de\ la\ media \\ CV > 1 & La\ media\ NO\ ES\ representativa \\ CV = 0 & La\ mesdia\ ES\ SUFICIENTEMENTE\ representativa \end{cases}$$

1.5. Cambios de escala y de origen

Las modificaciones de la muestra tienen un impacto definido en cada uno de los estadísticos descritos.
Se distinguen 2 tipos de modificaciones sobre una variable:
a) Cambio de escala (producto de la variable por un escalar).
b) Cambio de origen (suma de un escalar a la variable).

CAMBIOS DE ESCALA			CAMBIOS DE ORIGEN	
$x \to y = C \cdot x$ La variable se modifica mediante el producto por un escalar			$x \to y = C + x$ La variable se modifica mediante la suma de un escalar	
DEFINICIÓN	AFECTA	ESTADÍSTICOS	AFECTA	DEFINICIÓN
$\bar{y} = C \cdot \bar{x}$	Sí	Media	Sí	$\bar{y} = \bar{x} + C$
$Mo_y = C \cdot Mo_x$	Sí	Moda	Sí	$Mo_y = Mo_x + C$
$Me_y = C \cdot Me_x$	Sí	Mediana	Sí	$Me_y = Me_x + C$
$\sigma_y^2 = C^2 \cdot \sigma_y^2$	Sí	Varianza	NO	$\sigma_y^2 = \sigma_x^2$
$\sigma_y = \lvert C \rvert \cdot \sigma_x$	Sí	Desviación típica σ	NO	$\sigma_y = \sigma_x$
$CV_y = CV_x$	NO	Coeficiente de variación	Sí	$CV_y = \frac{\sigma_x}{\bar{x} + C}$

1.6. Datos estandarizados

Una forma de hacer comparables los datos procedentes de distintos conjuntos es la ESTANDARIZACIÓN.

Es una aplicación de las propiedades de ciertas distribuciones de variables aleatorias conocidas como NORMALES.

Los valores estandarizados se denominan z-valores.

La estandarización consiste en transformar los datos de un conjunto X de modo que el conjunto transformado Y responda a una distribución cuya media es $\bar{z} = 0$ y con desviación típica $\sigma_z = 1$.

Esto se consigue mediante una sencilla transformación:

$$\underbrace{x_i}_{\substack{Valores \\ de\ origen}} \rightarrow \underbrace{z_i = \frac{x_i - \bar{x}}{\sigma_x}}_{\substack{Valores \\ estandarizados}}$$

Nótense 3 propiedades:

1) La media del conjunto Z transformado es nula $\bar{z} = 0$:

$$\bar{z} = \frac{z_1 + z_2 + \cdots + z_N}{N} = \frac{\frac{x_1 - \bar{x}}{\sigma_x} + \frac{x_2 - \bar{x}}{\sigma_x} + \cdots + \frac{x_N - \bar{x}}{\sigma_x}}{N} = \frac{\frac{x_1 + x_2 + \cdots x_3}{\sigma_x} - \frac{N\bar{x}}{\sigma_x}}{N}$$

Pero ya se ha demostrado:

$$\sum_{i=1}^{N} x_i = N\bar{x}$$

Entonces:

$$\bar{z} = \frac{\frac{x_1 + x_2 + \cdots x_3}{\sigma_x} - \frac{N\bar{x}}{\sigma_x}}{N} = \frac{\frac{N\bar{x}}{\sigma_x} - \frac{N\bar{x}}{\sigma_x}}{N} = 0$$

2) La desviación típica del conjunto Z transformado es $\sigma_z = 1$:

$$\sigma_z^2 = \frac{(z_1 - \bar{z})^2 + (z_2 - \bar{z})^2 + \cdots + (z_N - \bar{z})^2}{N} \underbrace{=}_{\bar{z}=0} \frac{(z_1 - 0)^2 + (z_2 - 0)^2 + \cdots + (z_N - 0)^2}{N}$$

De lo cual, tomando $z_i = \frac{x_i - \bar{x}}{\sigma_x}$:

$$\sigma_z^2 = \frac{\left(\frac{x_1 - \bar{x}}{\sigma_x}\right)^2 + \left(\frac{x_2 - \bar{x}}{\sigma_x}\right)^2 + \cdots + \left(\frac{x_N - \bar{x}}{\sigma_x}\right)^2}{N} = \frac{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_N - \bar{x})^2}{\sigma_x^2}}{N}$$

Entonces:

$$\sigma_z^2 = \frac{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_N - \bar{x})^2}{\sigma_x^2}}{N} = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_N - \bar{x})^2}{\sigma_x^2 N}$$

Desde lo cual se alcanza la expresión de σ_x^2 propiamente:

$$\sigma_z^2 = \frac{1}{\sigma_x^2} \underbrace{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_N - \bar{x})^2}{N}}_{\sigma_x^2} = \frac{\sigma_x^2}{\sigma_x^2} = 1$$

3) Esta expresión estandarizada de los valores permite, además, identificar rápidamente cuánto distan de la media (aislando x):

$$z_i = \frac{x_i - \bar{x}}{\sigma_x} \rightarrow x_i = \bar{x} + z_i \sigma_x$$

Es decir, el valor x_i dista de la media $\bar{x}$ un número z_i de desviaciones típicas σ_x.

1.7. Construcción de tabla de frecuencias

La tabla de frecuencias describe al DISTRIBUCIÓN DE FRECUENCIAS de la variable, es decir, el conjunto de valores que adopta la variable y con qué frecuencia los adopta.

Se considera una muestra representada en una tabla:

n_i	x_i
15	3
10	4
6	2
9	5

x_i = Valor que adopta la variable
n_i = Frecuencia absoluta

1. Se ORDENAN los datos de acuerdo a los valores que toma x_i de forma CRECIENTE para facilitar el posterior cálculo de estadísticos. Es MUY RECOMENDABLE dejar los VALORES de datos en la PRIMERA COLUMNA.

x_i	n_i
2	6
3	15
4	10
5	9

2. Se determina el tamaño de la muestra sumando el número de valores.

x_i	n_i
2	6
3	15
4	10
5	9
TOTAL	40

N = tamaño de la muestra (total de individuos)

3. Se calcula la frecuencia relativa ri que es la proporción entre cada frecuencia absoluta y el tamaño de la muestra. La suma de frecuencias relativas es 1 (el redondeo puede desviarlo).

x_i	n_i	r_i
2	6	6/40=0.15
3	15	15/40=0.375
4	10	10/40=0.25
5	9	9/40=0.225
TOTAL	40	1

r_i = Frecuencia relativa

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

4. Se calcula la frecuencia absoluta acumulada N_i que es la suma de las frecuencias absolutas de todos los valores menores y la del valor estudiado. La mayor frecuencia absoluta acumulada coincide con el tamaño de la muestra.

x_i	n_i	r_i	N_i
2	6	0.15	6
3	15	0.375	21
4	10	0.25	31
5	9	0.225	40
TOTAL	40	1	

N_i = Frecuencia absoluta acumulada:

$$N_j = n_1 + n_2 + n_3 + \cdots + n_j$$

5. Se calcula la frecuencia relativa acumulada R_i que es la suma de las frecuencias relativas de todos los valores menores y la del valor estudiado. La mayor frecuencia relativa acumulada coincide con 1.

