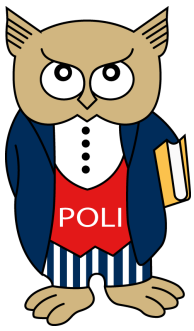
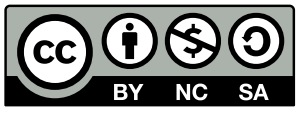


Probabilidad y Estadística - 2025-A

Sección 1: Estadística Descriptiva

Preparado por:

Cátedra de Probabilidad y Estadística - EPN



0. Índice general

1. Organización y Descripción de datos estadísticos

Resultados de aprendizaje

Conocimientos

1. Analizar la base conceptual de la estadística descriptiva.

Destrezas

1. Usar herramientas informáticas para el análisis estadístico de datos
2. Aplicar los conocimientos de técnicas y métodos de la Estadística Descriptiva.

Valores y aptitudes

1. Demostrar capacidad de auto aprendizaje
2. Demostrar responsabilidad en el cumplimiento de sus obligaciones

1.1. Introducción

La estadística resulta fundamental para conocer el comportamiento de ciertos eventos, por lo que ha adquirido un papel clave en la investigación. Se usa como un valioso auxiliar y en los diferentes campos del conocimiento y en las variadas ciencias. Es un lenguaje que permite comunicar información basada en datos cuantitativos.

Es tan importante que casi no existe actividad humana en que no esté involucrada la Estadística. Las decisiones más importantes de nuestra vida se toman con base en la aplicación de la Estadística. Pongamos algunos ejemplos.

La estadística es de gran importancia en la investigación científica debido a que:

- Permite una descripción más exacta.
- Nos obliga a ser claros y exactos en nuestros procedimientos y en nuestro pensar.
- Permite resumir los resultados de manera significativa y cómoda.

- Nos permite deducir conclusiones generales.

La evolución de la estadística ha llegado al punto en que su proyección se percibe en casi todas las áreas de trabajo. También abarca la recolección, presentación y caracterización de información para ayudar tanto en el análisis e interpretación de datos como en el proceso de la toma de decisiones. La estadística es parte esencial de la forma profesional, es hasta cierto punto una parte necesaria para toda profesión.

1.2. Definiciones Generales

1.2.1.- Estadística

El vocablo Estadística deriva etimológicamente del latín "status", que significa estado o situación, ya que en el principio los principales usos de la estadística provinieron de la motivación de los gobernantes y los imperios de conocer la extensión de sus dominios, riquezas y su población (conocer el estado de su gobierno). Imperios como el griego, egipcio y el romano realizaban censos de población cada cierto tiempo con los cuales recopilan grandes cantidades de datos relativos a su población, superficie, posesiones agrícolas y ganaderas, así como también de las riquezas de todos los territorios bajo su control, generalmente con carácter militar o fiscal para la recaudación de impuestos. Luego, con el paso del tiempo en los siglos XVIII y XIX científicos importantes de la época como Copérnico, Galileo, entre otros, contribuyen al desarrollo de lo que se conoce como el método científico y con la implantación del mismo la estadística se convierte en un factor protagonista de la investigación y generación de gran parte de todo el conocimiento que ahora tenemos. En el presente, una de las características fundamentales de la estadística es su transversalidad a través de todas las áreas de la ciencia, ya que su metodología es aplicable al estudio de muchas disciplinas tales como la física, economía, sociología, etc. La estadística aplicada adecuadamente nos ayuda a obtener conclusiones relevantes para el estudio de todo tipo de fenómenos observables que pueden ser medidos. Así pues, la estadística aparece, a lo largo de la historia como un poderoso instrumento utilizado por gobiernos e instituciones o como elemento auxiliar de las distintas ciencias, ayudando a estas a desentrañar las grandes preguntas que la curiosidad del ser humano siempre ha perseguido; es decir: qué variables intervienen en un fenómeno, qué leyes rigen el comportamiento de las mismas y qué relación de dependencia hay entre ellas.

Definición de Estadística

En la actualidad, podemos definir, en general, a la estadística como la ciencia que trata de la recopilación, organización, presentación, análisis e interpretación de datos que intervienen en un fenómeno, con el fin de realizar una adecuada descripción del mismo y así poder inferir, predecir resultados, comportamientos o tomar decisiones con respecto al fenómeno que se está investigando.

Ramas de la estadística

Dentro de la estadística se distinguen dos ramas fundamentales:

- **Estadística Descriptiva:** Es la parte de la estadística que incluye todos los métodos de recolec-

ción, organización, resumen y presentación de un conjunto de datos. Se trata principalmente de describir las características fundamentales de los datos y para ellos se suelen utilizar indicadores, gráficos y tablas; también se la conoce como el análisis exploratorio de datos.

- **Estadística Inferencial:** Es la parte de la estadística que incluye todos los métodos utilizados para poder hacer predicciones, generalizaciones y obtener conclusiones a partir de los datos obtenidos de la observación de algún fenómeno, usa como punto de partida el análisis descriptivo de las muestras con observaciones del fenómeno investigado y en base a los resultados de este análisis poder deducir aspectos generales del fenómeno en sí, mediante el uso de estos métodos.

Ejemplo de uso de la estadística

La estadística es ampliamente utilizada en muchas áreas de la ciencia, por ejemplo, en el análisis económico algunos ejemplos de su uso son:

- Elaboración de indicadores macroeconómicos.
- Predicciones acerca del comportamiento futuro de la demanda de productos o servicios.
- Organizar y presentar datos económicos como: la evolución de los precios, el PIB, etc.
- En la epidemiología, por ejemplo: para estudiar la distribución de las enfermedades y los posibles factores de riesgos asociados.
- En el área de la salud, por ejemplo: el uso de estadísticas sanitarias para saber la razón de la muerte de las personas o cuales son las causas de enfermedades y traumatismos. Para abordar de mejor manera los problemas de salud y priorizar el uso de recursos sanitarios muy valiosos. Para conocer las problemáticas de salubridad presentes en una comunidad, los factores de riesgo o predisposición a ciertas patologías y en la búsqueda de las respuestas a las mismas.

1.2.2.- Definiciones generales

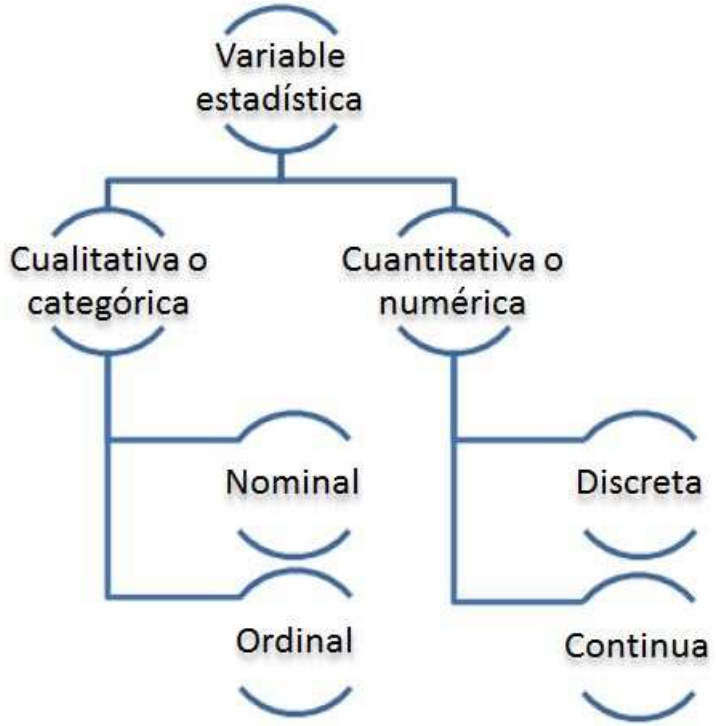
- **Población:** Conjunto o colección de objetos al que está referido un estudio estadístico. Puede estar constituida por cualquier tipo de elemento, es decir, por personas, pero también por objetos de cualquier tipo de naturaleza.
- **Muestra:** Cualquier subconjunto de una población. Si los elementos que componen la muestra son elegidos aleatoriamente y todos los elementos tienen la misma probabilidad de ser elegidos, entonces se trata de una muestra aleatoria simple.
- **Individuo:** Cada uno de los elementos que forman parte de la población, pudiendo ser algo con existencia real, como una persona, un automóvil o una casa, o algo más abstracto como la temperatura, una opinión, un voto o un sentimiento.
- **Variable:** Cualquier característica o propiedad que pueda ser estudiada en todos los elementos de la población, tales como el sexo, la edad, estatura, peso, color de pelo, nivel de estudios, entre otras.

