Estadistica Descriptiva

Marcelo Paz

Estadistica y Probabilidades

18 de diciembre de 2023

1. Definiciones generales

1. **Estadística:** Ciencia que estudia los métodos para recoger, organizar, resumir y analizar datos, así como para sacar conclusiones válidas y tomar decisiones razonables basadas en tal análisis.

2. **Estadística Descriptiva:** Parte de la estadística que se ocupa de la recolección, presentación, descripción, análisis y resumen de datos.

2. Tipos de variables

1. **Variable cualitativa:** Es aquella que no se puede medir numéricamente, sino que se clasifica en categorías.

2. **Variable cuantitativa:** Es aquella que se puede medir numéricamente.

   a) **Variable discreta:** Es aquella que puede tomar valores aislados, es decir, no puede tomar todos los valores posibles entre dos valores cualesquiera.

   b) **Variable continua:** Es aquella que puede tomar todos los valores posibles entre dos valores cualesquiera.

3. Graficos

3.1. Tabla de frecuencias:

Es una tabla que representa los datos de forma ordenada en columnas.

Signos Visibles	**Frecuencia Absoluta** f_i	**Frecuencia Acumulada** F_i	**Frecuencia relativa porcentual** fr_i
Dieta Severa	9	9	33 %
Uso de Ropa Holgada	6	15	22 %
Miedo a Engordar	3	18	11 %
Hiperactividad	4	22	15 %
Uso de Laxantes	5	27	19 %
Total	27		100 %

Figura 1: Tabla de frecuencias de datos **cualitativos**

Número de asignaturas reprobadas c_i	**Frecuencia Absoluta** f_i	**Frecuencia Acumulada** F_i	**Frecuencia Relativa porcentual** fr_i
0	26	26	28,8 %
1	17	43	18,8 %
2	21	64	23,3 %
3	11	75	12,2 %
4	7	82	7,7 %
5	4	86	4,4 %
6	3	89	3,3 %
7	1	90	1,1 %
Total	90		100 %

Figura 2: Tabla de frecuencias de datos **cuantitativos discretos**

Fronteras	**Frecuencia Absoluta** f_i	**Frecuencia Acumulada** F_i	**Frecuencia relativa porcentual** fr_i	**Marca de clase** m_i
1,5 - 2,5	5	5	10 %	2
2,5 - 3,5	14	19	28 %	3
3,5 - 4,5	6	25	12 %	4
4,5 - 5,5	25	50	50 %	5
Total	50		100 %	

Figura 3: Tabla de frecuencias de datos **cuantitativos continuos**

3.2. Histograma:

Es un gráfico de barras. Se construye ubicando en el eje horizontal a las fronteras y en el vertical a las frecuencias absolutas. Su particularidad es que las barras están pegadas, pues comparten un lado en común. Su utilidad se aprecia cuando se quiere estudiar la forma de la distribución. Esto es, cuando se quiere estudiar la simetría o sesgo de los datos