x_i	n_i	r_i	N_i	R_i
2	6	0.15	6	0.15
3	15	0.375	21	0.525
4	10	0.25	31	0.775
5	9	0.225	40	1
TOTAL	40	1		

R_i = Frecuencia relativa acumulada:

$$R_j = r_1 + r_2 + r_3 + \cdots + r_j$$

6. Se calcula la media aritmética. Para ello, es MUY ÚTIL añadir la columna con el producto entre cada valor y su frecuencia (accesoria a la media):

x_i	n_i	r_i	N_i	R_i	$x_i \cdot n_i$
2	6	0.15	6	0.15	12
3	15	0.375	21	0.525	45
4	10	0.25	31	0.775	40
5	9	0.225	40	1	45
TOTAL	40	1			

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i \cdot n_i}{N} = \frac{12 + 45 + 40 + 45}{40} = \frac{142}{40} = 3.55$$

7. Se calcula la varianza σ^2 (sigma al cuadrado). Para ello, se necesita $x_i^2 \cdot n_i$. Es MUY ÚTIL añadir una nueva columna (accesoria a la varianza) con el producto de la columna anterior (accesoria de la media) por x_i ya que:

$$x_i^2 \cdot n_i = (x_i \cdot n_i) \cdot x_i$$

x_i	n_i	r_i	N_i	R_i	$x_i \cdot n_i$	$x_i^2 \cdot n_i$
2	6	0.15	6	0.15	12	24
3	15	0.375	21	0.525	45	135
4	10	0.25	31	0.775	40	160
5	9	0.225	40	1	45	225
TOTAL	40	1				

$$\sigma^2 = \frac{\sum_{i=1}^{n} x_i^2 \cdot n_i}{N} - \bar{x}^2 = \frac{24 + 135 + 160 + 225}{40} - 3.55^2 = 0.9975$$

8. Desviación típica σ (sigma):
   Raíz de la varianza:

$$\sigma = \sqrt{\sigma^2} \rightarrow \sigma = \sqrt{0.9975} = 0.9987$$

9. La moda (Mo):
   Se identifica el valor asociado al MÁXIMO en la columna de FRECUENCIAS absolutas (NO en la de VALORES).

x_i NO!!	n_i SÍ!!	r_i	N_i	R_i	$x_i \cdot n_i$	$x_i^2 \cdot n_i$
2	6	0.15	6	0.15	12	24
3	15	0.375	21	0.525	45	135
4	10	0.25	31	0.775	40	160
5	9	0.225	40	1	45	225
TOTAL	40	1				

10. Mediana (Me):

- Valor que deja el 50% de las observaciones a su izquierda.
- Coincide con el segundo cuartil Q2.
- Para identificar la mediana, se recurre a la FRECUENCIA ACUMULADA.
- Para ello, es INDISPENSABLE que los valores de la columna xi ESTÉN ORDENADOS.
- Luego se divide entre 2 el tamaño muestral:

$$\frac{N}{2} = 20$$

- Se busca en la tabla QUÉ VALOR ES EL PRIMERO en dejar a su izquierda la mitad de los datos.

x_i	n_i	r_i	N_i	R_i	$x_i \cdot n_i$	$x_i^2 \cdot n_i$
2	6	0.15	6	0.15	12	24
3	15	0.375	21	0.525	45	135
4	10	0.25	31	0.775	40	160
5	9	0.225	40	1	45	225
TOTAL	40	1				

El 2 deja 6 a su izquierda → No incluye los 20.
El 3 deja 21 a su izquierda → Sí incluye los 20 → Me = 3.

11. Cuartiles y mediana

Q1 = valor que deja el 25% de las observaciones totales a su izquierda.
Se divide el tamaño muestral entre 4 se identifica qué valor incluye Q1:

$$\frac{N}{4} = \frac{40}{4} = 10$$

Se busca el primer valor con una frecuencia absoluta acumulada igual o mayor a 10.

x_i	n_i	r_i	N_i	R_i	$x_i \cdot n_i$	$x_i^2 \cdot n_i$
2	6	0.15	6	0.15	12	24
3	15	0.375	21	0.525	45	135
4	10	0.25	31	0.775	40	160
5	9	0.225	40	1	45	225
TOTAL	40	1				

$$\frac{N}{4} = \frac{40}{4} = 10 \rightarrow Q_1 = 3$$

Q2 = valor que deja el 50% de las observaciones a su izquierda.
Equivale a la mediana Me.

$$\frac{N}{4} \cdot 2 = \frac{40}{4} \cdot 2 = 20$$

Se busca el primer valor con una frecuencia absoluta acumulada igual o mayor a 20.

x_i	n_i	r_i	N_i	R_i	$x_i \cdot n_i$	$x_i^2 \cdot n_i$
2	6	0.15	6	0.15	12	24
3	15	0.375	21	0.525	45	135
4	10	0.25	31	0.775	40	160
5	9	0.225	40	1	45	225
TOTAL	40	1				

$$\frac{N}{4} \cdot 2 = \frac{40}{4} \cdot 2 = 20 \rightarrow Q_2 = 3$$

Q3 = valor que deja el 75% de las observaciones a su izquierda.

$$\frac{N}{4} \cdot 3 = \frac{40}{4} \cdot 3 = 30$$

Se busca el primer valor con una frecuencia absoluta acumulada igual o mayor a 30.

x_i	n_i	r_i	N_i	R_i	$x_i \cdot n_i$	$x_i^2 \cdot n_i$
2	6	0.15	6	0.15	12	24
3	15	0.375	21	0.525	45	135
4	10	0.25	31	0.775	40	160
5	9	0.225	40	1	45	225
TOTAL	40	1				

$$\frac{40}{4} \cdot 3 = 30 \rightarrow Q_3 = 4$$

Nótese que se puede calcular cualquier percentil que se desee:

EJEMPLO

Cálculo del Percentil 82 a partir de la frecuencia absoluta acumulada:

$$0.82 \cdot 40 = 32.8$$

Se busca el primer valor con una frecuencia absoluta acumulada igual o mayor a 32.8

x_i	n_i	r_i	N_i	R_i	$x_i \cdot n_i$	$x_i^2 \cdot n_i$
2	6	0.15	6	0.15	12	24
3	15	0.375	21	0.525	45	135
4	10	0.25	31	0.775	40	160
5	9	0.225	40	1	45	225
TOTAL	40	1				

El valor 5 deja a la izquierda el 82% de los datos.

12. Coeficiente de variación

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

En el caso del ejemplo anterior:

$$CV = \frac{\sigma}{\bar{x}} = \frac{0.9987}{3.55} = 0.28$$

Según el valor de CV se puede inferir:

$$\begin{cases} 0 < CV < 0.5 & Hay\ que\ cuestionarse\ la\ representatividad\ de\ la\ media \\ CV > 1 & La\ media\ NO\ ES\ representativa \\ CV = 0 & La\ mesdia\ ES\ SUFICIENTEMENTE\ representativa \end{cases}$$

Por tanto, en este ejemplo, la media no es especialmente representativa.