Ejemplo 1. En la siguiente tabla se observa algunos ejemplos de las definiciones planteadas.

Población	Muestra	Individuo	Variable
Los libros de una biblioteca	20 libros de la biblioteca	Cada libro	Costo de sustitución Área de estudio Frecuencia de salida
Personas mayores de 18 años	2000 personas usadas en el estudio	Cada persona que pertenece a la población adulta	Estatura Peso
Alumnos inscritos en la universidad	3000 estudiantes usados en el estudio	Cada estudiante	Promedio Número de créditos Facultad

1.2.3.- Variables y Atributos

Variable estadística es toda característica medible objeto de nuestro estudio en los elementos de la muestra y que puede tomar un conjunto de valores.



Variables Cualitativas, Categóricas o Atributos: Son aquellas variables que no pueden ser descritas numéricamente. Se utiliza la palabra, el sustantivo, adjetivo y adverbio fundamentalmente, se clasifican en:

- **Ordinales:** Sugieren una ordenación o son susceptibles de ella; por ejemplo, el grado militar, el nivel de estudios, grado de satisfacción, entre otras.
- **Nominales:** Si sólo admiten una mera ordenación alfabética, pero no establece orden por su

naturaleza, por ejemplo el color de pelo, sexo, estado civil, etc.

Variables Cuantitativas o Numéricas: Son las que pueden ser descritas por medio de números, pudiendo ser:

- **Cuantitativas discretas:** Aquellas a las que se les puede asociar un número entero, es decir, aquellas que por su naturaleza no admiten un fraccionamiento de la unidad, por ejemplo, número de hermanos, páginas de un libro, etc.
- **Cuantitativas continuas:** Aquellas que por su naturaleza admiten que entre dos valores cualesquiera la variable pueda tomar cualquier valor intermedio, por ejemplo, peso, altura, tiempo, etc.

Para observar algunos ejemplos se puede revisar el siguiente enlace (Variables).

El valor cero de la variable cualitativa representa ausencia del atributo que se mide, y solo en este caso se tiene la propiedad de la proporcionalidad.

Un **valor atípico** es una observación extrañamente grande o pequeña. Los valores atípicos pueden tener un efecto desproporcionado en los resultados estadísticos, como la media, lo que puede conducir a interpretaciones engañosas.

1.3. ANÁLISIS DE DATOS

En estadística descriptiva se tienen tres formas generales para presentar un conjunto de datos y estas son: en forma de datos individuales, en forma de datos agrupados y mediante representaciones gráficas.

- **Datos individuales:** Cuando los datos se presentan explícitamente como una lista de valores.
- **Datos agrupados:** Cuando los datos están presentados mediante tablas, como en una tabla de frecuencias.
- **Representaciones gráficas:** Cuando un conjunto de datos se presentan gráficamente mediante histogramas, diagramas de barra, etc.

1.3.1.- TABLAS DE FRECUENCIAS

Una tabla de frecuencias o distribución de frecuencias es una herramienta que se emplea para resumir, mediante una tabla, numerosos datos de manera que se ponga de manifiesto la localización y la dispersión de las observaciones. Con una tabla de frecuencias se pueden resumir datos categóricos, nominales u ordinales. Si los datos son continuos se pueden resumir utilizando la misma técnica una vez que se los ha dividido mediante intervalos de clase.

Una tabla de frecuencias consta de dos columnas, en la primera se especifican los valores distintos en los datos (x_i) de forma ascendente y en la segunda la frecuencia absoluta con la que aparecieron dichos valores (f_i).

A partir de la frecuencia absoluta, suelen construirse otras estadísticas, como:

- La frecuencia relativa (f_r), que consiste simplemente en presentar la frecuencia absoluta en términos porcentuales. Considerando como el cien por ciento al tamaño de la muestra (N).
- La frecuencia absoluta acumulada (F), que consiste en ir realizando una suma acumulada de las frecuencias absolutas a través de las categorías, ya sea en forma ascendente o descendente. Y, de una forma similar se puede construir también la frecuencia relativa acumulada (F_r).

Ejemplo 2. Supongamos que el número de hijos de una muestra de 20 familias es el siguiente:

2	1	1	3	1	2	5	1	2	3
4	2	3	2	1	4	2	3	2	1

Realizar el análisis descriptivo.

Solución. Ahora vamos a proceder a realizar un análisis descriptivo del comportamiento del número de hijos por familia, tenemos que el tamaño de la muestra es $N = 20$, para resumir el conjunto de datos individuales utilizaremos una tabla de frecuencias, podemos ver que el menor número de hijos que se observó fue de 1 y el máximo 5; por lo tanto, el rango es $5 - 1 = 4$, de aquí:

x_i	f_i	$f_{r_i} = f_i \big/ N$	F_i	$F_{r_i} = F_i \big/ N$
1	6	0.30	6	0.30
2	7	0.35	13	0.65
3	4	0.20	17	0.85
4	2	0.10	19	0.95
5	1	0.05	20	1.00

Donde

$$F_i = \sum_{m=1}^{i} f_m \quad \text{y} \quad F_{r_i} = \sum_{m=1}^{i} f_{r_m}.$$

□

Agrupamiento mediante intervalos de clase

Si el número de valores distintos que toma la variable estadística X es demasiado grande o la variable es continua, se realiza un agrupamiento de los datos en intervalos y se hace un recuento del número de observaciones que caen dentro de cada uno de ellos. Para agrupar los datos mediante intervalos de clase no existe un único procedimiento y la forma en como se construyen estos intervalos depende básicamente de los objetivos de la investigación, sin embargo, aquí se especifica un procedimiento que se puede seguir para realizar este agrupamiento:

1. Determinar el recorrido o rango, de los datos.
2. Decidir el número k de intervalos de clase en que se van a agrupar los datos, $5 \leqslant k \leqslant 20$.

   Una regla que a veces se suele seguir es elegir $k = \sqrt{N}$.

3. Determinar la amplitud A (constante) de cada intervalo.

$$A = \frac{\text{Rango}}{k}$$

En la práctica, la amplitud de los intervalos no necesariamente tiene que ser igual. Si un intervalo de clase carece de datos es recomendable reorganizar la amplitud de los intervalos de clase.

4. Determinar los extremos de los intervalos de clase.

   - Límite inferior: L_i
   - Límite superior: $L_{i+1} = L_i + A$

   por convención los intervalos se definirán de la siguiente forma $[L_i - L_{i+1})$, se incluye el extremo inferior y el extremo superior no se incluye.

