Probabilidad y Estadística - UTN

Estadística Descriptiva - Clase 3

Datos agrupados - Tabla de frecuencias

- Cuando el número de datos de un conjunto es considerado como grande, por ejemplo mayor a 20, es convieniente agruparlos para lograr un mejor análisis e interpretación.
- Para variables cualitativas, o cuantativas discretas con dominio de pocos valores, se utiliza una **tabla de datos agrupados puntualmente**
- Para variables cuantitativas con dominio de muchos valores o variables cuantitativas continuas se utiliza una **tabla de datos agrupados en intervalos**

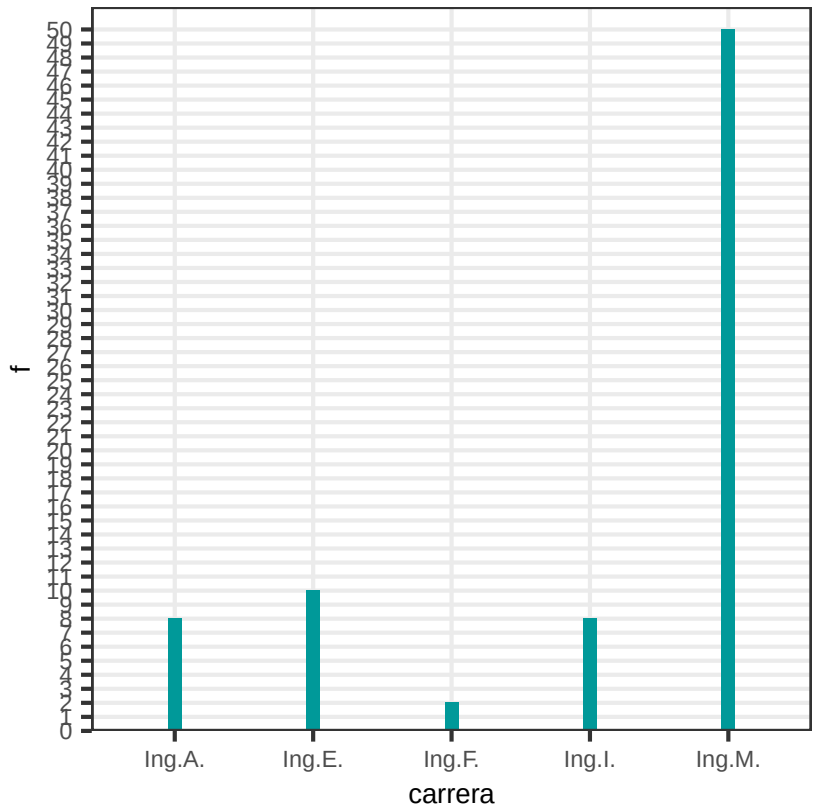
Datos agrupados puntualmente - Variable cualitativa

Se realizó una encuesta a un grupo de 78 estudiantes de la UTN Regional Haedo de las comisiones XX, 2°3° y 2°9° con el objetivo de conocer sus carreras de estudio y la cantidad de transportes públicos que utilizan para llegar hasta la universidad. Los resultados se muestran a continuación.

Tabla1. Distribución de frecuencias de las carreras de estudio.

	Carrera	f	F	fr	Fr
1	Ing.A.	8	8	0.10	0.10
2	Ing.E.	10	18	0.13	0.23
3	Ing.F.	2	20	0.03	0.26
4	Ing.I.	8	28	0.10	0.36
5	Ing.M.	50	78	0.64	1.00