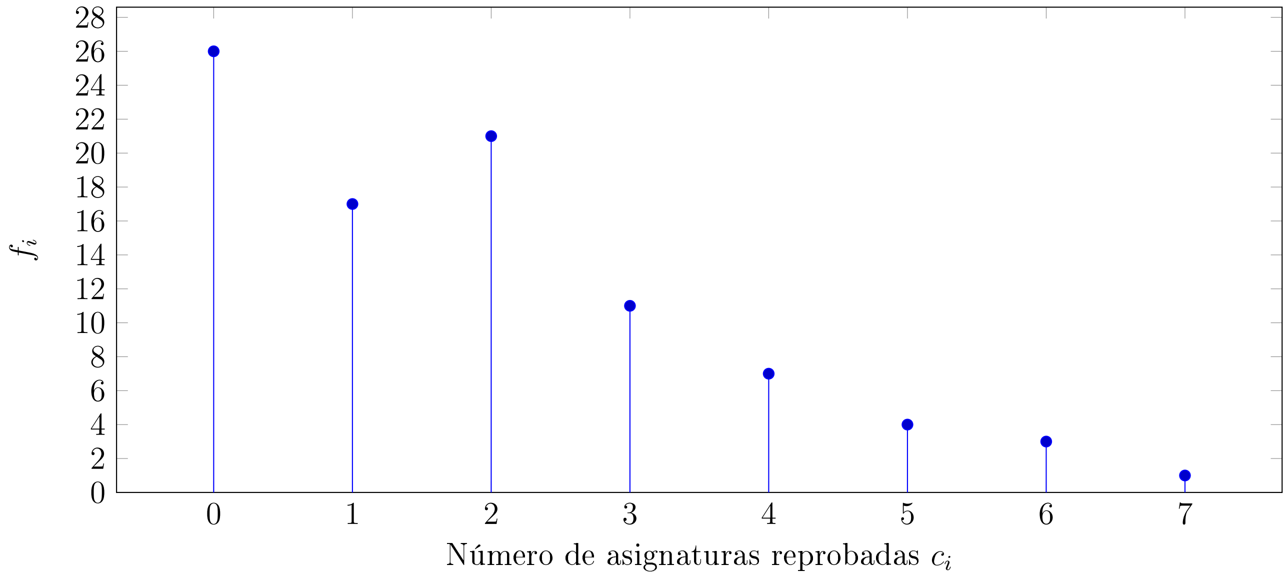

Figura 4: Histograma de datos **cuantitativos discretos**

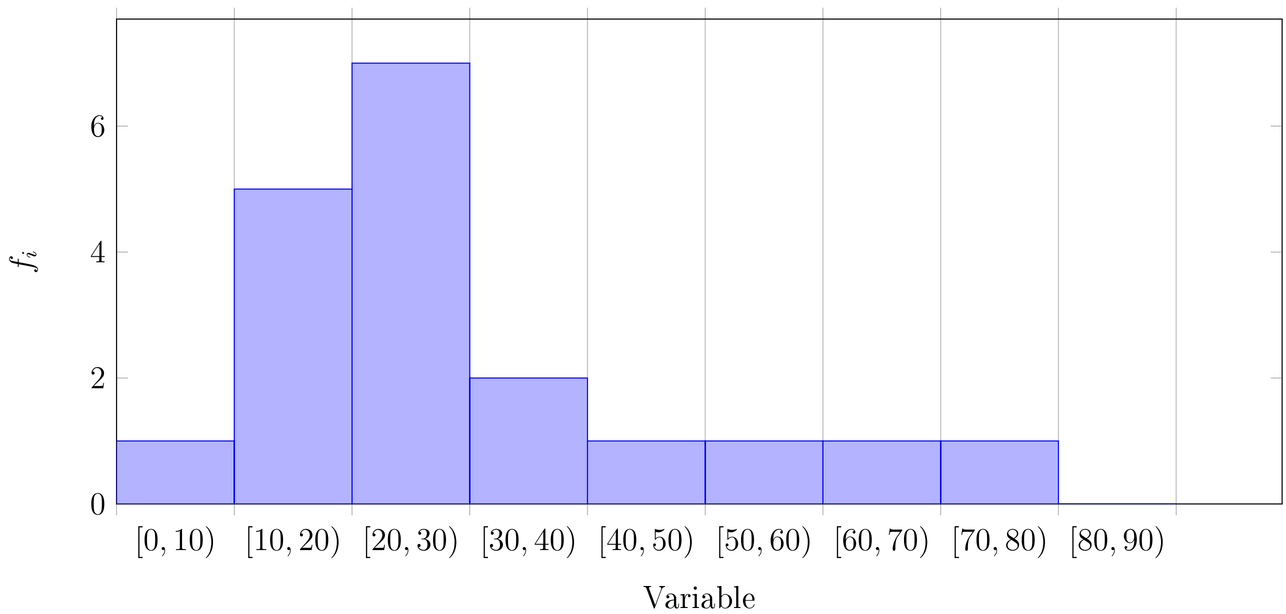

Figura 5: Histograma de datos **cuantitativos continuos**

3.3. Polígono de frecuencia:

Es un gráfico que consiste en destacar las marcas de clase de cada intervalo frente a sus correspondientes frecuencias absolutas. Cada marca de clase se une con segmentos de rectas que generan la curva. Para cerrar el área se utilizan las marcas de clase ficticias: La primera se crea restando la amplitud a la marca de clase del primer intervalo y la segunda se crea sumando la amplitud a la marca de clase del último intervalo. Su utilidad se aprecia cuando se quiere comprar dos o más distribuciones de frecuencias. Generalmente se dibuja sobre el Histograma.

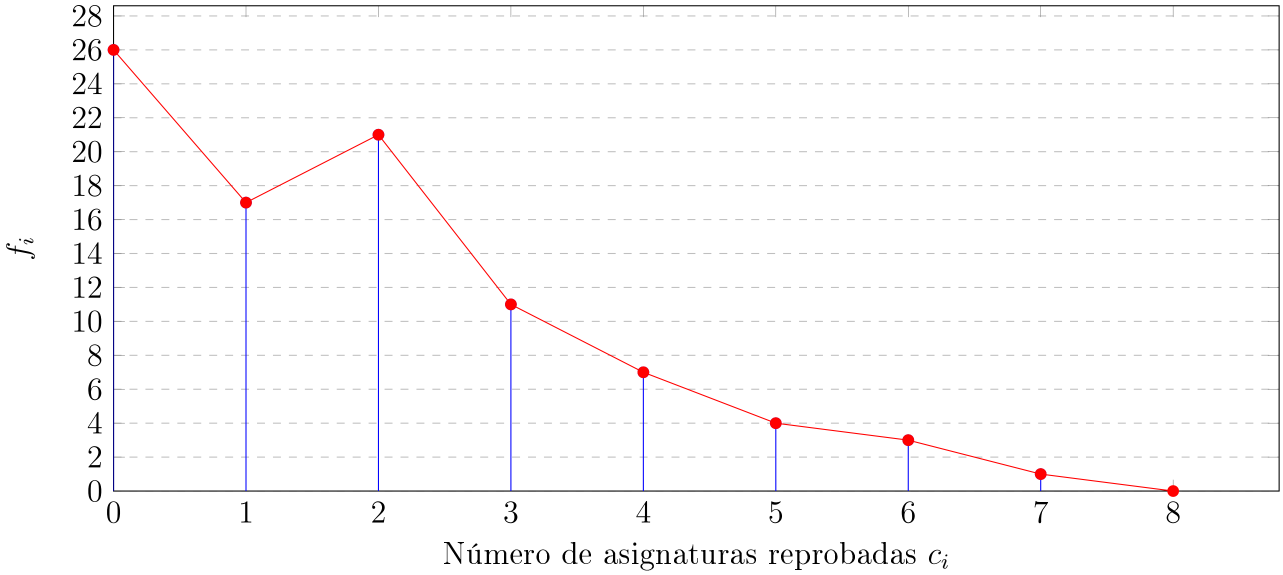

Figura 6: Histograma y Poligono de frecuencias de datos **cuantitativos discretos**