5. Calcular las marcas o puntos medios de clase de cada intervalo: m_i

$$m_i = \frac{L_i + L_{i+1}}{2}$$

Finalmente, la distribución de frecuencias en intervalos de clase tendría la siguiente estructura:

$[\boldsymbol{L_i - L_{i+1}})$	$\boldsymbol{m_i}$	$\boldsymbol{f_i}$	$\boldsymbol{f_{r_i}}$	$\boldsymbol{F_i}$	$\boldsymbol{F_{r_i}}$
$[L_1 - L_2)$	m_1	f_1	f_{r_1}	F_1	F_{r_1}
$[L_2 - L_3)$	m_2	f_2	f_{r_2}	F_2	F_{r_2}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$[L_k - L_{k+1}]$	m_k	f_k	f_{r_k}	F_k	F_{r_k}

De lo anterior podemos notar que:

1. **Frecuencias absolutas** f_i:

$$0 \leqslant f_i \leqslant N \quad \text{y} \quad \sum_{i=1}^{k} f_i = N.$$

2. **Frecuencias relativas** f_{r_i}:

$$f_{r_i} = \frac{f_i}{N}; \quad 0 \leqslant f_{r_i} \leqslant 1 \quad \text{y} \quad \sum_{i=1}^{k} f_{r_i} = 1.$$

3. **Frecuencias absolutas acumuladas** F_i

$$F_i = F_{i-1} + f_i; \quad F_1 = f_1 \quad \text{y} \quad F_k = N.$$

4. **Frecuencias relativas acumuladas** F_{r_i}

$$F_{r_i} = \frac{F_i}{N}; \quad F_{r_i} = F_{r_{i-1}} + f_{r_i}; \quad F_{r_1} = f_{r_1} \quad \text{y} \quad F_{r_k} = 1.$$

Ejemplo 3. En la siguiente tabla se listan los datos medidos por James Short en 1763 sobre el paralaje del Sol en segundos de arco. El paralaje es el ángulo subtendido por la Tierra vista desde el Sol. Se midió observando tránsitos de Venus desde diferentes posiciones y permitió la primera medida de la distancia Tierra-Sol, que es la unidad básica de la escala de distancias en el Sistema Solar (la unidad astronómica).

Datos (en segundos de arco):

8.63	10.16	8.50	8.31	10.80	7.50	8.12
8.42	9.20	8.16	8.36	9.77	7.52	7.96
7.83	8.62	7.54	8.28	9.32	7.96	7.47

Resumir este conjunto de datos agrupándolos mediante intervalos de clase.

En este caso, tenemos un conjunto de datos continuos, y procedemos de la siguiente manera:

1. Rango $= x_{\text{máx}} - x_{\text{mín}} = 10.80 - 7.47 = 3.33$.

2. Número de intervalos: $k = \sqrt{21} = 4.53$, luego, $k = 5$. Como se redondea por exceso, la amplitud del intervalo multiplicada por el número de intervalos será mayor que el recorrido y no tendremos problemas en los extremos.

3. Amplitud del intervalo: $3.33/5 = 0.666$, en consecuencia, tomemos la amplitud igual a 0.7.

   Si tomamos $L_1 = 7.47$ entonces el último extremo será $7.47 + (5 \times 0.7) = 10.97$ que resulta ser mayor que 10.80 (máximo). Ahora ya podemos calcular los extremos para cada intervalo de clase y las marcas de clase correspondientes.

4. Recuento y construcción de la tabla,

$L_i - L_{i+1}$	m_i	f_i	f_{r_i}	F_i	F_{r_i}
$[7.47 - 8.17)$	7.82	9	0.429	9	0.429
$[8.17 - 8.87)$	8.52	7	0.333	16	0.762
$[8.87 - 9.57)$	9.22	2	0.095	18	0.857
$[9.57 - 10.27)$	9.92	2	0.095	20	0.952
$[10.27 - 10.97)$	10.62	1	0.048	21	1
Total		21	1		

   donde, el primer intervalo $[7.47 - 8.17)$ y su punto medio se determinan así:

   - $L_1 = x_{\text{mín}} = 7.47$
   - $L_2 = 7.47 + 0.7 = 8.17$
   - $m_1 = \dfrac{7.47 + 8.17}{2} = 7.82$

   de la misma forma se determinan límites inferior, superior y punto medio para cada intervalo. Se cuenta el número de observaciones para cada intervalo y se ubica como frecuencia absoluta de cada intervalo. □

1.3.2.- Representaciones gráficas para variable cuantitativa de datos sin agrupar

Diagrama de puntos

Un diagrama de puntos es una gráfica utilizada para ilustrar un número reducido de datos, la cual permite identificar con facilidad dos características:

- La localización de los datos.
- La dispersión o variabilidad de los datos.

Diagrama de barras

- Se utiliza para representar datos de VARIABLES DISCRETAS.
- Se representan en el eje de abscisas los distintos valores de la variable.
- Sobre cada uno de estos valores se levanta una barra de longitud igual a la frecuencia correspondiente.
- Se pueden representar tanto las frecuencias absolutas n_i como las relativas f_i.

Polígono de frecuencias

- Se obtiene uniendo con rectas los extremos superiores de las barras del diagrama anterior.

Ejemplo 4. En el ejemplo 2, del número de hijos por familia, la representación gráfica mediante un diagrama de puntos es la siguiente:

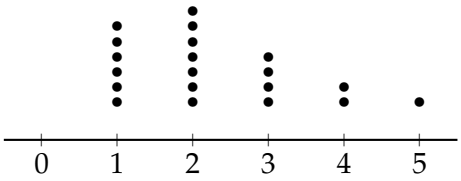


Figura 1.1: Diagrama de puntos de la variable número de hijos por familia.

Se observa que 6 familias (30 %) en la muestra tuvieron sólo 1 hijo y que la mayoría de las familias 13 (65 %) tuvieron como máximo 2 hijos. Finalmente, su diagrama de barras y respectivo polígono de frecuencias se pueden observar en la figura 1.2.

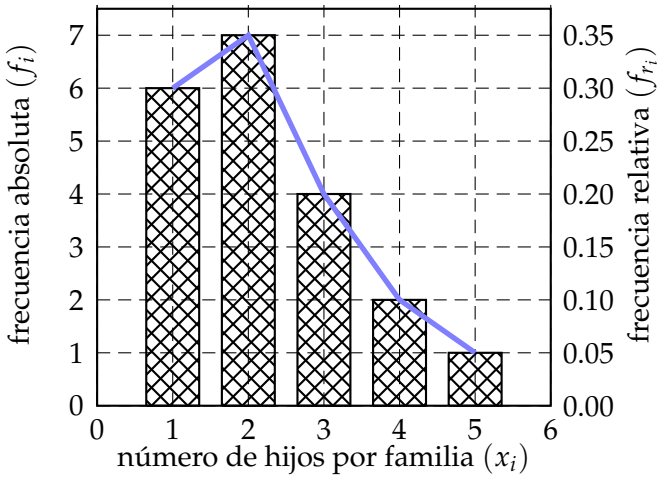


Figura 1.2: Diagrama de barras y polígono de frecuencias de la variable número de hijos por familia.

El gráfico o diagrama de barras también se lo puede presentar girado 90 grados (sobre todo cuando la variable tiene más categorías).

1.3.3.- Representaciones gráficas para variable cuantitativa de datos agrupados

Histograma

- Se utilizan principalmente para datos de variable continua.
- Es un conjunto de rectángulos adyacentes, cada uno de los cuales representa un intervalo de clase.
- La base de cada rectángulo es proporcional a la amplitud del intervalo.
- La altura de cada rectángulo corresponde a la frecuencia absoluta o relativa.

Polígono de Frecuencias

- Se obtiene uniendo con rectas los puntos medios de cada segmento superior de los rectángulos en el histograma.

Polígono de Frecuencias Acumuladas u Ojiva

- Sirve para representar las frecuencias acumuladas de datos agrupados por intervalos. El polígono parte de una altura cero para el extremo inferior del primer intervalo, sobre el extremo superior de cada intervalo se levanta una línea vertical de altura la frecuencia (absoluta o relativa) acumulada de ese intervalo. Evidentemente, la altura que se alcanza al final del polígono es N, para frecuencias absolutas, o 1, para frecuencias relativas.