Gráfico 1. Diagrama de bastones que representa a la tabla 1

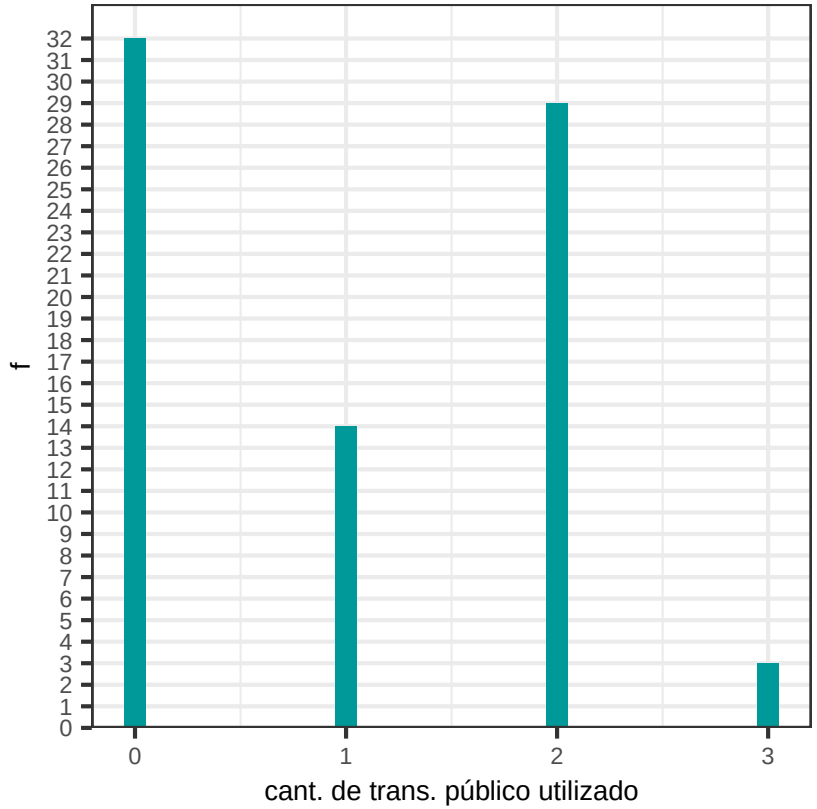


Datos agrupados puntualmente - Variable cuantitativa discreta

Tabla2. Distribución de frecuencias de la cant. de transporte público utilizado (x)

	x	f	F	fr	Fr
1	0	32	32	0.41	0.41
2	1	14	46	0.18	0.59
3	2	29	75	0.37	0.96
4	3	3	78	0.04	1.00

Gráfico 2. Diagrama de bastones que representa a la tabla 2.



Referencias:

f : frecuencia absoluta.

fr : frecuencia relativa.

F : frecuencia acumulada.

Fr : frecuencia acumulada relativa.

Datos agrupados puntualmente

Ejercicios:

1. ¿Qué cantidad de estudiantes tiene que utilizar por lo menos un transporte público?
2. ¿Qué porcentaje utiliza a lo sumo uno?
3. ¿Qué proporción utliza entre 1 y 3?
4. ¿Qué cantidad utiliza uno o más de dos?
5. De los que utilizan menos de tres, ¿qué porcentaje utiliza por lo menos uno?
6. Calcular el promedio, desvío estándar y mediana de la variable bajo estudio. Interpretar los resultados. (Los cálculos de promedio y desvío se presentan a continuación)

Medidas descriptivas para datos agrupados puntualmente

- **Promedio o media muestral**

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

Donde:

$\mathbf{x_i}$: valores de la variable, con $i = 1, 2, ..., m$.

$\mathbf{f_i}$: frecuencia absoluta correspondiente al valor x_i, con $i = 1, 2, ..., m$.

n: tamaño de la muestra

m: tamaño del dominio de la variable.

Medidas descriptivas para datos agrupados puntualmente

- **Varianza**

$$s^2_{n-1} = \frac{\sum_{i=1}^{m}(x_i - \overline{x}) * f_i}{n-1}$$

- **Desvío estándar**

$$s_{n-1} = \sqrt{s^2_{n-1}}$$

Las referencias son similares a las de promedio presentadas anteriormente.