COROLARIO

x_i = Valor que adopta la variable.
n_i = Frecuencia absoluta del valor i.
N = Tamaño de la muestra.
N_i = Frecuencia absoluta acumulada del valor i.
r_i = Frecuencia relativa del valor i.
R_i = Frecuencia relativa acumulada del valor i.
Me = Mediana (Q2)
σ = Desviación típica poblacional.
σ^2 = Varianza poblacional.
Mo = Moda

ENTREGABLE 1 – Cuestionario

PREGUNTA 1

Dadas las siguientes observaciones sobre el número y la superficie de los pisos de una inmobiliaria:

n_i	x_i
3	71
9	88
7	113
5	75

Se observa es que las columnas de la tabla están permutadas respecto el orden habitual de presentación. Conviene reescribirla para facilitar su lectura:

valor	frecuencia
x_i	n_i
71	3
88	9
113	7
75	5

Calculad:

a) La media:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i \cdot n_i}{n} = \frac{71 \cdot 3 + 88 \cdot 9 + 113 \cdot 7 + 75 \cdot 5}{3 + 9 + 7 + 5} = 90.4583\ m^2$$

b) La desviación estándar poblacional:

$$\sigma = \sqrt{\sigma^2}$$

Es decir:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} x_i^2 \cdot n_i}{n} - \bar{x}^2}$$

En este caso:

$$\sigma = \sqrt{\frac{71^2 \cdot 3 + 88^2 \cdot 9 + 113^2 \cdot 7 + 75^2 \cdot 5}{24} - 90.4583^2} = 15.7347$$

c) La mediana Me:

$$\frac{n}{2} = \frac{24}{2} = 12 \rightarrow Q_2 = 88.000\ m^2$$

$El\ valor\ x_i = 88\ es\ el\ menor\ valor\ que\ deja\ a\ su\ izquierda\ 12\ valores.$

PREGUNTA 2

Los siguientes datos corresponden a los gastos mensuales en libros de 6 estudiantes de la UOC:

$$\{1,4,10,4,2,10\}$$

Calculad su media, desviación estándar (poblacional) y la mediana.

En primer lugar, se construye la tabla de frecuencias con las columnas adecuadas para facilitar el cálculo de la media, la desviación estándar y la mediana.

x_i	n_i	N_i	$x_i \cdot n_i$	$x_i^2 \cdot n_i$
1	1	1	1	1
2	1	2	2	4
4	2	4	8	32
10	2	6	20	200
TOTAL	6	6	31	237

a) Cálculo de la media:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i \cdot n_i}{n} = \frac{31}{6} = 5.\widehat{16}$$

b) Cálculo de la desviación estándar (poblacional):

$$\sigma_x = \sqrt{\sigma_x^2} \rightarrow \sigma_x = \sqrt{\frac{\sum_{i=1}^{n} x_i^2 \cdot n_i}{n} - \bar{x}^2}$$

$$\sigma_x = \sqrt{\frac{237}{6} - 5.\widehat{16}^2} = \sqrt{12.8052} = 3.5784$$

c) Cálculo de la mediana:

$$\frac{n}{2} = \frac{6}{2} = 3 \rightarrow Q_2 = 4$$

$$El\ valor\ x_i = 4\ es\ el\ menor\ valor\ que\ deja\ a\ su\ izquierda\ 3\ valores$$

Para entenderlo, se representan los valores en 2 conjuntos con la misma cantidad de observaciones: uno por debajo de la mediana, y otro por encima.

Inferiores a Q_2	1	2	4				
Mediana				Q_2			
Superiores a Q_2					4	10	10

El valor exactamente entre 4 y 4 es el propio 4.

PREGUNTA 3

Se supone una muestra de tamaño 102 con media es 44.451, mediana 42, moda 14 y desviación estándar 21.733.

Para este ejercicio, se debe considerar:

CAMBIOS DE ESCALA			CAMBIOS DE ORIGEN	
$x \to y = C \cdot x$ La variable se modifica mediante el producto por un escalar			$x \to y = C + x$ La variable se modifica mediante la suma de un escalar	
DEFINICIÓN	AFECTA	ESTADÍSTICOS	AFECTA	DEFINICIÓN
$\bar{y} = C \cdot \bar{x}$	Sí	Media	Sí	$\bar{y} = \bar{x} + C$
$Mo_y = C \cdot Mo_x$	Sí	Moda	Sí	$Mo_y = Mo_x + C$
$Me_y = C \cdot Me_x$	Sí	Mediana	Sí	$Me_y = Me_x + C$
$\sigma_y^2 = C^2 \cdot \sigma_y^2$	Sí	Varianza	NO	$\sigma_y^2 = \sigma_x^2$
$\sigma_y = \lvert C \rvert \cdot \sigma_x$	Sí	Desviación típica σ	NO	$\sigma_y = \sigma_x$
$CV_y = CV_x$	NO	Coeficiente de variación	Sí	$CV_y = \frac{\sigma_x}{\bar{x} + C}$

a. Si se suma 35 a todas las observaciones:

Se trata de un **CAMBIO DE ORIGEN**

La nueva media es: $\bar{x} + C = \bar{y} \to 44.451 + 35 = 79.451$
La nueva mediana es: $Me_x + C = Me_y \to 42 + 35 = 79$
La nueva moda es: $Mo_x + C = Mo_y \to 14 + 35 = 49$
La nueva desviación estándar es: $\sigma_y = \sigma_x \to \sigma_y = 21.733$ **(NO VARÍA)**

b. Si se multiplican por 4 todas las observaciones:

Se trata de un **CAMBIO DE ESCALA**

La nueva media es: $\bar{x} \cdot C = \bar{y} \to 44.451 \cdot 4 = 177.804$
La nueva mediana es: $Me_x \cdot C = Me_y \to 42 \cdot 4 = 168$
La nueva moda es: $Mo_x \cdot C = Mo_y \to 14 \cdot 4 = 56$
La nueva desviación estándar es: $\sigma_x \cdot |C| = \sigma_y \to 21.733 \cdot |4| = 86.932$

c. Si se multiplican por -8 todas las observaciones:

Se trata de un **CAMBIO DE ESCALA**

La nueva media es: $\bar{x} \cdot C = \bar{y} \to 44.451 \cdot (-8) = -355.608$
La nueva mediana es: $Me_x \cdot C = Me_y \to 42 \cdot (-8) = -362$
La nueva moda es: $Mo_x \cdot C = Mo_y \to 14 \cdot (-8) = -112$
La nueva desviación estándar es: $\sigma_x \cdot C = \sigma_y \to 21.733 \cdot |(-8)| = 173.864$

d. Si se suma 31 a las 89 observaciones más grandes de la muestra, la nueva mediana es:

Por definición de mediana, se pueden agrupar la mitad de los valores por encima de la mediana y la otra mitad por debajo.

La muestra es de 102 observaciones, por tanto, la mediana es el valor que ocupa la posición $\frac{102}{2} = 51$ en la colección ordenada de la muestra.

Si se modifican las 89 observaciones mayores, el valor de la mediana (que ocupa la posición 51) se verá afectada por esta transformación de la muestra, ya que está incluido entre ellos.
Por tanto, la nueva mediana será 42 + 31 = 73.

Para no verse afectada la Me, como máximo, se deberían haber modificado hasta las 50 observaciones mayores.