Ejemplo 5. En el ejemplo 3 habíamos construido la distribución de frecuencias de la distancia Tierra-Sol, su representación gráfica viene dada por el siguiente histograma (figura 1.3):

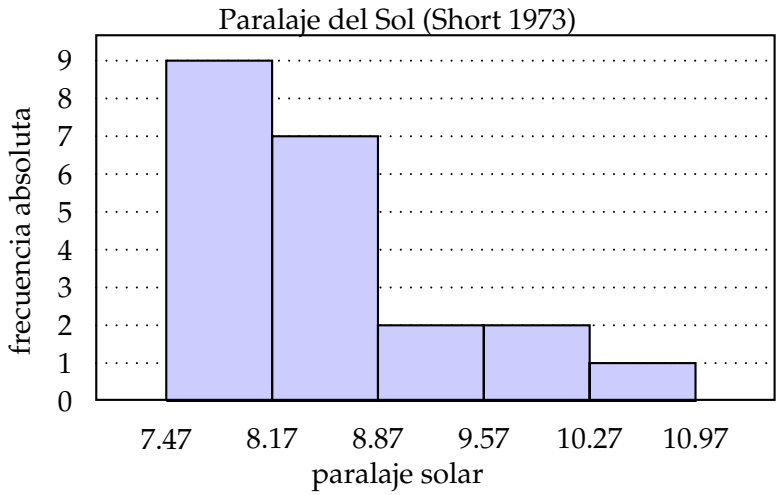


Figura 1.3: Histograma de las medidas de la paralaje del Sol.

Se puede observar que las medidas del paralaje del Sol más frecuentes están entre 7.47 y 8.17, y que medidas de la paralaje mayores a 10.27 son menos frecuentes. El polígono de frecuencias acumuladas (Ojiva) resultante vendría dado por (figura 1.4):

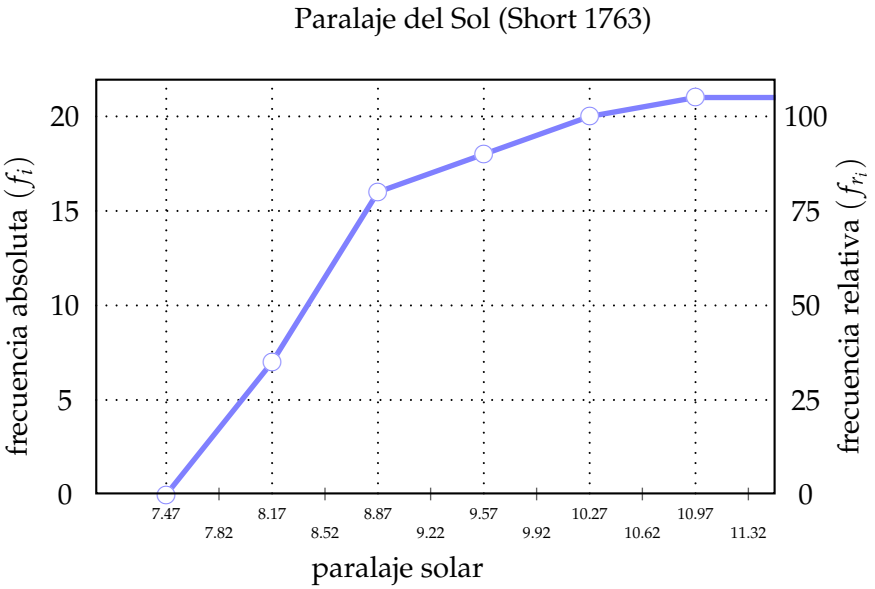


Figura 1.4: Polígono de frecuencias acumuladas de las medidas del paralaje del Sol.

Del polígono de frecuencias acumuladas, podemos concluir que de todas las medidas del paralaje del Sol en la muestra alrededor del 75 % son menores a 8.87.

1.3.4.- Representaciones gráficas para variables cualitativas

Existe una gran variedad de representaciones para variables cualitativas, de las cuales vamos a describir las dos más usadas.

Diagrama de barras o columnas

- Representar en el eje de abscisas las diferentes categorías y levantar sobre cada una de ellas un rectángulo o columna.
- La altura de cada rectángulo es la frecuencia (absoluta o relativa) de dicha categoría.

Diagrama de sectores o pastel

- Se representa el valor de cada categoría como un sector de un círculo completo.
- El área de cada sector es proporcional a la frecuencia de la categoría en cuestión. Se multiplica 360° por la frecuencia relativa correspondiente.
- Proporciona una idea visual muy clara de cuáles son las categorías mas representativas.

Ejemplo 6. Las notas de una asignatura de Física del curso académico 95/96 se distribuyeron de acuerdo a la siguiente tabla para los alumnos presentados en junio:

Nota	f_i	f_{r_i}	F_i	F_{r_i}
Suspenso (SS)	110	0.46	110	0.46
Aprobado (AP)	90	0.38	200	0.84
Notable (NT)	23	0.10	223	0.94
Sobresaliente (SB)	12	0.05	235	0.99
Matrícula de Honor (MH)	2	0.01	237	1.00

En este caso, Nota es una variable cualitativa, la misma consta de categorías que determinan el desempeño de los estudiantes en base a sus resultados SS, AP, NT, SB y MH. Se observa que el 5 % de los estudiantes fueron sobresalientes y sólo el 1 % estuvieron en el grupo de honor. Los diagramas de barras y de sectores correspondientes son los que se presentan en figura 1.5).

Para el diagrama de sectores es necesario saber el ángulo de cada uno de los sectores, para ello se tiene

Nota	**Ángulo [°]**
SS	165.6
AP	136.8
NT	36
SB	18
MH	3.6

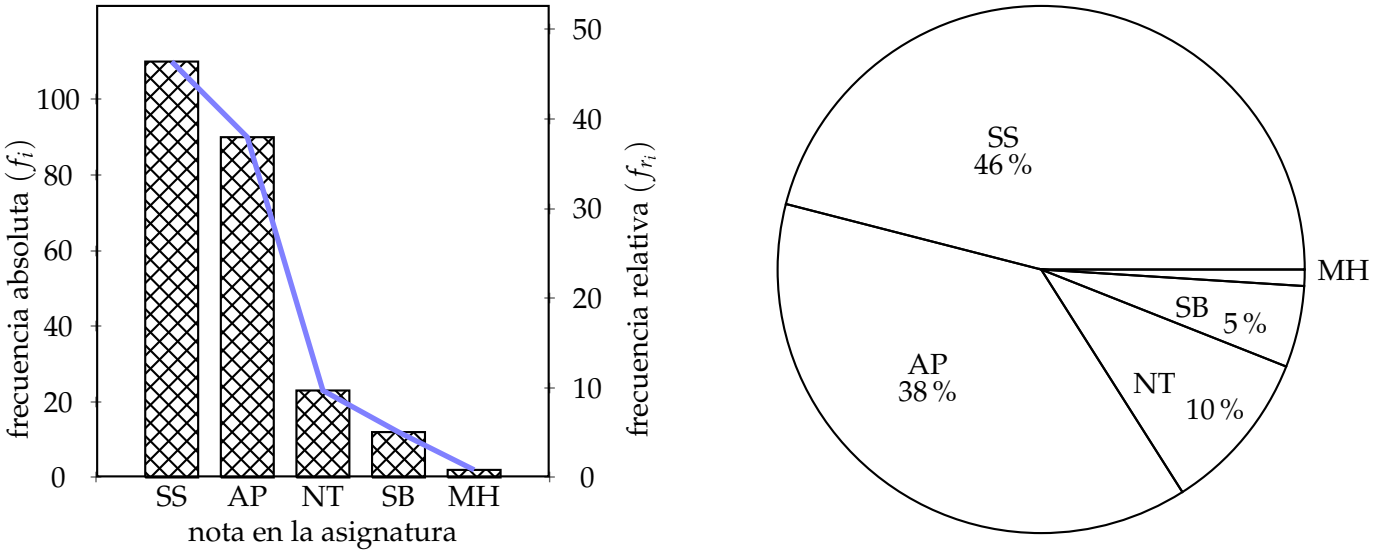


Figura 1.5: Diagrama de barras (izquierda) y de sectores (derecha).