Tabla 3. Distribución de frecuencias de las alturas de los estudiantes

Datos agrupados en intervalos - Tabla de frecuencias

- Cada dato del conjunto será ubicado en el intervalo que corresponda.
- Hay diferentes maneras para determinar la cantidad de intervalos necesarios de utilizar. Una de ellas es la **regla de Sturges**:

$$\text{cant. de intervalos: } k = 1 + log_2(n)$$

De todos modos, uno podría decidir la cantidad de intervalos a utilizar, y según la distribución obtenida modificarlos para un mejor análisis.

Datos agrupados en intervalos - Tabla de frecuencias

- El ancho del intervalo (a) se calcula de la siguiente manera:

$$a = \frac{x_{Máx} - x_{min}}{k} = \frac{rango}{k}$$

En este caso, todos los intervalos serán de igual logitud, pero no necesariamente debe ser así, los intervalos pueden ser de longitudes diferentes.

No es una condición estricta que el límite inferior del primer intervalo comience en el mínimo de los datos ni que el límite superior del último intervalo termine en el máximo, pero se debe construir la tabla de tal manera que todos los datos estén contenidos en algún intervalo.

Datos agrupados en intervalos

De la muestra realizada y presentada anteriormente, también se obtuvo como información las alturas de las/os 78 estudiantes, cuyos resultados son los siguientes (se muestran los 8 primeros datos):

```
## [1] 1.73 1.60 1.73 1.65 1.78 1.77 1.69 1.71
```

- **Variable:** altura de un estudiante de la UTN.
- **Tipo de variable:** cuantitativa continua.
- **Población:** estudiantes de la UTN Regional Haedo.

Datos ordenados en intervalos

Para obtener una mejor interpretación de los datos se agrupan en intervalos de clases.

	x (altura)	f	Pm	fr	F	Fr
1	(1.5,1.54]	1	1.52	0.01	1	0.01
2	(1.54,1.58]	2	1.56	0.03	3	0.04
3	(1.58,1.62]	6	1.60	0.08	9	0.12
4	(1.62,1.66]	7	1.64	0.09	16	0.21
5	(1.66,1.7]	6	1.68	0.08	22	0.28
6	(1.7,1.74]	16	1.72	0.21	38	0.49
7	(1.74,1.78]	19	1.76	0.24	57	0.73
8	(1.78,1.82]	11	1.80	0.14	68	0.87
9	(1.82,1.86]	6	1.84	0.08	74	0.95
10	(1.86,1.9]	4	1.88	0.05	78	1.00

Tabla 3. Distribución de frecuencias de las alturas de los estudiantes

Datos agrupados en intervalos

Referencias:

Pm: Punto medio del intervalo.

f: frecuencia absoluta.

fr: frecuencia relativa.

F: frecuencia acumulada.

Fr: frecuencia acumulada relativa.

Histograma

Para representar una tabla de frecuencias y observar el comportamiento de los datos de la muestra, se puede utilizar un **histograma**.