PREGUNTA 4

Los siguientes datos corresponden a la cantidad de asignaturas superadas de 9 estudiantes de la UOC. Calculad sus cuartiles y la mediana:
{12,5,19,10,20,12,10,6,1}

Se observa un número IMPAR de observaciones (n= 9), por tanto, se debe poder observar Q_2 (coincide con la mediana) en la serie ordenada.

En primer lugar, se ordenan los valores:

$$\{1,5,6,10,10,12,12,19,20\}$$

La posición central dejará 4 valores a su izquierda y 4 a su derecha. Por tanto, se trata de la posición 5.

$$\{\underbrace{1,5,6,10}_{<Q_2}, \underbrace{10}_{Q_2}, \underbrace{12,12,19,20}_{>Q_2}\}$$

$$La\ mediana\ ocupa\ la\ posición\ 5 \rightarrow \mathrm{Me} = 10$$

Se puede representar como tabla

Inferiores a Q_2	1	5	6	10					
Mediana					10				
Superiores a Q_2						12	12	19	20

$$Por\ tanto, la\ mediana\ es\ 10.$$

En segundo lugar, se calculan los cuartiles subdividiendo los conjuntos en que la muestra queda dividida por la mediana (valores inferiores a Q_2 y valores superiores a Q_2). Es decir:

Se toma el subconjunto izquierdo para calcular Q_1:

1	5	Q_1	6	10

Q_1 ocupa el valor central del subconjunto izquierdo: la posición 2.5 (exactamente entre el segundo y el tercer valor: entre 5 y 6). Se interpola:

$$Q_1 = \underbrace{5}_{Valor\ previo\ a\ Q_1} + \underbrace{0.5}_{\substack{Se\ busca\ un\\ valor\ central\\ "a\ MEDIO\ camino"}} \cdot \underbrace{(6 - 5)}_{Diferencia\ recorrida}$$

$$Q_1 = 5 + 0.5 \cdot 1 = 5 + 0.5 = 5.5$$

Esto es equivalente a hacer su promedio:

$$Q_1 = \frac{5+6}{2} = 5.5$$

Y ahora el subconjunto derecho para calcular Q_3:

12	12	Q_3	19	20

Q_3 ocupa el valor central del subconjunto izquierdo: la posición 2.5 (exactamente entre el segundo y tercer valor: entre 12 y 19). Se interpola:

$$Q_3 = \underbrace{12}_{Valor\ PREVIO\ a\ Q_3} + \underbrace{0.5}_{\substack{Se\ busca\ un\\ valor\ central\\ "a\ MEDIO\ camino"}} \cdot \underbrace{(19 - 12)}_{Diferencia\ recorrida}$$

$$Q_3 = 12 + 0.5 \cdot 7 = 12 + 3.5 = 15.5$$

De forma alternativa, para verlo desde la derecha:

$$Q_3 = \underbrace{19}_{Valor\ POSTERIOR\ a\ Q_3} - \underbrace{0.5}_{\substack{Se\ busca\ un\\ valor\ central\\ "a\ MEDIO\ camino"}} \cdot \underbrace{(19 - 12)}_{Diferencia\ recorrida}$$

$$Q_1 = 19 - 0.5 \cdot 7 = 19 - 3.5 = 15.5$$

PREGUNTA 5

Los siguientes datos corresponden a la cantidad de asignaturas superadas por 8 estudiantes de la UOC:

x_i	n_i
4	1
9	4
15	1
17	2

Calculad sus cuartiles y la mediana.

Se observa un número PAR de observaciones (n = 8), por tanto, Q_2 se ha de poder calcular como el promedio de las 2 observaciones adyacentes al valor que deja 4 observaciones a cada lado (que es propiamente Me).

En primer lugar, se ordenan los valores extraídos de la tabla para calcular Q_2:

$$\{4,9,9,9,9,15,17,17\}$$

Se observa que el valor 9 ocupa la 4 posición: deja 4 valores por encima de sí y 4 valores por debajo de sí. La mediana es 9.

$$\{\underbrace{4,9,9,9}_{<Q_2}, \underbrace{9}_{Me}, \underbrace{15,17,17}_{>Q_2}\}$$

Alternativamente, se divide la muestra **ORDENADA** en 2 conjuntos con el MISMO número de observaciones y se evalúa el valor que ocupa la posición central. Para una serie de n observaciones, el valor en la posición central NO SE OBSERVA, pero deja $\frac{n}{2}$ a cada lado. en este caos, deja 4.

Inferiores a Q_2	4	9	9	9					
Mediana					Q_2				
Superiores a Q_2						9	15	17	17

La mediana es el valor que ocupa la quinta posición: entre 9 y 9.

$$Por\ tanto, la\ mediana\ es\ 9.$$

En segundo lugar, se calculan los cuartiles subdividiendo los conjuntos en que la muestra queda dividida por la mediana (valores inferiores a Q_2 y valores superiores a Q_2). Es decir:

Se toma el subconjunto izquierdo para calcular Q_1:

4	9	Q_1	9	9

Q_1 ocupa el valor central del subconjunto izquierdo: la posición 2.5 (exactamente entre el segundo y el tercer valor: entre 9 y 9). Interpolando o haciendo el promedio entre 9 y 9:

$$Q_1 = \underbrace{9}_{Valor\ previo\ a\ Q_1} + \underbrace{0.5}_{\substack{Se\ busca\ un\\ valor\ central\\ "a\ MEDIO\ camino"}} \cdot \left(\underbrace{9 - 9}_{Diferencia\ recorrida} \right)$$

$$Q_1 = 9$$

Y ahora el subconjunto derecho para calcular Q_3:

9	15	Q_3	17	17

Q_3 ocupa el valor central del subconjunto izquierdo: la posición 2.5 (exactamente entre el segundo y tercer valor: entre 15 y 17).
Se interpola o se promedia entre 15 y 15:

$$Q_3 = \underbrace{15}_{Valor\ previo\ a\ Q_3} + \underbrace{0.5}_{\substack{Se\ busca\ un\\ valor\ central\\ "a\ MEDIO\ camino"}} \cdot \left(\underbrace{17 - 15}_{Diferencia\ recorrida} \right)$$

$$Q_3 = 15 + 1 = 16$$

PREGUNTA 6

Se estudian las preferencias de nadadores federados en cuanto a equipación.

Completa las siguientes oraciones:

1. Un ejemplo de muestreo por cuotas sería…
2. Si tenemos una lista de los nadadores federados y enviamos un correo a los 30 primeros de la lista…
3. Si asignamos un número a cada uno de los N nadadores federados, elegimos al azar uno A de entre y a partir de este escogemos de la forma A+x(N/K) done K es el tamaño de la muestra…
4. Si seleccionamos 4 comarcas al azar y después tomamos 59 nadadores federados de estas comarcas…

a. No estamos haciendo un muestreo aleatorio.
b. Estamos haciendo un muestreo sistemático.
c. Estamos haciendo un muestreo por conglomerados.
d. Enviar entrevistadores hasta encontrar 100 mujeres y 200 hombres entre los nadadores federados.