El grupo de estudiantes suspensos representa la mayor parte del total por lo tanto le corresponde el sector mayoritario del circulo (46 %).

1.3.5.- Medidas tendencia central

Las medidas descriptivas es un conjunto de estadísticas que permiten resumir el comportamiento de un conjunto de datos, entre estas tendremos las medidas de tendencia central, de dispersión y de forma.

En el proceso de resumir los datos para describir la información, surgen otros procedimientos englobados en lo que se ha dado por llamar medidas de tendencia central, cuyo objetivo o propósito es hallar un valor tendencia central o categoría que sea "representativo" de todo el conjunto de datos. Este valor "representativo" de las características y atributos de todo el conjunto de datos, es lo que se conoce como el promedio o centro de la distribución de datos.

Sin embargo, es muy común observar en libros y publicaciones en general, que dependiendo del área de estudio, por ejemplo, en el campo económico, utilizan la técnica del valor modal para establecer el promedio de los ingresos familiares, o si se revisan aplicaciones en el área de la bioestadística, para establecer el valor promedio de medidas antropométricas, usan el método de la mediana.

Así, el valor promedio o central de un conjunto de datos, es un valor que trata de caracterizar o representar a todos los valores de la muestra, y no necesariamente siempre es la media aritmética.

El valor central, o simplemente el promedio de la distribución de datos, pueden obtenerse por tres métodos o procedimientos:

- la media aritmética,
- la mediana, y
- la moda.

De estos tres procedimientos, el más conocido y utilizado (por sus propiedades como estadístico) es la media aritmética, y de allí la razón para creer que la media aritmética es el promedio, siendo en realidad que la media aritmética es uno de los procedimientos o métodos para llegar a obtener el valor central o promedio de nuestro conjunto de datos.

Y, ante la pregunta "¿por qué hay varios métodos para calcular el promedio?", se debe al tipo de variable que se dispone.

Los procedimientos de cálculo del promedio, se definen, considerando que se tiene una variable X con una muestra de n valores $(x_1, x_2, \ldots, x_n)$, de la siguiente manera:

- **Media aritmética:** Se define la media aritmética (o simplemente media) para datos sin agrupar como:

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

o también

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

donde m es el número de categorías. Y para datos agrupados mediante intervalos de clase,

$$\overline{x} = \sum_{i=1}^{k} m_i \cdot f_{r_i}$$

- **Mediana:** Es el valor que divide a la distribución de datos en dos partes iguales. Pero, para establecer tal valor, los datos deben ser primeramente ordenados, ya sea en forma ascendente o descendente. Así, se tiene que de todo el conjunto de datos, el 50 % está por debajo de la mediana, y el otro 50 % está por encima de la mediana.

  El procedimiento a seguir cuando los datos son individuales es:

  - ordene las n observaciones de menor a mayor.
  - mediana muestral es igual a la observación en la posición $\dfrac{n+1}{2}$, si n es impar.
  - mediana muestral es igual al promedio de dos observaciones en las posiciones $\dfrac{n}{2}$ y $\dfrac{n+2}{2}$, si n es par.

  Si los datos están resumidos en intervalos de clase, la mediana se determina por interpolación, así:

  - Se determina la primera clase cuya frecuencia acumulada sea mayor o igual a $\dfrac{n}{2}$ dicho intervalo se denomina clase mediana.
  - La mediana Me se calcula con la fórmula,

$$\text{Me} = L_{i-1} + \frac{\dfrac{n}{2} - F_{i-1}}{f_i} A$$

  donde:

L_{i-1}: límite inferior de la clase mediana;

F_{i-1}: frecuencia acumulada del intervalo inmediatamente anterior a la clase mediana;

f_i: frecuencia absoluta de la clase mediana;

A: amplitud de la clase mediana.

- **Moda:** La moda Mo es aquel valor que tiene mayor frecuencia absoluta. Hay ocasiones en las cuales los datos pueden tener dos o más modas, o no puede existir, cuando todos los datos tienen igual frecuencia.

Otras medidas de posición

Cuantiles: Otras medidas resumen (no de tendencia central), pero sí de posicionamiento a lo largo de la distribución de los datos que ayudan a describir éstos, son los denominados cuantiles, entre los más frecuentemente utilizados tenemos:

- Cuartiles: son los valores del conjunto de datos que dividen a la distribución ordenada de datos en cuatro partes iguales.
- Quintiles: son los valores del conjunto de datos que dividen a la distribución ordenada de datos en cinco partes iguales.
- Deciles: son los valores del conjunto de datos que dividen a la distribución ordenada de datos en diez partes iguales, y finalmente,
- Percentiles: son los valores del conjunto de datos que dividen a la distribución ordenada de datos en cien partes iguales.

De todos analizaremos dos en detalle:

- Percentiles: Los percentiles P_k, son cada uno de los 99 valores que dividen a la distribución de los datos en 100 partes iguales. Para el cálculo del percentil de orden k se procede de la siguiente manera:
  - Si los n datos no están agrupados, se efectúa la siguiente descomposición:

$$\frac{nk}{100} = j + r$$

donde:
j: la parte entera de $\frac{nk}{100}$;
r: la parte fraccionaria de $\frac{nk}{100}$.

  - Calculamos el percentil de la siguiente forma:

$$P_k = \begin{cases} \frac{x_j + x_{j+1}}{2} & \text{si } r = 0 \\ x_{j+1} & \text{si } r > 0, \end{cases}$$

- Si los datos están agrupados en intervalos de clases, se calcula mediante:

$$P_k = L_{i-1} + \frac{\frac{nk}{100} - F_{i-1}}{f_i} A$$

donde:
k: es el orden del percentil;
L_{i-1}: límite inferior de la clase de interés, cuya frecuencia acumulada es la primera mayor o igual a $\frac{nk}{100}$;
F_{i-1}: frecuencia acumulada hasta L_{i-1};
f_i: frecuencia absoluta de la clase de interés;
A: amplitud de la clase de interés.

- Cuartiles: Son valores que dividen a la distribución de los datos en 4 partes, cada una de las cuales abarca al 25 % de los mismos. Los cuartiles son 3:
  - El cuartil inferior Q_1, que deja a su izquierda el 25 % de los datos y se cumple que $Q_1 = P_{25}$.
  - El cuartil medio Q_2, que deja a su izquierda el 50 %, que coincide con la mediana de los datos y se cumple que $Q_2 = \text{Me} = P_{50}$.
  - El cuartil superior Q_3, que deja a su izquierda el 75 % de los datos y se cumple que $Q_3 = P_{75}$.

El DIAGRAMA DE CAJAS se construye con L_s, L_i, Q_1, Q_2, Q_3 y RIQ, donde

$$RIQ = Q_3 - Q_1; \quad L_i = Q_1 - 1.5RIQ, \quad \text{y} \quad L_s = Q_3 + 1.5RIQ.$$

Los datos que se no se encuentren en intervalo $[L_i, L_s]$ se denominan atípicos; es más los datos que se encuentran fuera del intervalo $[L_I, L_S]$ se denominan atípicos extremos; donde $L_I = Q_1 - 3RIQ$ y $L_S = Q_3 + 3RIQ$.

Los datos pueden resultar de distribución simétrica o asimétrica ya sea negativa o positiva como se muestra en la figura 1.6.