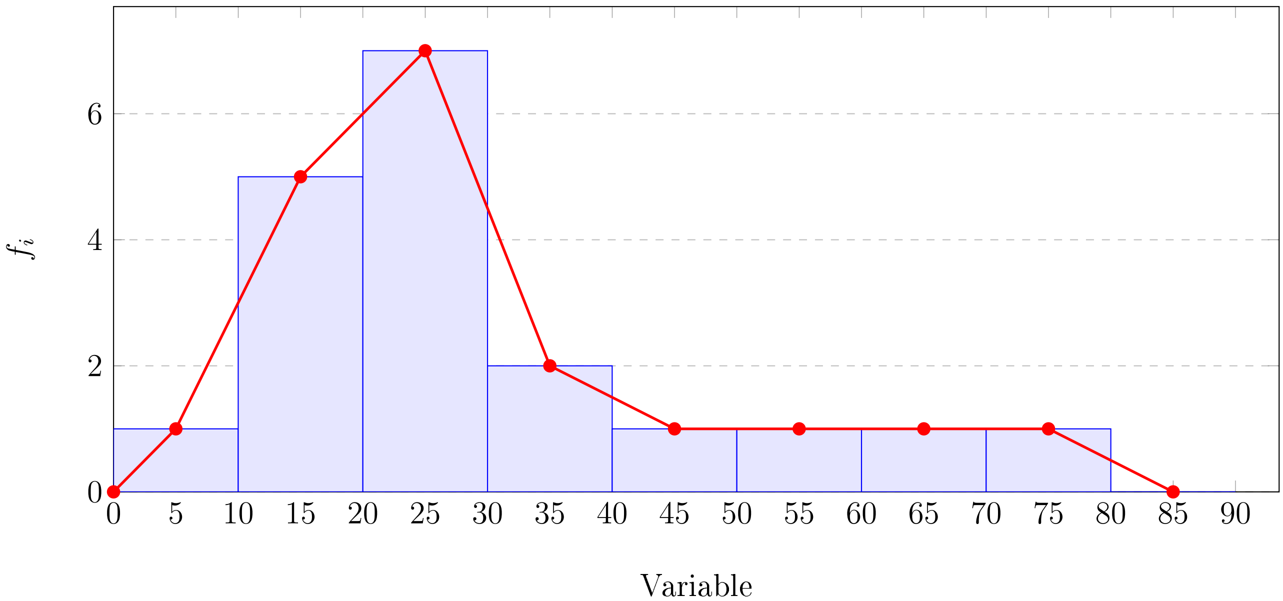

Figura 7: Histograma y Poligono de frecuencia de datos **cuantitativos continuos**

3.4. Ojiva:

Es un gráfico de frecuencia acumulada. Se gráfica única y exclusivamente con las fronteras en el eje horizontal. Su utilidad se aprecia cuando se cuenta con medidas de posición tales como: Cuartiles, Deciles y Percentiles.