1 – d) El muestreo por cuotas se suele asociar a la distinción de las partes de la muestra por alguna de sus características. En este caso, el sexo.
2 – a) Tomar los 30 primeros no permite que la probabilidad de todos los elementos de la muestra de ser examinados sea igual. Por tanto, NO ES ALEATORIO.
3 – b) Esta es la definición de la técnica de muestreo sistemático.
4 – c) El muestreo por conglomerados se caracteriza por asumir que las partes de la muestra (cada Proción de 59 observaciones) son comparables pese a estar diferenciadas por alguna característica, en este caso, la geográfica.

ENTREGABLE 1 – Práctica de R

PREGUNTA 1

Encontrad los resúmenes numéricos y un histograma de la variable PM10Concentration1999 y comentad el resultado.

La tabla de datos se ha importado como fichero .csv introduciendo el comando:

```
datos<-read.table("C:/Users/Tete/Downloads/AP.csv",
header=TRUE,sep=";",
na.strings="NA",
fileEncoding="UTF-8",
quote="\"",
colClasses=c(rep("character",4),rep("numeric",2),rep("character",2)))
```

Nótese que la tabla ha sido asignados a la variable "datos".

Los resúmenes numéricos (que incluyen cuartiles, valores mínimo y máximo y media) de la variable evaluada se han obtenido con el método summary(). Es decir:

```
summary(datos$PM10Concentration1999)
```

Que ha devuelto:

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    6.0    24.0    38.0    51.1    71.0   359.0
```

Se observa que:

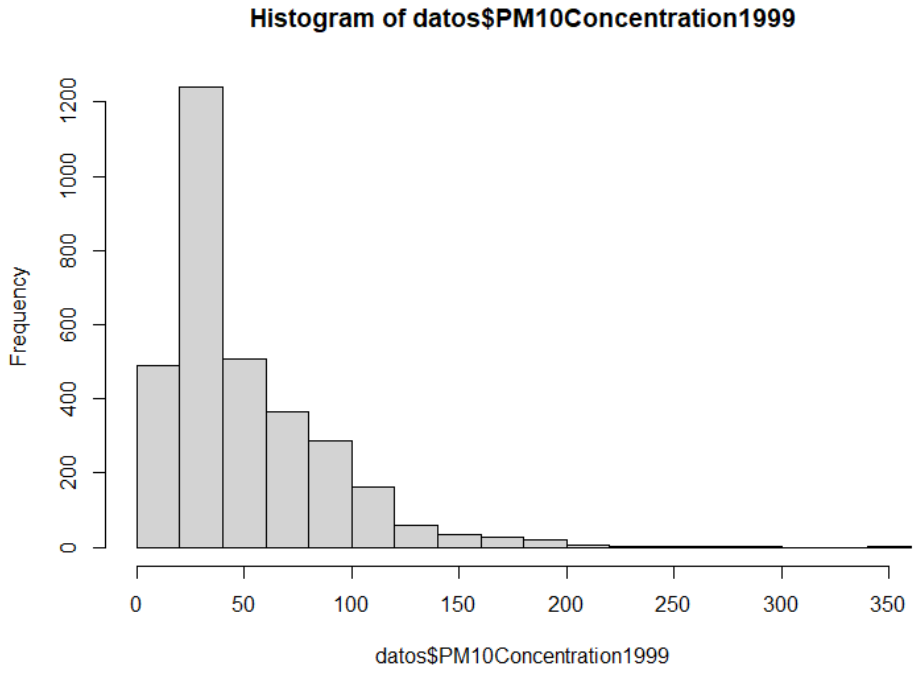
- El valor mínimo es 6.0.
- Que el valor 24.0 deja a su izquierda al menos el 25% de las observaciones.
- Que la mediana (38.0) es inferior a la media (71.0).
- Que el valor 71.0 deja a su izquierda al menos el 75% de las observaciones.
- El valor máximo es 359.0.

Además, en el panel entorno de variables se muestra un tamaño muestral de 3218 observaciones.

Para realizar el histograma se recurre a hist() y se escribe:

```
hist(datos$PM10Concentration1999)
```

Que devuelve el histograma:



De la observación del histograma se desprende que los datos de la variable estudiada están fuertemente desplazados a la izquierda, es decir, se observa una distribución unimodal con cola hacia la derecha, lo cual se relaciona con coeficiente de asimetría inferior a 0.

PREGUNTA 2

Encontrad los resúmenes numéricos y un histograma de la variable PM10Concentration1999 en las ciudades de España y comparadlo con el resultado anterior.

De nuevo, el resumen numérico para la variable estudiada se ha realizado con summary():

```
summary(datos$PM10Concentration1999[datos$Country=="Spain"])
```

En esta ocasión, se debe acceder a un subconjunto de datos etiquetados por el valor "Spain" en el campo "Country", es decir, se necesita tomar un fragmento de una columna de la tabla. Para ello, se ha empleado la sintaxis "tabla$columnaTomada[tabla$campoAFiltrar==valorDeseado].

Lo cual ha devuelto:

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  27.00   33.50   40.00   40.35   45.50   67.00
```

Se observa que:

- El valor mínimo es 27.
- Que el valor 33.5 deja a su izquierda al menos el 25% de las observaciones.
- Que la mediana (40) es muy cercana a la media (40.35).
- Que el valor 45.50 deja a su izquierda al menos el 75% de las observaciones.
- El valor máximo es 67.0.

En el panel entorno de variables se muestra un tamaño muestral de 55 observaciones, a lo cual también se puede acceder mediante:

```
length(datos$PM10Concentration1999[datos$Country=="Spain"])
```

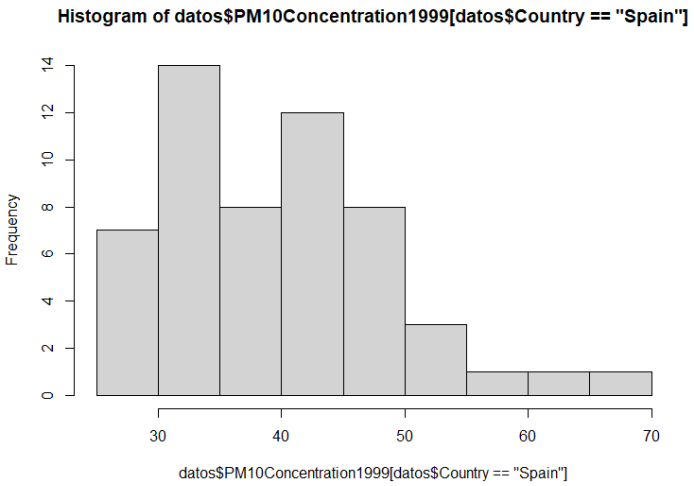
Que devuelve:

```
[1] 55
```

Para representar gráficamente el subconjunto filtrado, se ha escrito:

```
hist(datos$PM10Concentration1999[datos$Country=="Spain"])
```

Que devuelve el histograma:



A continuación, se comparan los datos de los resúmenes numéricos para todas las ciudades del fichero con los obtenidos para las ciudades de España:

	Todas las ciudades (n=3218)	Ciudades de España (n=55)
Mínimo	6.00	27.00
Primer cuartil	24.00	33.50
Mediana	38.00	40.00
Tercer Cuartil	71.00	45.50
Máximo	359.00	67.00
Media	51.10	40.35