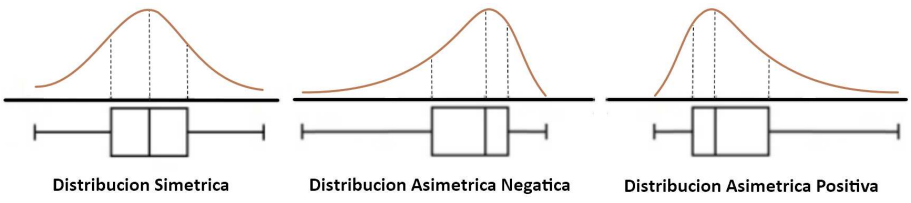


Figura 1.6: Relación entre la simetría y el diagrama de caja.

Ejemplo 7. (Cont. Ejemplo 2) Se tiene que el número de hijos en la muestra de 20 familias son los siguientes:

2	1	1	3	1	2	5	1	2	3
4	2	3	2	1	4	2	3	2	1

Encuentre la media, mediana, moda, los cuartiles del número de hijos y el diagrama de caja.

Solución. Vamos a encontrar las medidas de tendencia central y posición del número de hijos en las familias, en este caso utilizaremos los métodos definidos para un conjunto de datos individuales, por lo tanto:

- Cálculo de la media aritmética
$$\overline{x} = \frac{1}{n}\sum_{i=1}^{20} x_i = 2.25$$
La media representa una especie de centro de gravedad del conjunto de observaciones, en este caso podemos decir que el número promedio de hijos por familia fue de 2.25, otra forma de interpretar este valor sería que si el promedio de hijos por familia es de 2.25 entonces 40 familias tendrían un total de $40 \times 2.25 = 90$ hijos en promedio.

- Para el cálculo de la mediana se necesita tener el conjunto de datos ordenados en orden ascendente:
$$1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 2 \quad 2 \quad 2 \quad 2 \quad 2 \quad 2 \quad 2 \quad 3 \quad 3 \quad 3 \quad 3 \quad 4 \quad 4 \quad 5$$
En este caso, tenemos que el conjunto de datos tiene un número de observaciones par ($n = 20$), por lo tanto
$$\text{Me} = \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2} = \frac{x_{10} + x_{11}}{2} = \frac{2+2}{2} = 2$$

- Para este conjunto de datos el valor que más se repite es 2, por lo tanto
$$\text{Mo} = 2$$

- Los cuartiles son los valores que dividen en cuatro partes iguales al conjunto de datos en términos de su número de observaciones, como $np = 20 \times 0.25 = 5$ y $np = 20 \times 0.75 = 15$, entonces:
$$\begin{aligned} Q_1 &= \frac{x_{np} + x_{np+1}}{2} = \frac{x_5 + x_6}{2} = \frac{1+1}{2} = 1 \\ Q_2 &= \text{Me} = 2 \\ Q_3 &= \frac{x_{np} + x_{np+1}}{2} = \frac{x_{15} + x_{16}}{2} = \frac{3+3}{2} = 3 \end{aligned}$$
Por debajo del primer cuartil se tiene el 25 % de las observaciones, del segundo cuartil el 50 % y del tercer cuartil el 75 %.

Ahora, determinaremos el mismo grupo de medidas pero utilizaremos datos agrupados, utilizando una tabla de frecuencias para agrupar los datos tenemos:

x_i	f_i	f_{r_i}	F_i	F_{r_i}
1	6	0,3	6	0,3
2	7	0,35	13	0,65
3	4	0,2	17	0,85
4	2	0,1	19	0,95
5	1	0,05	20	1

Por lo tanto, las medidas de tendencia central y posición son:

- Para la media tenemos
$$\overline{x} = \frac{1}{n}\sum_{i=1}^{q} x_i \times f_i$$
donde q es el número de clases en la tabla de frecuencias, por lo tanto
$$\overline{x} = \frac{1}{20}\sum_{i=1}^{5} x_i \times f_i = \frac{1}{20}\left[(1\times 6) + (2\times 7) + \ldots + (5\times 1)\right] = 2.25.$$

- Para el cálculo de la mediana se puede utilizar el mismo método del ejemplo anterior, como tenemos $n = 20$, entonces
$$\text{Me} = \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2} = \frac{x_{10} + x_{11}}{2}$$
las observaciones 10 y 11 pertenecen a la segunda clase, por lo tanto,
$$\text{Me} = \frac{2+2}{2} = 2.$$
Por otro lado, utilizando el método de datos agrupados se tiene que
$$F(\text{Me}) = 0.5(20) = 10 \;\equiv\; \text{Me} = 2.$$

- La moda es el valor que más se repite en el conjunto de observaciones, podemos ver que la mayor frecuencia se presenta en la segunda clase, por lo tanto
$$\text{Mo} = 2.$$

- Para el cálculo de los cuartiles se tiene que
$$\begin{aligned} F(Q_1) = 0.25(20) = 5 &\;\equiv\; Q_1 = 1 \\ Q_2 = \text{Me} &\;\equiv\; Q_2 = 2 \\ F(Q_3) = 0.75(20) = 15 &\;\equiv\; Q_3 = 3 \end{aligned}$$

- Vamos a determinar si existen datos atípicos, para ello tenemos que $RIQ = Q_3 - Q_1 = 2$, luego, determinemos
$$L_i = Q_1 - 1.5RIQ = -2 \quad \text{y} \quad L_s = Q_3 + 1.5RIQ = 6.$$

Se observa que tanto $x_{\text{máx}} < L_s$ y $x_{\text{mín}} > L_i$, en consecuencia, no existen datos atípicos y los bigotes de la caja tienen como extremos exactamente estos dos valores.

A continuación, se muestra el diagrama de caja respectivo (figura 1.7).

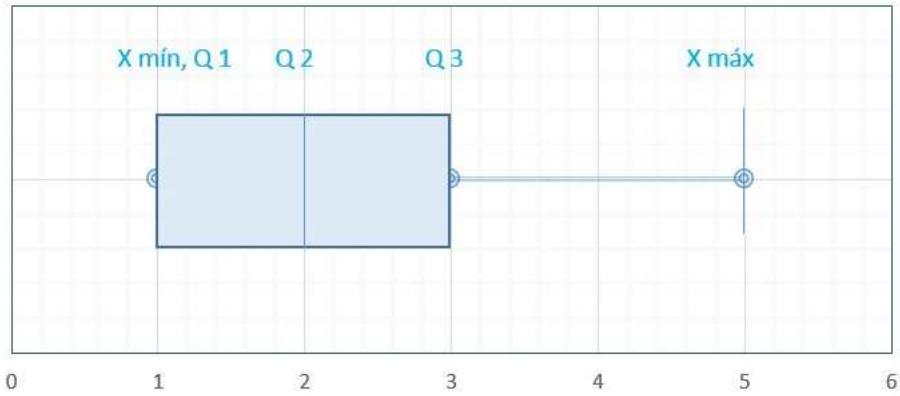


Figura 1.7: Diagrama de caja de la variable número de hijos.

Se observa que los datos tiene una distribución asimétrica positiva, además se observa que los datos están distribuidos simétricamente entre el cuartil 1 y el 3, es decir, la misma cantidad de datos se ubican entre Q_1 y Q_2; y Q_2 y Q_3. También hay una dispersión de los datos entre Q_3 y $x_{\text{máx}}$ mayor a la observada entre los cuartiles. □

Ejemplo 8. (Cont. Ejemplo 3) Se tiene la siguiente muestra de medidas de la paralaje del Sol respecto a la Tierra.