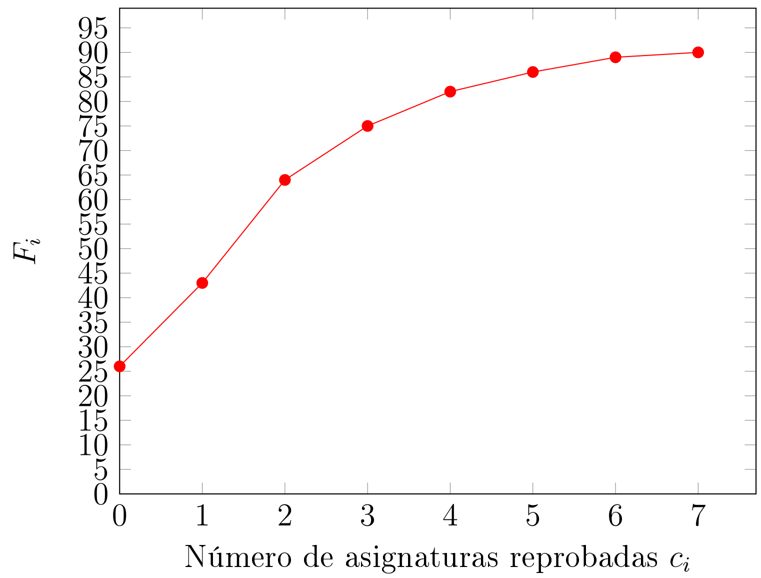

Figura 8: Ojiva de datos **cuantitativos discretos**

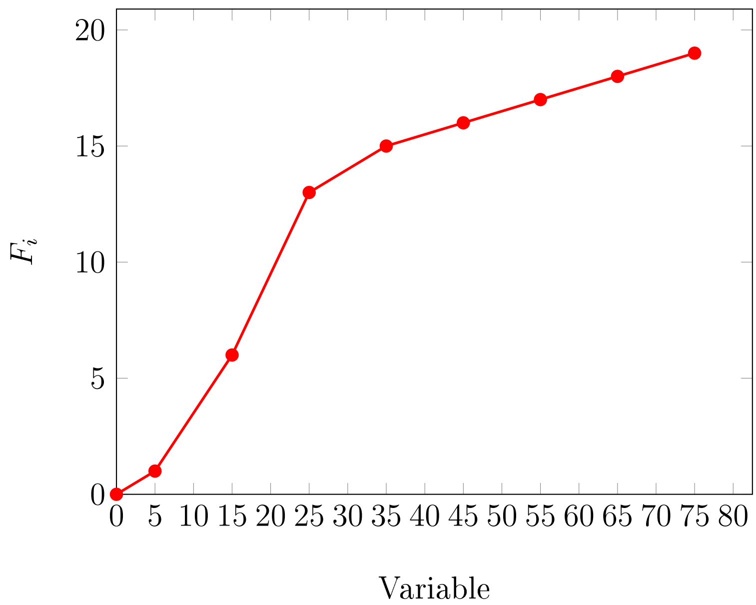

Figura 9: Ojiva de datos **cuantitativos continuos**

4. Medidas de tendencia central

4.1. Media aritmética:

Es la suma de todos los datos dividida por el total. Cambia un poco en su forma según como estén presentados los datos.

1. **Datos no agrupados:**

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

   * x_i es cada dato.

2. **Datos agrupados(discreta):**

$$\overline{x} = \frac{\sum_{i=1}^{k} f_i \cdot c_i}{n}$$

   * c_i es cada dato.

3. **Datos agrupados(continua):**

$$\overline{x} = \frac{\sum_{i=1}^{k} f_i \cdot m_i}{n}$$

   * m_i es la marca de clase de cada dato.

* f_i es la frecuencia absoluta de cada dato.
* n es el total de datos.
* k es el total de intervalos.

4.2. Mediana:

Es el valor que ocupa la posición central en un conjunto de datos ordenados. Cambia un poco en su forma según como estén presentados los datos.

1. **Datos no agrupados:** es necesario ordenar de menor a mayor los datos.

$$M_e = \begin{cases} x_{\frac{n+1}{2}} & \text{si } n \text{ es impar} \\ \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2} & \text{si } n \text{ es par} \end{cases}$$

2. **Datos agrupados(discreta):** M_e Es el valor de la clase o categoria (c_i) donde se encuentra la mitad de los datos en la columna de la frecuencia acumulada.

3. **Datos agrupados(continua):**

$$M_e = FI_k + \left(\frac{\frac{n}{2} - F_{k-1}}{f_k} \right) \cdot A_k$$

* FI_k es la Frontera Inferior de la clase mediana.
* F_{k-1} es la Frecuencia absoluta acumulada de la clase anterior a la clase mediana.
* f_k es la frecuencia absoluta de la clase.
* A_k es la amplitud de la clase mediana.
* n es el total de datos.

4.3. Moda:

Es el valor que más se repite en un conjunto de datos. Cambia un poco en su forma según como estén presentados los datos.

1. **Datos no agrupados:** es necesario ordenar de menor a mayor los datos.

$$M_o = \text{Valor que más se repite}$$

2. **Datos agrupados(discreta):** M_o Es el valor de la clase o categoria (c_i) donde se encuentra la mayor frecuencia absoluta.

3. **Datos agrupados(continua):**

$$M_o = FI_k + \left(\frac{a}{a+b} \right) \cdot A_k$$

* FI_k es la Frontera Inferior de la clase modal.
* a es la Diferencia entre la frecuencia absoluta de la clase modal y la frecuencia absoluta de la clase anterior a la clase modal.
* b es la Diferencia entre la frecuencia absoluta de la clase modal y la frecuencia absoluta de la clase posterior a la clase modal.
* A_k es la amplitud de la clase modal.

4.4. Resumen

Descripción	**Media** $\overline{x}$	**Mediana** M_e	**Moda** M_o
Datos NO agrupados	$\overline{x} = \dfrac{\sum_{i=1}^{n} x_i}{n}$	Si el total de observaciones es impar, la mediana es el valor que se encuentra justo en la mitad del conjunto previamente ordenado de menor a mayor. Si el total de observaciones es par, la mediana es el promedio de las dos observaciones centrales del conjunto previamente ordenado.	Valor que más se repite. **Unimodal:** una moda. **Bimodal:** dos modas. **Multimodal:** mas de 2 modas
Datos agrupados (discreta)	$\overline{x} = \dfrac{\sum_{i=1}^{n} f_i c_i}{n}$	Es el valor de la clase o categoria (c_i) donde se encuentra la mitad de los datos en la columna de la frecuencia acumulada.	Es el valor de la clase o categoria (c_i) donde se encuentra la frecuencia absoluta mas alta.
Datos agrupados (continua)	$\overline{x} = \dfrac{\sum_{i=1}^{n} f_i m_i}{n}$	$M_e = FI_k + \left(\dfrac{\frac{n}{2} - F_{k-1}}{f_k} \right) \cdot A_k$	$M_o = FI_k + \left(\frac{a}{a+b} \right) \cdot A_k$

Figura 10: Tabla Resumen **Medidas de Tendencia Central**

5. Medidas de dispersión

Nos indican que tanto se alejan los datos del centro. Las más utilizadas en Estadística Descriptiva son:

5.1. Varianza

Es el promedio cuadrado de las distancias entre cada observación y el promedio de ellos. Se denota por S^2 . Su gran desventaja es que crece conforme crecen los datos y también puede ser cero si estos, son muy parecidos entre si.

Se utiliza la media $\overline{x}$

1. **Datos no agrupados:**

$$s^2 = \frac{\displaystyle\sum_{i=1}^{n}(x_i - \overline{x})^2}{n}$$

* x_i es cada dato.

2. **Datos agrupados (discreta):**

$$s^2 = \frac{\displaystyle\sum_{i=1}^{n}(c_i - \overline{x})^2 f_i}{n}$$

* c_i es cada dato.