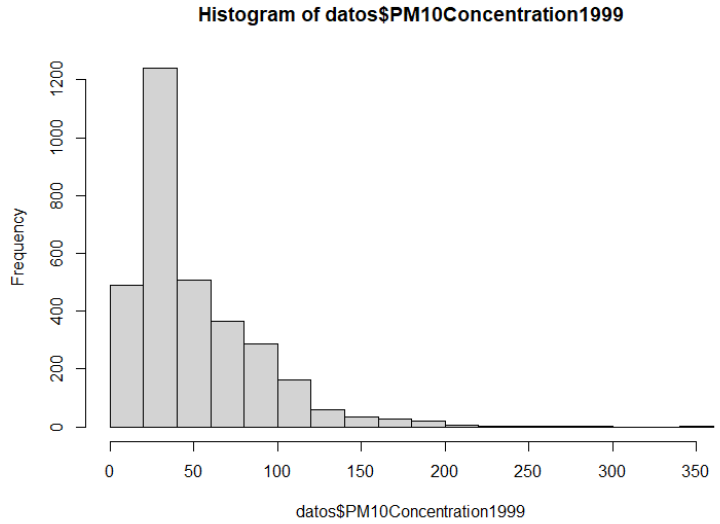
Se concluye que:

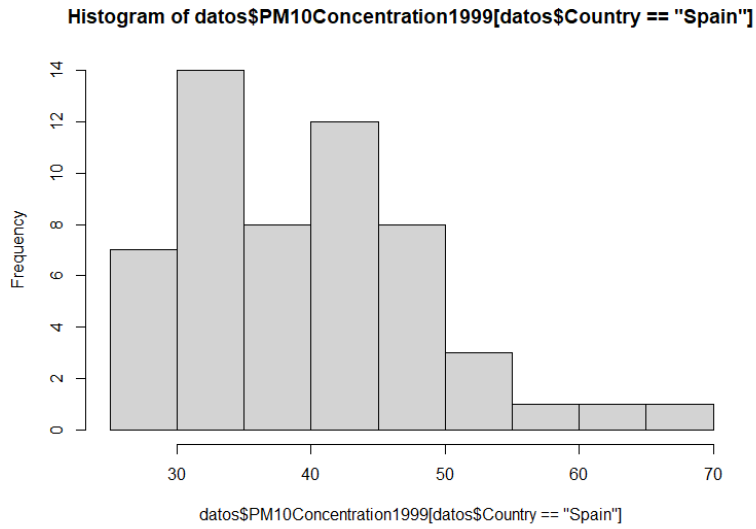
1. En primer lugar, se observa que los valores extremos mínimo y máximo se han acercado (uno ha subido y otro ha bajado).

2. En segundo lugar, se observa que la modificación de la muestra ha tenido un impacto significativo en el valor de la media.

3. Por último, es llamativa la relativamente leve alteración del valor de la mediana en comparación con la alteración de los otros estadísticos calculados.

En cuanto a los histogramas:





Se observa una distribución relativamente similar entre ambos, con una asimetría hacia la derecha mucho más marcada en el caso de los datos generales y mucho más suave en el caso de los datos de España.

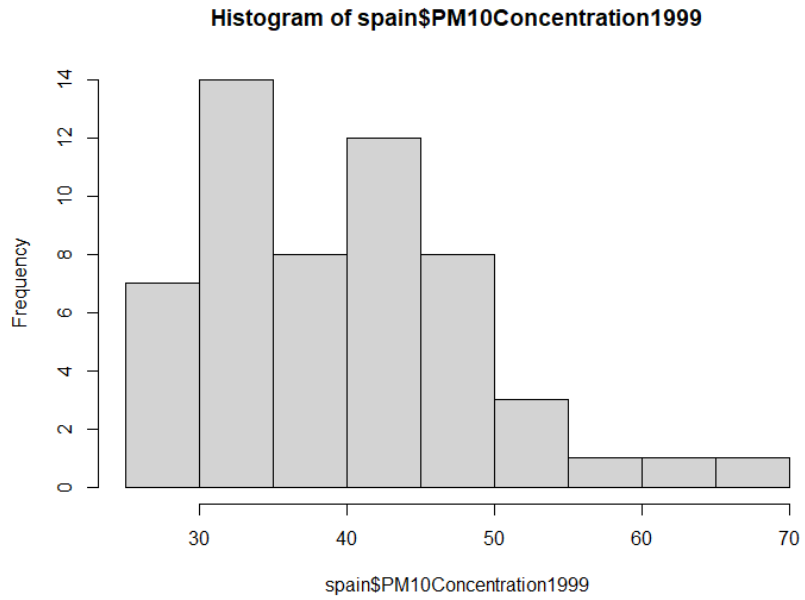
La representación refleja lo mencionado respecto los resúmenes numéricos: la muestra de España presenta una dispersión mucho menor y su media se encuentra desplazada hacia la izquierda respecto la media del conjunto.

APÉNDICE A LA PREGUNTA 2

Una forma alternativa de obtener el subconjunto de datos asociados al valor "Spain" de la variable "Country", es decir, de hacer el filtrado, sería el comando "subset". Por ejemplo, el siguiente comando devuelve el histograma anterior:

```
spain<-subset(datos,Country=="Spain")
```

hist(spain$PM10Concentration1999)



Asimismo, el filtrado de datos podría haberse realizado por el campo "Cod", es decir:

```
summary(datos$PM10Concentration1999[datos$Cod=="ESP"])
```

Devuelve los mismos datos que filtrando por "Country":

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  27.00   33.50   40.00   40.35   45.50   67.00
```

PREGUNTA 3

Encontrad el valor máximo de la variable PM10Concentration1999 y la ciudad y el correspondiente país donde se da este valor.

En primer lugar, se procede a localizar en la tabla el valor máximo de la variable PM10Concentration1999.

Para ello, se recurre al comando which.max, que devuelve la posición en que se encuentra el valor máximo dentro de la variable suministrada, que en este caso es una columna.

Por tanto, which.max devolverá la fila en la cual se encuentra el valor máximo.

Una vez localizada la posición de la variable, se accede a la fila que ocupa en la tabla mediante el comando datos$City[elemento_deseado] donde "elemento_deseado" es el resultado de which.max aplicado a la columna PM10Concentration1999.
Escrito en un solo comando:

```
print(datos$City[which.max(datos$PM10Concentration1999)])
```

Devuelve:

```
[1] "Nyala"
```

Se puede verificar que es correcto accediendo a la fila 2688 de la tabla:

```
print(datos[2668,])
```

Que devuelve:

```
Cod Country Citycode  City Population2000 PM10Concentration1999
Region          IncomeGroup
2668 SDN   Sudan  7360006 Nyala         306972              359 Sub-
Saharan Africa Lower middle income
```

Y contrastando el resultado con el valor máximo de la columna consultada:

```
max(datos$PM10Concentration1999)
```

Que devuelve:

```
[1] 359
```

De este modo queda validada la ciudad y se revela el país en que se encuentra el valor máximo de la variable consultada, que coincide a su vez con el resultado de:

```
print(datos$Country[2668])
```

Que es:

```
[1] "Sudan"
```

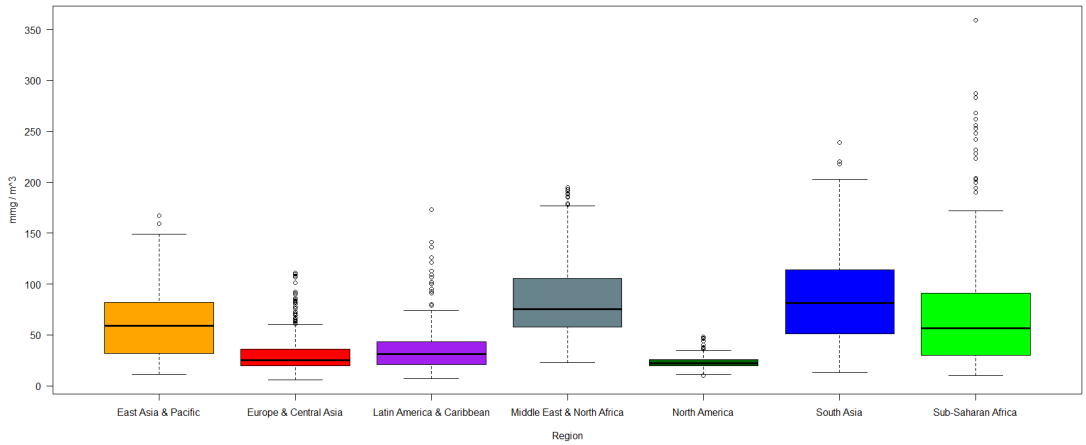
PREGUNTA 4

Obtened un Boxplot distinguiendo por regiones geográficas de la variable PM10Concentration1999. Podéis usar "las=2" en el boxplot para las etiquetas verticales. Para ampliar el área del gráfico, usad la instrucción par(mar=c(12,5,1,1)) o similar antes del boxplot. Comentad el resultado.

Para obtener un diagrama de cajas se ha escrito:

```
par(mar=c(12,8,2,2))
boxplot(datos$PM10Concentration1999~datos$Region,col=c("
orange","red"," purple","light blue4","dark green","blue","
green"),ylab="mmg / m^3",xlab="Region")
```

Se ha obtenido:



En primer lugar, sobre los valores de la media, se observa:

- Que la media en las regiones de Europa y Asia Central y Norte América presentan la media más baja, juntamente con América Latina y Caribe.
- Que las regiones con una media más elevada son Oriente Próximo y Norte de África junto con el Sur de Asia.

En segundo lugar, sobre la dispersión:

- La dispersión de los datos es mucho menor en Europa y Asia Central y Norte América que en el resto de las regiones (cajas MUY achatadas, ya que los valores Q1 y Q3 son próximos a la Mediana y, a la vez, bigotes mucho más cortos que en resto de regiones).
- La dispersión es mucho más elevada (bigotes muy largos y abundantes valores atípicos excluidos por encima) en el Sur de Asia y Oriente Próximo y Norte de África y especialmente en África Subsahariana, en la cual se observan valores atípicos mucho más elevados que en el resto.

En tercer lugar, los valores mínimos son muy similares entre las diversas regiones y ligeramente más elevados en Oriente Próximo y Norte de África.
Por último, los valores máximos más altos se observan en África Subsahariana, a la cual siguen Asia del Sur y Oriente Próximo y el Norte de África. Los máximos más modestos de todo el conjunto de datos se encuentran en Europa y Asia Central y Norte América.
Esta representación permite comparar de forma especialmente visual el rango intercuartílico asociado a cada región representada.
Cabe destacar que el método boxplot() agrupa los valores centrales y excluye valores atípicos mostrados como puntos independientes de la caja.

II. Muestreo

2.1. Concepto de muestreo

La población es el conjunto de individuos objeto de estudio.
La muestra (*sample*) es cualquier subconjunto de la población.

Las técnicas de muestreo (*sampling*) permiten la extracción de una muestra a partir de la población.

Se entiende por CENSO la lista de todos los individuos de la población.

Para una población de N individuos, de la cual se desea una muestra de k individuos existen $\binom{N}{k}$ (el número combinatorio se lee N sobre k) combinaciones diferentes.

El SESGO es aquella propiedad de la muestra POCO REPRESENTATIVA (en la cual algunos sectores de la población están sobrerrepresentados y otros infrarrepresentados).

2.2. Muestreo aleatorio simple

2.2.1. DEFINICIÓN DE MUESTREO ALEATORIO SIMPLE

Cumple 2 propiedades:

- Todos los elementos de la población tienen la MISMA PROBABILIDAD de ingresar en la muestra.
- La selección de individuos se realiza de uno en uno y con REPOSICIÓN (cada vez que uno se elige, se devuelve a la población, no se excluye) es decir, la selección de uno NO CONDICIONA la probabilidad de selección de los otros. Es decir, la selección es SIEMPRE un suceso INDEPENDIENTE.

Para la selección aleatoria, se pueden usar bombos con bolas, papeletas o tablas de dígitos aleatorios.

2.2.2. TABLAS DE DÍGITOS ALEATORIOS

Son tablas de números generadas por ordenador y validadas mediante la superación de pruebas de aleatoriedad e independencia.

Se usan para generar la lista de POSICIONES DEL CENSO DE UNA POBLACIÓN que integrarán la MUESTRA.

Se usa, por ejemplo, un dado para tomar qué DÍGITO servirá como punto de partida y, desde ahí, se forman grupos de números de la extensión adecuada para elegir las POSICIONES del CENSO que se tomarán para la muestra.

EJEMPLO
Se desea extraer una muestra de 4 individuos de una población de 900.
Si se dispone de la tabla de dígitos:

78027 23894 10394 72957

Se lanza un dado para identificar el punto de partida y se saca un 5.
Se parte, por tanto, de:

7802**7** 23894 10394 72957

Se forman grupos de 3 dígitos inferiores a 200 a partir del 7 indicado:
Se toma 7 23:

7802**7** **23**894 10394 72957

Se selecciona el individuo que ocupa la posición 723 del censo.
Se sigue con el 894, el 103, se DESCARTA el 947 (excede el censo) y el 295.

Por tanto, para formar la muestra, se eligen los individuos de las posiciones:
723, 894, 103 y 295

2.3. Muestreo sistemático

Se desea una muestra de tamaño k a partir de una población de N individuos.

1. Para ello, se numera el censo.
2. Se calcula un número m tal que:

$$m = \left[\frac{N}{k}\right]$$

Donde:

$$\left[\frac{N}{k}\right] = Parte\ ENTERA\ del\ cociente\ \frac{N}{k}$$

Nótese el uso del operador PARTE ENTERA: [*] devuelve la parte entera de *.

Es decir, se divide el tamaño de la población entre el tamaño de la muestra deseada y se toma LA PARTE ENTERA (se TRUNCA a 0 decimales).

3. Se elige un número entero L aleatoriamente entre 0 y m, que será la posición del primer individuo a ingresar en la muestra.
4. Se suma m al número L para obtener el siguiente individuo de la muestra.
5. Se repite hasta conseguir k individuos en la muestra.

EJEMPLO

Se tiene una población CENSADA de 1400 individuos.
Se desea una muestra de 12 individuos obtenida mediante muestreo sistemático.

1. Se calcula m:

$$m = \left[\frac{N}{k}\right] = \left[\frac{1400}{12}\right] = 116$$

2. Se toma, aleatoriamente, un número L entre 0 y m, por ejemplo, L = 39.
3. Se suma m + L para obtener la posición en el censo del primer individuo de la muestra:

$$m + L = 116 + 39 = 155$$

4. Se repite con 2m + L para el SEGUNDO:

$$2m + L = 2 \cdot 116 + 39 = 271$$

5. Se repite con 3m + L para el TERCERO:

$$3m + L = 3 \cdot 116 + 39 = 387$$

6. Y así hasta obtener 12.

2.4. Muestreo estratificado

En lugar de dar prioridad a la aleatoriedad, antepone la REPRESENTATIVIDAD de la muestra.

Se basa en el CONOCIMIENTO PREVIO de la población. En concreto, en el conocimiento de la SIMILITUD de los valores de la variable estudiada en ciertos segmentos de la población denominados ESTRATOS.

Estas similitudes entre ciertos segmentos de la población son una fuente de SESGO. Por ejemplo, hacer un sondeo de telefónico de voto en el año 1921 orientaría la preferencia, ya que el segmento de población con acceso al teléfono no representa la heterogeneidad de la población.

Para superar ese sesgo, se emplean los ESTRATOS.
Un estrato es cada uno de los subconjuntos disyuntos (excluyentes) de la población integrado por individuos que comparten cierta CARACTERÍSTICA COMÚN que condiciona el valor de la variable estudiada. Es decir, un estrato es un segmento homogéneo de la población estudiada.

Se aspira a componer la muestra con PROPORCIONES de cada estrato similares a las proporciones observadas en la población. Es decir, que la MUESTRA REPRODUZCA LA ESTRUCTURA DE LA POBLACIÓN.

Para ello, se extrae mediante muestreo aleatorio simple una proporción de individuos de cada estrato lo más fiel posible al tamaño relativo de ese estrato en la población. Se gana precisión:
- Cuanto más homogéneos sean en sí mismos los estratos.
- Cuanto más distintos sean entre ellos.

EJEMPLO

Se quiere obtener una muestra representativa del siguiente colectivo de 228 estudiantes, formada por 30 individuos.

En este caso, los estratos ya están hechos: son los cursos.

Se calcula el peso de cada curso en la población y se asigna el número de representantes en la muestra:

	Estudiantes	Proporción %	En la muestra de 30
Estudiantes de primer curso	68	$68/228 = 29.8\%$	$(68/228) \cdot 30 \approx 9$
Estudiantes de segundo curso	60	$60/228 = 26.3\%$	$(60/228) \cdot 30 \approx 8$
Estudiantes de tercer curso	52	$52/228 = 22.8\%$	$(52/228) \cdot 30 \approx 7$
Estudiantes de cuarto curso	48	$48/228 = 21.05\%$	$(48/228) \cdot 30 \approx 6$

2.5. Muestreo por conglomerados

Se aplica cuando NO se puede disponer del CENSO.
Un conglomerado es una unidad física (generalmente, geográfica) en que se distribuyen los individuos de la población.
Por tanto, los conglomerados son agrupaciones naturales de individuos.
El muestreo por conglomerados se basa en la selección aleatoria de conglomerados, de cada uno de los cuales se extrae una muestra.

Se aspira a que CADA CONGLOMERADO sea tan heterogéneo como la población entera, que REFLEJE LA ESTRUCTURA DE LA POBLACIÓN.

Que no todos los conglomerados estén reflejados en la muestra final conduce a 2 limitaciones:
- Si los conglomerados son MUY DISTINTOS ENTRE ELLOS, la REPRESENTATIVIDAD de la muestra será BAJA.
- Si los conglomerados son MUY HOMOGÉNEOS en sí mismos, habrá características SOBRERREPRESENTADAS y SUBREPRESENTADAS

Para ejecutarlo:
- Se identifican qué agrupaciones naturales son los conglomerados.
- Se seleccionan al azar algunos conglomerados.
- Se toma de cada conglomerado escogido una muestra aleatoria simple.
- La muestra final la integran la unión de las muestras procedentes de cada conglomerado.

→ Nótese que:
- Un estrato aspira a ser HOMOGÉNEO en sí mismo y lo más DISTINTO a los demás.
- Un conglomerado aspira a representar la HETEROGENEIDAD poblacional.

	Estratos	Conglomerados
Todos están representados en la muestra final	Sí	No
Para consigo mismo	Homogéneo	Reflejo de la heterogeneidad de la población
Entre ellos	Lo más distinto posible de los otros estratos	Lo más similar posible a los otros conglomerados

2.6. Muestreo polietápico

Combinación de los muestreos por conglomerados y estratificado.
Se estratifican los conglomerados, es decir, se agrupan en estratos.
Se obtienen muestras aleatorias de cada estrato de los conglomerados seleccionados.

Responde a la necesidad de mitigar el sesgo que se produce en caso de que los CONGLOMERADOS sean MUY HOMOGÉNEOS y, por tanto, no representen la heterogeneidad de la población total.

2.7. Muestreo por cuotas

Se prefiere como aproximación PRELIMINAR.

Se ejecuta cuando no es posible acceder al CENSO (lista) de la población, pero sí se dispone de información sobre la distribución de sus individuos. Las CUOTAS son un cierto número de INDIVIDUOS dentro de ciertas CATEGORÍAS (suelen ser género o edad).

En base al conocimiento de la distribución, se definen categorías, dentro de cada una de las cuales hay una proporción de los individuos.

Se aspira a que la muestra refleje esa misma proporción de cada categoría.

Para ello, se asigna cada CUOTA, es decir, el número de individuos de cada categoría que permite reflejar la estructura de la población en la muestra.

Se suele asignar luego el muestreo a ENTREVISTADORES que muestrean hasta que satisfacen las sucesivas cuotas.

2.8. Resumen muestreos

Muestreo	Características
Aleatorio simple	- Selección equiprobable de individuos. - Selección independiente (con reposición)
Sistemático	- Se selecciona el primero al azar - Se seleccionan los sucesivos a intervalos fijos definidos por L + m con m = [N/k] y L aleatorio entre 0 y m.
Estratificado	- Población dividida en ESTRATOS disjuntos de rasgos diferenciados, es decir, HOMOGÉNEOS. - La muestra se extrae por muestreo aleatorio simple de CADA UNO estrato. - La PROPORCIÓN de cada estrato en la muestra representa la estructura de la población.
Conglomerados	- Población distribuida en regiones de proximidad, CONGLOMERADOS cada uno de los cuales representa la HETEROGENEIDAD de la población. - Se seleccionan ALGUNOS de ellos, con individuos de cada uno de los cuales se integra la muestra final.
Polietápico	- Estratificación de conglomerados.
Por cuotas	- Se conoce la proporción de CATEGORÍAS en la población - La PROPORCIÓN de individuos de cada categoría en la muestra (ese número de individuos es una CUOTA) refleja la proporción de esa categoría en la población. - El entrevistador selecciona individuos hasta alcanzar las cuotas.