Datos (en segundos de arco):

8.63	10.16	8.50	8.31	10.80	7.50	8.12
8.42	9.20	8.16	8.36	9.77	7.52	7.96
7.83	8.62	7.54	8.28	9.32	7.96	7.47

1. Encuentre la media, mediana, moda, los cuartiles y el diagrama de caja de la medidas del paralaje del Sol respecto de la Tierra con los datos individuales.
2. Encuentre la media, mediana, moda, los cuartiles y el diagrama de caja de la medidas del paralaje del Sol respecto de la Tierra con los datos agrupados.

Solución.

1. Vamos a determinar las medidas de tendencia central y de posición usando los datos individuales, para ello, ordenemos lo datos de menor a mayor:

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
7.47	7.50	7.52	7.54	7.83	7.96	7.96	8.12	8.16	8.28	8.31	8.36	8.42	8.50	8.62	8.63	9.20	9.32	9.77	10.16	10.80

Determinemos la media

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

$$= \frac{1}{20}(7.47 + 7.50 + \ldots + 10.80)$$
$$= 8.4966.$$

Vamos a determinar los cuartiles de los datos de forma individual, así tenemos que que determinar el número np en el caso del primer cuartil es 5.25 en el caso del segundo es 10.5 y en el caso del tercero es 17.75, en consecuencia,

$$Q_1 = x_{\lceil np \rceil} = x_6 = 7.96$$
$$Q_2 = x_{\lceil np \rceil} = x_{11} = 8.31$$
$$Q_3 = x_{\lceil np \rceil} = x_{16} = 8.63$$

Recuerde que Me $= Q_2 = 8.31$. También $RIQ = Q_3 - Q_1 = 0.67$ de donde

$$L_i = Q_1 - 1.5RIQ = 6.955 \quad \text{y} \quad L_s = Q_3 + 1.5RIQ = 9.635$$

Se observa que $x_{\text{mín}} > L_i$ por lo que no hay datos atípicos por debajo de L_i, mientras que $x_{\text{máx}} > L_s$, por tanto existen datos atípicos

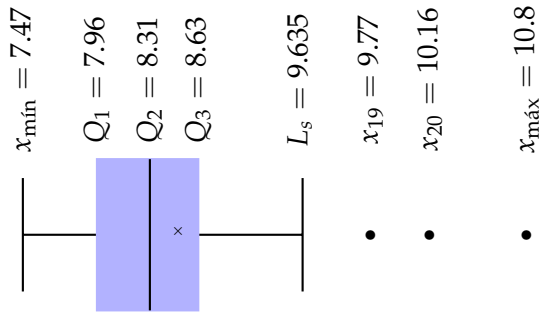


Figura 1.8: Diagrama de caja de la variable paralaje del Sol respecto a la Tierra.

2. Vamos a determinar las medidas de tendencia central utilizando datos agrupados mediante intervalos de clase, en el ejemplo 3 se obtuvo para este conjunto de datos la siguiente distribución de frecuencias:

i	$L_i - L_{i+1}$	m_i	f_i	f_{r_i}	F_i	F_{r_i}
1	7.405 - 8.105	7.755	7	0.333	7	0.333
2	8.105 - 8.805	8.455	9	0.429	16	0.762
3	8.805 - 9.505	9.155	2	0.095	18	0.857
4	9.505 - 10.205	9.855	2	0.095	20	0.952
5	10.205 - 10.905	10.555	1	0.048	21	1
			21	1		

cuyo polígono de frecuencias acumuladas (Ojiva) resultante se puede ver en la figura 1.9.

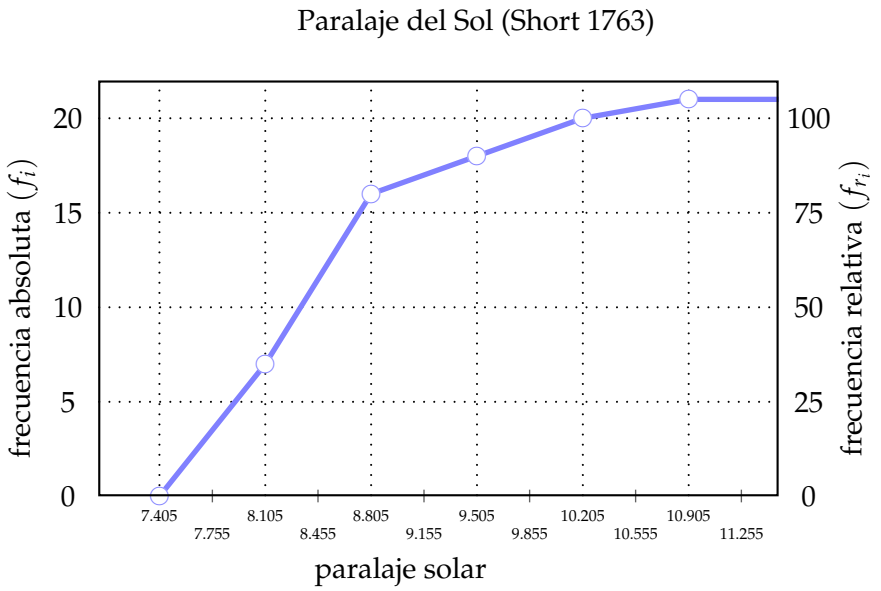


Figura 1.9: Polígono de frecuencias acumuladas de las medidas de la paralaje del Sol.

Para determinar las medidas de tendencia central y posición tenemos:

- La media calculada a partir de una distribución de frecuencias viene dada por:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{q} m_i \times f_i$$

donde, m_i es el punto medio de cada clase y q es el número de clases, por lo tanto

$$\overline{x} = \frac{1}{21}\sum_{i=1}^{5} m_i \times f_i = \frac{1}{21}\left[(7.755 \times 7) + (8.455 \times 9) + \ldots + (10.555 \times 1)\right] = 8.52$$

Obsérvese que la media calculada de esta forma difiere de la media que hemos calculado en la primera parte. Esto se debe a la agrupación de los datos en intervalos.

- Para la mediana se tiene que

$$F_r(\text{Me}) = 0.5$$

Ya que ningún intervalo de clase tiene un extremo que tenga una frecuencia relativa acumulada de 0.5 es necesario utilizar una interpolación a partir del polígono de frecuencias, como en el intervalo de clase $(8.105, 8.805]$ la frecuencia relativa se acumuló desde 0.333 hasta 0.762, entonces la mediana pertenecerá a este intervalo, por lo tanto

$$F_r(\text{Me}) - F_{r_2} = m(x - L_2)$$

donde,

$$m = \frac{F_{r_3} - F_{r_2}}{L_3 - L_2} = \frac{0.762 - 0.333}{8.805 - 8.105} \approx 0.613,$$

entonces, la ecuación de interpolación resultante viene dada por

$$F_r(\text{Me}) - 0.333 = 0.613(\text{Me} - 8.105)$$

por lo tanto, para la mediana tenemos

$$0.5 - 0.333 = 0.613(\text{Me} - 8.105) \equiv \text{Me} = \frac{0.5 - 0.333}{0.613} + 8.105$$
$$\equiv \text{Me} \approx 8.38.$$

También diferente de la mediana que se obtuvo en la primera parte.