3. **Datos agrupados (continua):**

$$s^2 = \frac{\displaystyle\sum_{i=1}^{n}(m_i - \overline{x})^2 f_i}{n}$$

* m_i es la marca de clase de cada dato.

* f_i es la frecuencia absoluta de cada dato.
* n es el total de datos.

5.2. Desviación Estandar

Es la raíz cuadrada de la varianza. Su gran ventaja por sobre ésta, es que entrega sus resultados en la misma unidad de medida que la variable.

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

* s^2 es la varianza.

5.3. Coeficiente de Variación

Se define como el cuociente entre la desviación estándar y el promedio de los datos. Generalmente se entrega en porcentaje. Su gran ventaja es que sus resultados carecen de unidad de medida, por lo que permite comparar datos aunque estén en distinta unidades de medida.

$$CV = \frac{s}{\overline{x}} \cdot 100$$

* s es la desviación estándar.
* $\overline{x}$ es la media.

5.4. Rango

Es la diferencia entre el máximo y el mínimo de los datos. Su utilidad se aprecia cuando tenemos más información de la variable.

$$R = x_{max} - x_{min}$$

* x_{max} es el valor máximo de los datos.
* x_{min} es el valor mínimo de los datos.

6. Medidas de posición

Son aquellas que permiten conocer con mayor detalle a una variable. Entre se encuentran los:

6.1. Cuartiles

Dividen al conjunto de datos en cuatro partes porcentualmente iguales.

$$Q_k \quad , \text{k} = 1,2,3 \quad , \text{donde: } Q_1 = 25\,\% \quad Q_2 = 50\,\% \quad Q_3 = 75\,\%$$

1. Para datos cuantitativos discretos:

$$Q_k = \frac{kn}{4}$$

2. Para datos cuantitativos continuos:

$$Q_k = FI_k + \left(\frac{\frac{kn}{4} - F_{k-1}}{f_k} \right) \cdot A_k$$

* FI_k es la Frontera Inferior de la clase del k-esimo cuartil.
* F_{k-1} es la frecuencia acumulada hasta la clase anterior a la clase del k-esimo cuartil.
* f_k es la frecuencia absoluta de la clase del k-esimo cuartil.
* A_k es la amplitud de la clase del k-esimo cuartil.

6.2. Deciles

Dividen al conjunto de datos en diez partes porcentualmente iguales.

$$D_k \quad , \text{k} = 1,2,...,8,9 \quad , \text{donde: } D_1 = 10\,\% \quad ... \quad D_9 = 90\,\%$$

1. Para datos cuantitativos discretos:

$$D_k = \frac{kn}{10}$$

2. Para datos cuantitativos continuos:

$$D_k = FI_k + \left(\frac{\frac{kn}{10} - F_{k-1}}{f_k} \right) \cdot A_k$$

* FI_k es la Frontera Inferior de la clase del k-esimo decil.
* F_{k-1} es la frecuencia acumulada hasta la clase anterior a la clase del k-esimo decil.
* f_k es la frecuencia absoluta de la clase del k-esimo decil.
* A_k es la amplitud de la clase del k-esimo decil.

6.3. Percentiles

Dividen al conjunto de datos en cien partes porcentualmente iguales.

$$P_k \quad , \text{k} = 1,2,...,98,99 \quad , \text{donde: } P_1 = 1\,\% \quad ... \quad P_{99} = 99\,\%$$

1. Para datos cuantitativos discretos:

$$P_k = \frac{kn}{100}$$

2. Para datos cuantitativos continuos:

$$P_k = FI_k + \left(\frac{\frac{kn}{100} - F_{k-1}}{f_k} \right) \cdot A_k$$

* FI_k es la Frontera Inferior de la clase del k-esimo percentil.
* F_{k-1} es la frecuencia acumulada hasta la clase anterior a la clase del k-esimo percentil.
* f_k es la frecuencia absoluta de la clase del k-esimo percentil.
* A_k es la amplitud de la clase del k-esimo percentil.