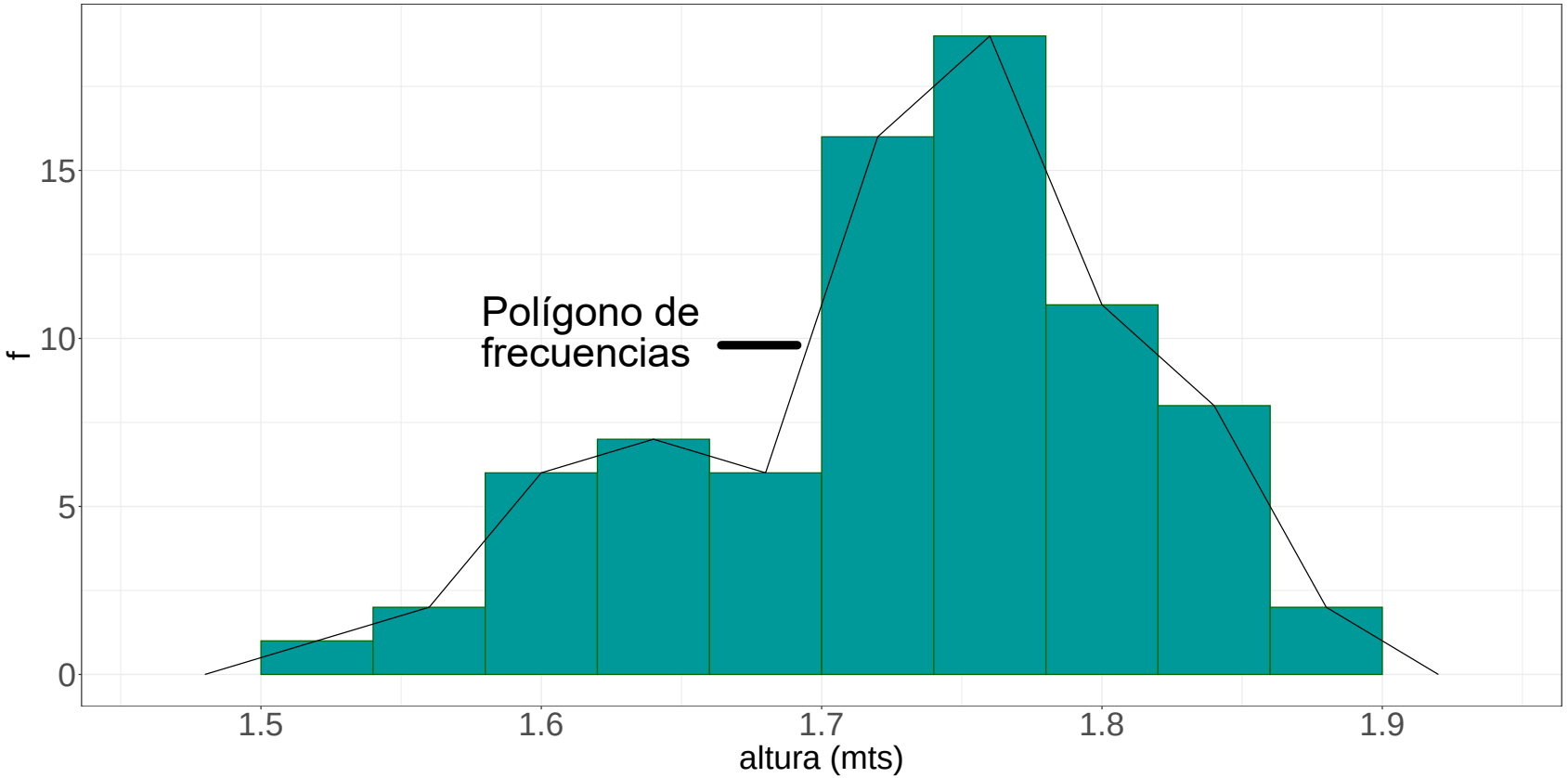
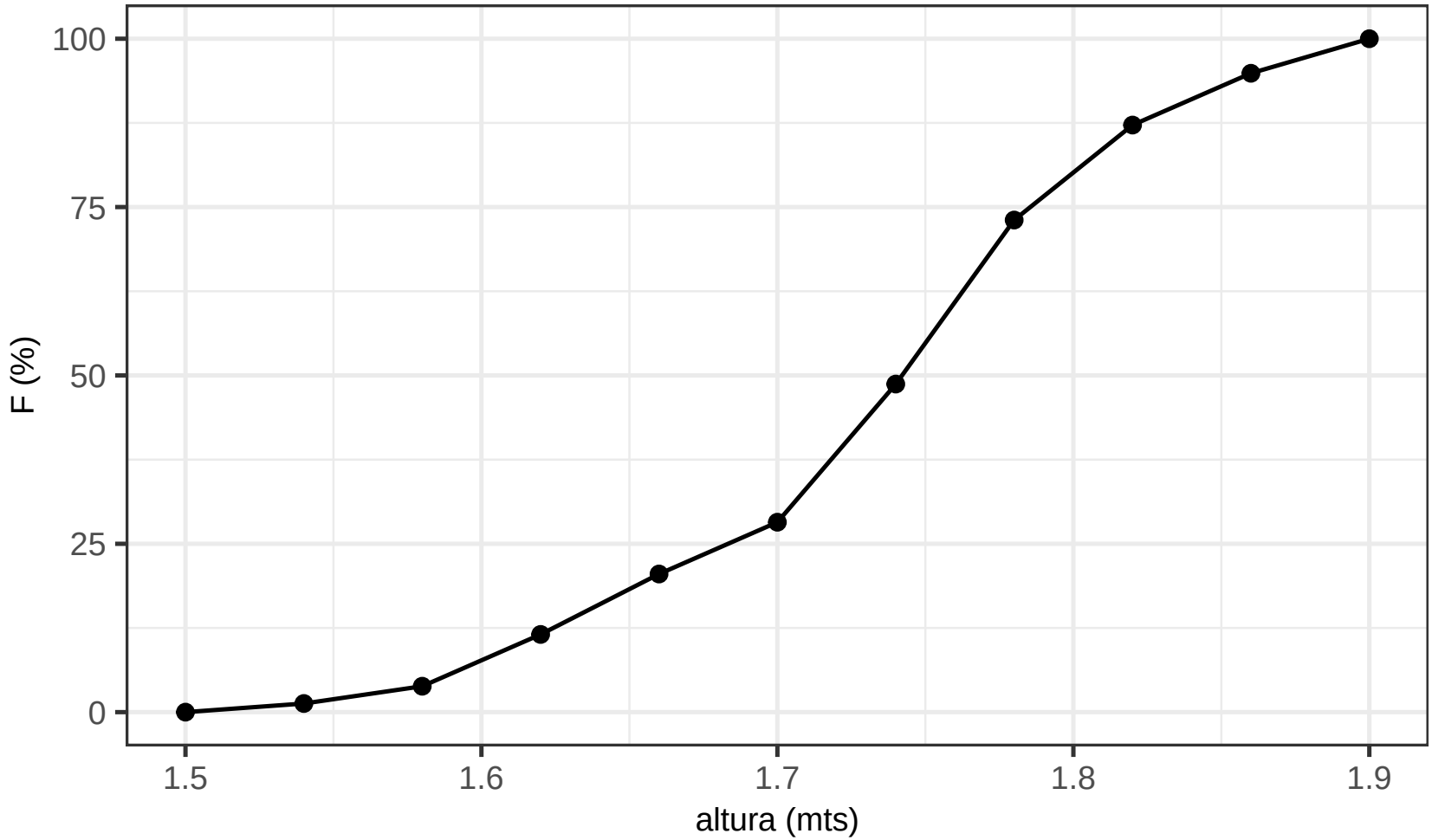