7. Simetría

$$f_1 = f_k \quad f_2 = f_{k-1} \quad f_3 = f_{k-2} \quad \text{... etc}$$

7.1. Unimodal

$$\overline{x} = M_e = M_o$$

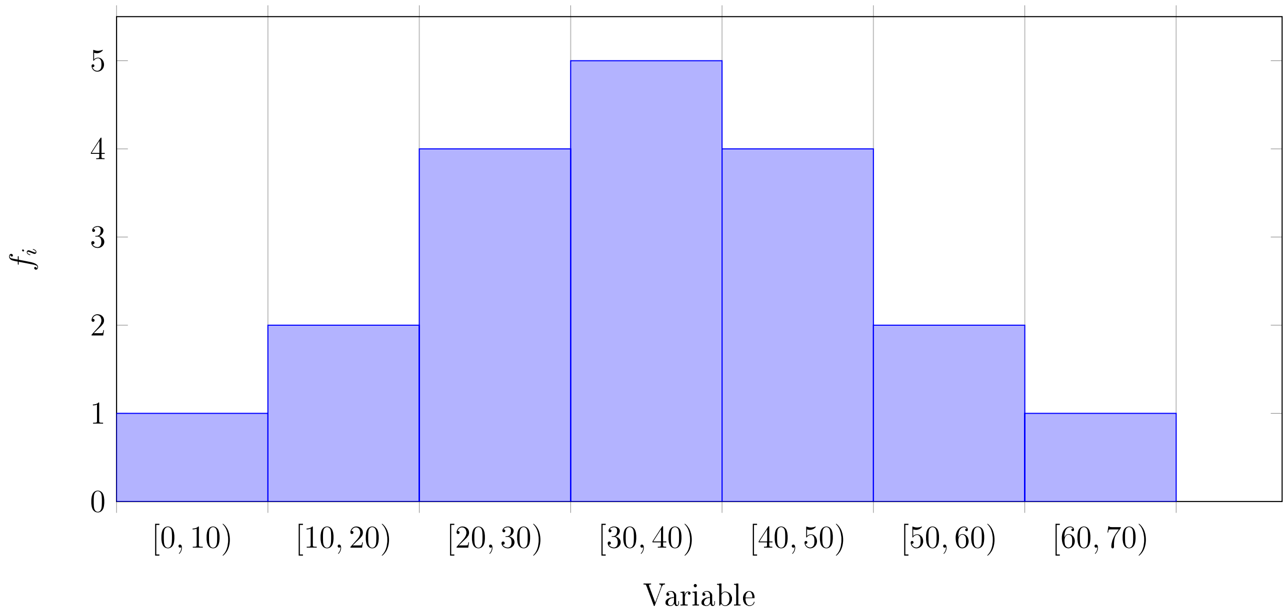

Figura 11: Distribución **simétrica y unimodal**

7.2. Bimodal

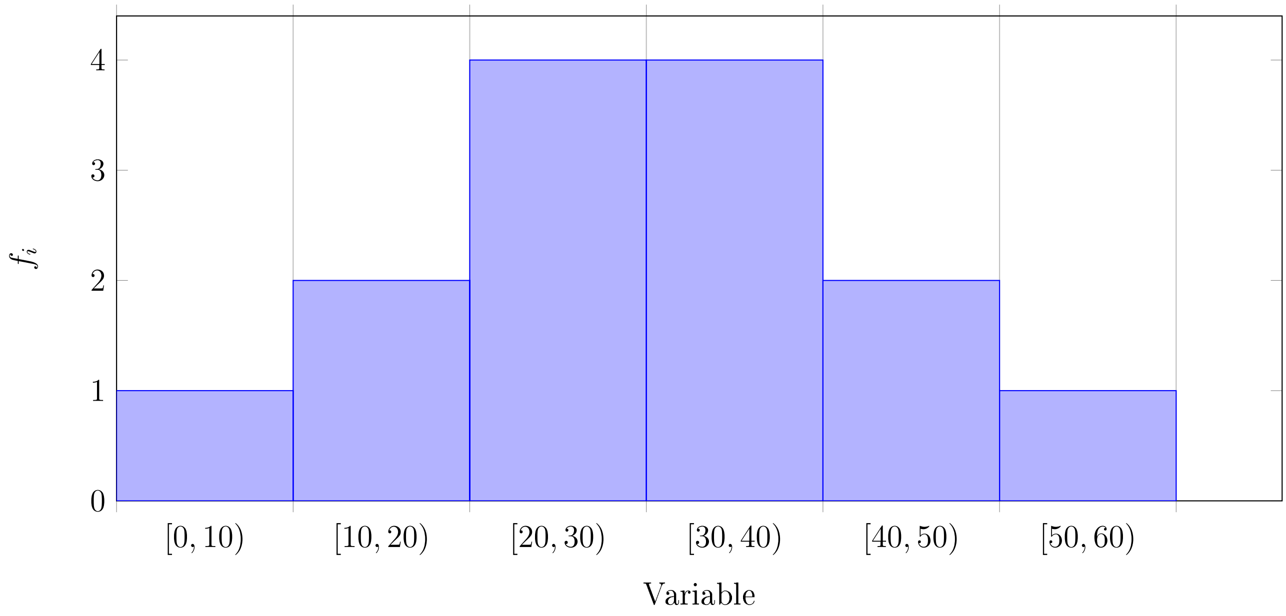

Figura 12: Distribución **simétrica y bimodal en el centro**

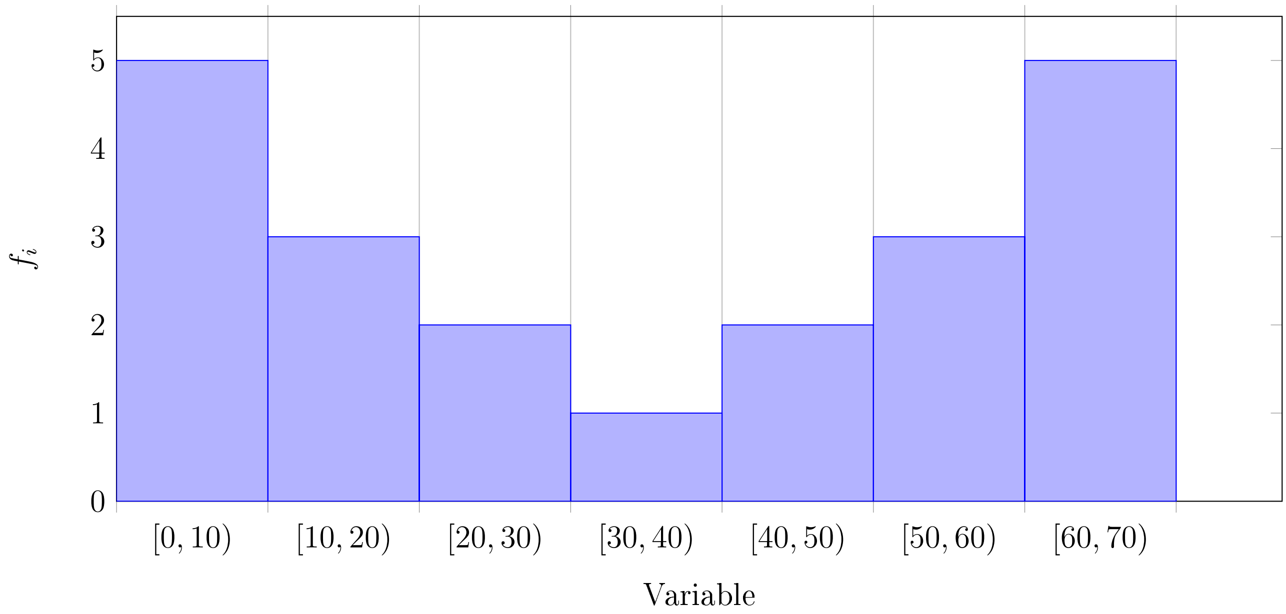

Figura 13: Distribución **simétrica y bimodal en los extremos**

8. Sesgo (Asimétria)

El sesgo es un comportamiento que se da en las medidas de tendencia central y es de la siguiente forma:

8.1. Positivo o a la derecha

$$\overline{x} > M_e > M_o$$

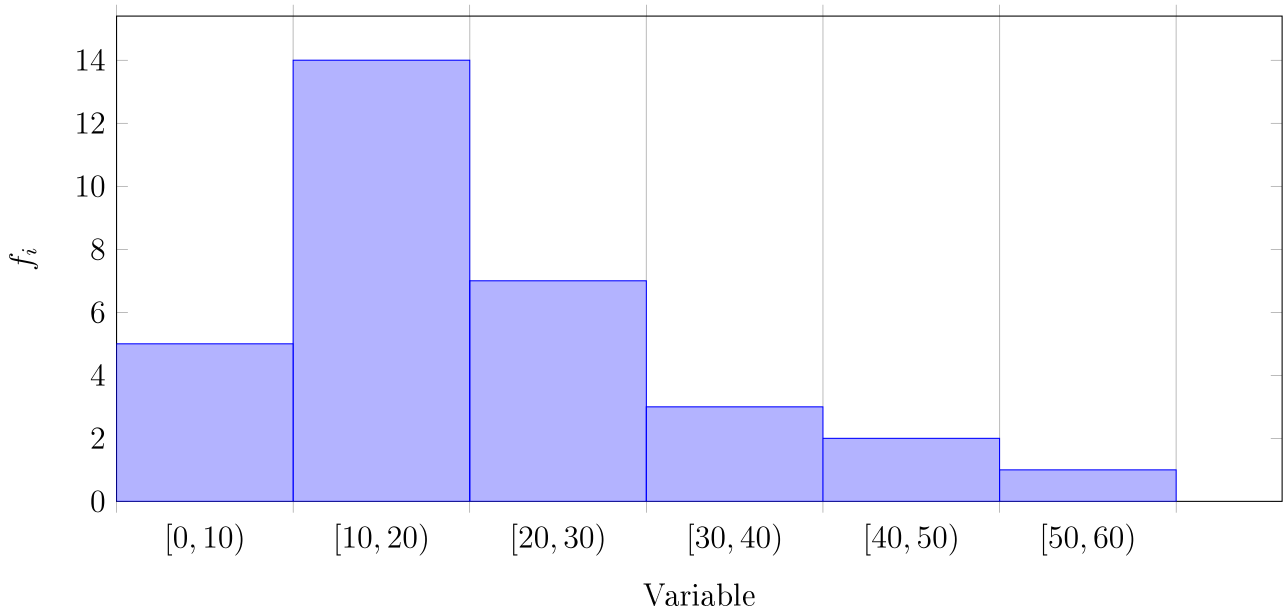

Figura 14: Distribución **asimétrica positiva**

8.2. Negativo o a la izquierda

$$\overline{x} < M_e < M_o$$

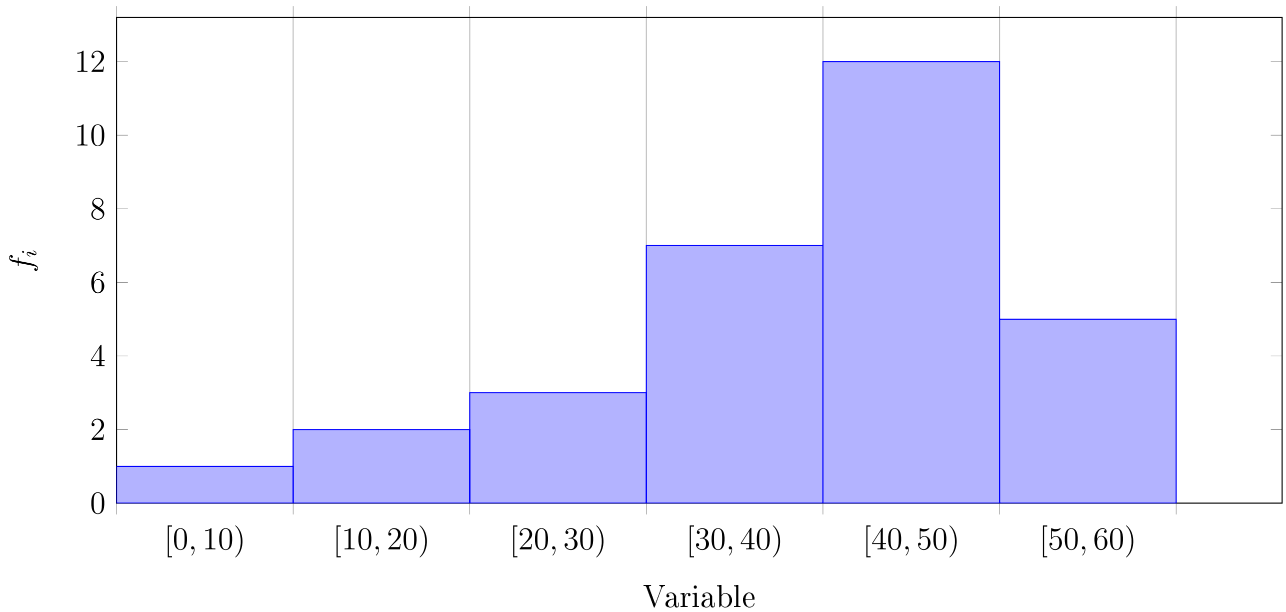

Figura 15: Distribución **asimétrica positiva**

8.3. Coeficientes de asimétria

$$CA_1 = \frac{\overline{x} - M_o}{s} \quad ; \quad CA_2 = \frac{3(\overline{x} - M_e)}{s} \quad ; \quad CA_3 = \frac{Q_1 - 2Q_2 + Q_3}{Q_3 - Q_1}$$

* donde $Q_3 - Q_1$ se conoce como rango intercuartílico.
* $CA_1 \neq CA_2 \neq CA_3$, pero coiciden en signo.

1. Si el **coeficiente es positivo**: se dice que la asimetría es positiva y el sesgo va a la derecha.
2. Si el **coeficiente es negativo**: se dice que la asimetría es negativo y el sesgo va a la izquierda.