Gráfico 3. Histograma que representa a la tabla 3.

Gráfico de frecuencias acumuladas u Ojiva



Una de las utilidades de este gráfico es poder observar las diferentes medidas de posición, como por ejemplo la mediana.

Medidas descriptivas para datos agrupados en intervalos

- **Promedio o media muestral**

$$\overline{x} = \frac{\sum_{i=1}^{k} Pm_i * f_i}{n}$$

Donde:

Pm : es el punto medio de cada intervalo intervalo.

f : la frecuencia absoluta de cada intervalo.

n : tamaño de la muestra

k : cantidad de intervalos

Medidas descriptivas para datos agrupados en intervalos

- **Moda o Modo**

1. Se busca el intervalo modal, que será el de mayor frecuencia.
2. Se calcula la Moda de la siguiente manera:

$$Mo = l.inf + \frac{\triangle_1}{\triangle_1 + \triangle_2} * a_i$$

Donde:

l.inf = límite inferior del intervalo que contiene la moda (intervalo modal).

$\triangle_1 = f_{mo} - f_{mo-1}$,

$\triangle_2 = f_{mo} - f_{mo+1}$

$\mathbf{a_{mo}}$: amplitud del intervalo que contiene a la moda.

f_{mo} : frecuencia absoluta del intervalo modal.