9. Regla Empírica

Si una distribución es simétrica, unimodal de forma acampanada, se dice **Aproximadamente Normal**

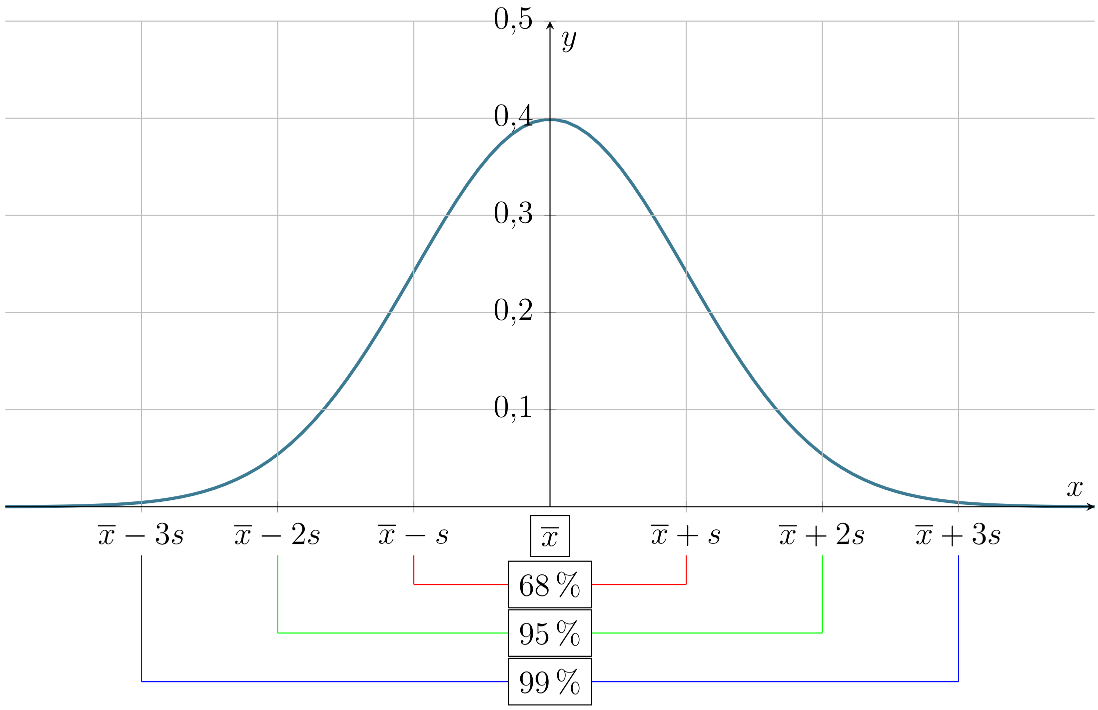

Figura 16: Regla Empírica

10. Teorema de Chebyshev

Si la distribución es asimétrica o tiene algún tipo de sesgo:

Dado un numero $k >= 1$ y un conjunto de n mediciones $x_1, x_2, ..., x_n$, por lo menos $\left(1 - \frac{1}{K^2}\right)$ % de las mediciones estara en $(\overline{x} - Ks, \overline{x} + Ks)$

$$\text{Si k=1} \quad \left(1 - \frac{1}{K^2}\right)\% = 0\,\%$$

$$\text{Si k=2} \quad \left(1 - \frac{1}{K^2}\right)\% = 75\,\%$$

$$\text{Si k=2,6} \quad \left(1 - \frac{1}{K^2}\right)\% = 85,21\,\%$$

$$\text{Si k=3} \quad \left(1 - \frac{1}{K^2}\right)\% = 88,9\,\%$$