Tabla 3. Distribución de frecuencias de las altura de los estudiantes

f_{mo-1} frecuencia absoluta del intervalo anterior al intervalo modal.

f_{mo+1} frecuencia absoluta del intervalo posterior al intervalo modal.

Medidas descriptivas para datos agrupados en intervalos

- **Mediana**

Para calcular la mediana de una conjunto de datos representados en una tabla de frecuencias, se procede de la siguiente manera:

1. Se calcula la posición de la mediana en el conjunto de datos ordenados: $\frac{n}{2}$
2. Se ubica al intervalo que contiene a la mediana.
3. Se interpola dentro del intervalo de la siguiente manera:

$$\widetilde{\mathbf{x}} = \mathbf{l.inf} + \frac{\frac{\mathbf{n}}{\mathbf{2}} - \mathbf{F_{i-1}}}{\mathbf{f_i}} * \mathbf{a_i}$$

Donde:

l.inf : límite inferior del int. que contiene a la mediana.

$\mathbf{F_{i-1}}$: frecuencia acumulada hasta el intervalo anterior.

$\mathbf{f_i}$: frecuencia absoluta del intervalo que contiene a la mediana.

$\mathbf{a_i}$: amplitud del intervalo.

Medidas descriptivas para datos agrupados en intervalos

- **Percentil**

El cálculo de un percentil para datos agrupados es similar al de la mediana, solamente varía en la posición.

1. Se calcula la posición del percentil k en el conjunto de datos: $\frac{k*n}{100}$
2. Se ubica al intervalo que contiene al percentil.
3. Se interpola dentro del intervalo de la siguiente manera:

$$P_k = l.inf. + \frac{\frac{k*n}{100} - F_{i-1}}{f_i} * a_i$$

Donde

l.inf. : límite inferior del intervalor que contiene al percentil.

F_{i-1} : frecuencia acumulada hasta el intervalo anterior.

f_i : frecuencia absoluta del intervalo que contiene a la mediana.

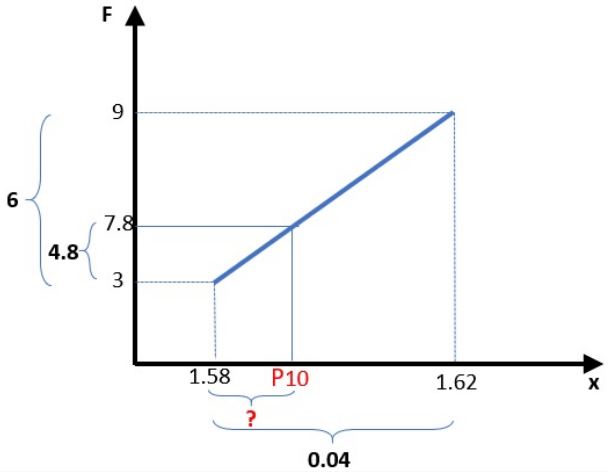
a_i : amplitud del intervalo.

Cálculo de alguna medida de posición - Ejemplo

A continuación se muestra, a modo de ejemplo, como calcular el P_{10} de la tabla 3 (alturas agrupadas), mediante la interpolación en el intervalo que lo contiene.

	x (altura)	f	Pm	fr	F	Fr
1	(1.5,1.54]	1	1.52	0.01	1	0.01
2	(1.54,1.58]	2	1.56	0.03	3	0.04
3	(1.58,1.62]	6	1.60	0.08	9	0.12
4	(1.62,1.66]	7	1.64	0.09	16	0.21

Observamos que el P_{10} se encuentra en el 3° intervalo, debido que la posición que ocupa en el conjunto de 78 datos es 7.8, esto se obtiene a partir de $\frac{\mathbf{10*78}}{\mathbf{100}}$.



Por regla de Thales:

$$\frac{0.04}{6} = \frac{a}{4.8} \Rightarrow a = 0.032$$

$$P_{10} = 1.58 + 0.032 = \mathbf{1.612}$$

(La fórmula de percentil realiza este mismo cálculo)

Medidas descriptivas para datos agrupados en intervalos

- **Varianza**

$$s^2 = \frac{\sum_{i=1}^{m}(Pm_i - \overline{x}) * f_i}{n-1}$$

Donde

Pm_i : punto medio del intervalo.

f_i : frecuencia absoluta del intervalo.

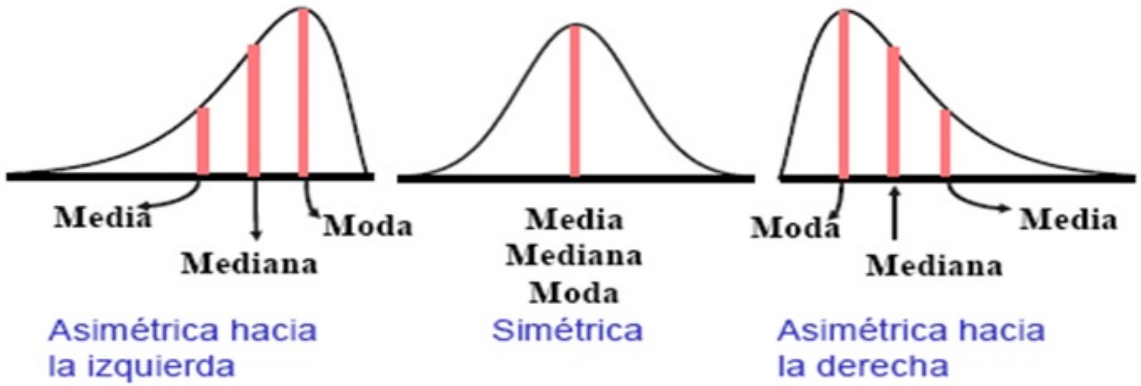
- **Desvío estándar**

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

Coeficiente de asimetría

Es una medida de forma que indica cómo se disrtibuyen los datos.

$$g = \frac{\frac{\sum_{i=1}^{m}(Pm_i - \bar{x})^3}{n}}{S^3} * f_i$$

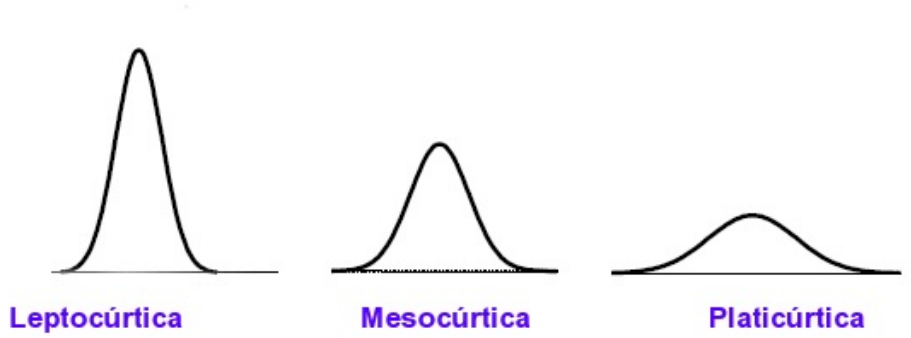


- Si $g > 0 \rightarrow$ Asimetría positiva o a derecha.
- Si $g = 0 \rightarrow$ Distribución simétrica.
- Si $g < 0 \rightarrow$ Asimetría negativa o a izquierda.

Coeficiente de curtosis

Es una medida de forma que indica qué tan puntiaguda es la distribución de los datos.

$$q = \frac{\frac{\sum_{i=1}^{m}(Pm_i - \bar{x})^4}{n}}{S^4} * f_i$$



- Si $q > 3 \rightarrow$ Leptocúrtica.
- Si $q = 3 \rightarrow$ Mesocúrtica.
- Si $q < 3 \rightarrow$ Platicúrtica.