- La moda en una distribución de frecuencias es el punto medio del intervalo modal, podemos ver que la mayor frecuencia se presenta en segundo intervalo de clase, por lo tanto

$$\text{Mo} = 8.455.$$

- Para el cálculo de los cuartiles se tiene que

$$F_r(Q_1) = 0.25,$$
$$Q_2 = \text{Me} = 8.38, \text{y}$$
$$F_r(Q_3) = 0.75$$

Para Q_1 y Q_3 es necesario utilizar el mismo procedimiento que para la mediana, para Q_1 la ecuación de interpolación viene dada por:

$$0.25 - 0 = 0.476(Q_1 - 7.405)$$

por lo tanto,

$$Q_1 = \frac{0.25}{0.476} + 7.405 \equiv Q_1 \approx 7.93.$$

por otro lado, para Q_3 se tiene

$$0.75 - 0.333 = 0.613(Q_3 - 8.105)$$

de donde,

$$Q_3 = \frac{0.75 - 0.333}{0.613} + 8.105 \equiv Q_3 \approx 8.79.$$

- Vamos a determinar si existen datos atípicos, para ello tenemos que $RIQ = Q_3 - Q_1 = 0.86$, luego, determinemos

$$L_i = Q_1 - 1.5RIQ = 6.64 \quad \text{y} \quad L_s = Q_2 + 1.5RIQ = 10.08.$$

Se observa que $x_{\text{mín}} > L_i$ por lo que no existe datos atípicos por debajo de L_i, que en este caso es $x_{\text{mín}}$. Ahora, puesto que $x_{\text{máx}} > L_s$ existen datos atípicos por encima de L_s. En este caso los datos atípicos son

$$10.16 \quad \text{y} \quad 10.80.$$

A continuación, se muestra el diagrama de caja respectivo (figura 1.10).

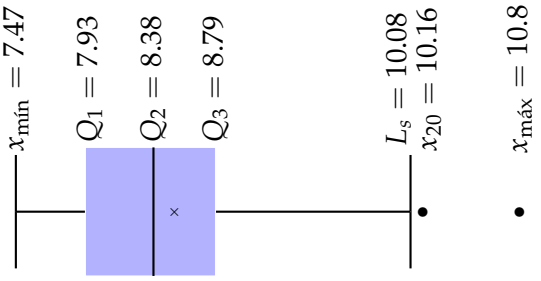


Figura 1.10: Diagrama de caja de la variable paralaje del Sol respecto a la Tierra.

Cuando se agrupan los datos, el número de datos atípicos se reduce a dos. Se observa que los datos tiene una distribución asimétrica positiva, que el bigote de la derecha $(x_{\text{máx}} - Q_3)$ es más largo que el bigote de la izquierda $(Q_1 - x_{\text{mín}})$, esto significa que los datos están mucho mas dispersos en ese intervalo que con respecto al otro bigote donde están más "juntos". Además, se observa que los datos están distribuidos simétricamente entre el cuartil Q_1 y Q_2; y también entre entre Q_2 y Q_3. □

1.3.6.- MEDIDAS DE DISPERSIÓN

Permiten medir el grado de dispersión alrededor del centro,estas medidas tienen la propiedad de que si los datos están ampliamente extendidos, la medida será alta; y cuando los datos se encuentren muy agrupados, será baja. Existen varias medidas de dispersión,

- Desviación estándar: La desviación estándar, notada como s, de un conjunto de n mediciones individuales $x_1, x_2, \ldots, x_n$ es la raíz cuadrada de la suma de los cuadrados de las desviaciones de las mediciones, respecto al promedio $\overline{x}$, dividida entre $n-1$; es decir:

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

  Si los datos están agrupados:

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{k} f_i(x_i - \overline{x})^2},$$

  La desviación estándar es siempre positiva y tiene las mismas unidades que la variable X.

- Varianza: La varianza s^2 es el cuadrado de la desviación estándar, es decir:

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

- Rango y Rango intercuartil:

  - El rango de n mediciones es la diferencia entre el valor máximo y el valor mínimo de la variable X,

$$\text{Rango} = x_{\text{máx}} - x_{\text{mín}}$$

- El rango intercuartil, denotado por RIQ de un conjunto de datos es igual a la diferencia entre los cuartiles superior e inferior, es decir,

$$RIQ = Q_3 - Q_1$$

- Coeficiente de variación: El coeficiente de variación, notado por CV, es igual a la desviación estándar dividida por la media, es decir,

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

Si el coeficiente variación es menor que 20 % se puede asumir homogeneidad; por otro lado, si el coeficiente variación es mayor que 20 % se asume heterogeneidad.

El coeficiente de variación permite comparar las dispersiones de dos muestras distintas, con distribuciones similares, y que se trate de una misma variable para las dos muestras, siempre que sus medias sean positivas.

Si la desviación de los datos con relación a la media es grande, entonces el promedio no tiene mérito; en caso contrario, el promedio es altamente significativo.

Se puede calcular el coeficiente de variación para dos muestras y concluir que el coeficiente de variación mayor coincide con la muestra de mayor variabilidad relativa.

Ejemplo 9. (Cont. Ejemplo 2) Se tiene que el número de hijos en la muestra de 20 familias son los siguientes:

2	1	1	3	1	2	5	1	2	3
4	2	3	2	1	4	2	3	2	1

Encuentre el rango, varianza y desviación estándar del número de hijos.

Solución. Vamos a determinar las medidas de dispersión del número de hijos en las familias de la muestra.

- La primera medida es el rango o recorrido, en este caso viene dado por:

$$\text{Rango} = x_{\text{máx}} - x_{\text{mín}} = 5 - 1 = 4.$$

- Como tenemos una muestra se tiene que la varianza viene dada por:

$$s^2 = \frac{1}{n-1}\sum_{k=1}^{n}(x_k - \overline{x})^2 = \frac{1}{20-1}\left[(2-2.25)^2 + (1-2.25)^2 + \ldots + (1-2.25)^2\right] = 1.35.$$

- La desviación estándar de la muestra es la raíz cuadrada de la varianza, por lo tanto:

$$s = \sqrt{1.35} = 1.16.$$

La desviación estándar representa que tan alejados en promedio están los datos con respecto de su media, en este caso se tiene que en promedio el número de hijos de las familias distan de su media en 1.16.

Ahora, determinemos las mismas medidas pero utilizando datos agrupados, del ejemplo 7 vimos que

x_i	f_i	f_{r_i}	F_i	F_{r_i}
1	6	0.3	6	0.3
2	7	0.35	13	0.65
3	4	0.2	17	0.85
4	2	0.1	19	0.95
5	1	0.05	20	1

Por lo tanto, las medidas de dispersión vienen dadas por:

- El rango o recorrido:
$$R = x_{\text{máx}} - x_{\text{mín}} = 5 - 1 = 4.$$
- La varianza para datos agrupados viene dada por:
$$s^2 = \frac{1}{n-1}\sum_{i=1}^{q}(x_i - \overline{x})^2 \times f_i$$
donde q es el número de clases en la tabla de frecuencias, entonces:
$$\begin{aligned} s^2 &= \frac{1}{20-1}\sum_{i=1}^{5}(x_i - \overline{x})^2 \times f_i \\ &= \frac{1}{20-1}\left[(1-2.25)^2 \times 6 + (2-2.25)^2 \times 7 + \ldots + (5-2.25)^2 \times 1\right] \\ &= 1.35 \end{aligned}$$
- La desviación estándar es entonces:
$$s = \sqrt{1.35} = 1.16.$$ □

Ejemplo 10. (Cont. Ejemplo 3) Se tiene la siguiente muestra de medidas de la paralaje del sol respecto a la tierra.
Datos (en segundos de arco):

8.63	10.16	8.50	8.31	10.80	7.50	8.12
8.42	9.20	8.16	8.36	9.77	7.52	7.96
7.83	8.62	7.54	8.28	9.32	7.96	7.47

1. Encuentre el rango, varianza y desviación estándar del número de hijos.

Solución. Vamos a determinar las medidas de dispersión utilizando datos agrupados mediante intervalos de clase, en el ejemplo 3 se determinó la siguiente distribución de frecuencias:

i	$L_i - L_{i+1}$	m_i	f_i	fr_i	F_i	F_{r_i}
1	7.405 - 8.105	7.755	7	0.333	7	0.333
2	8.105 - 8.805	8.455	9	0.429	16	0.762
3	8.805 - 9.505	9.155	2	0.095	18	0.857
4	9.505 - 10.205	9.855	2	0.095	20	0.952
5	10.205 - 10.905	10.555	1	0.048	21	1
			21	1		

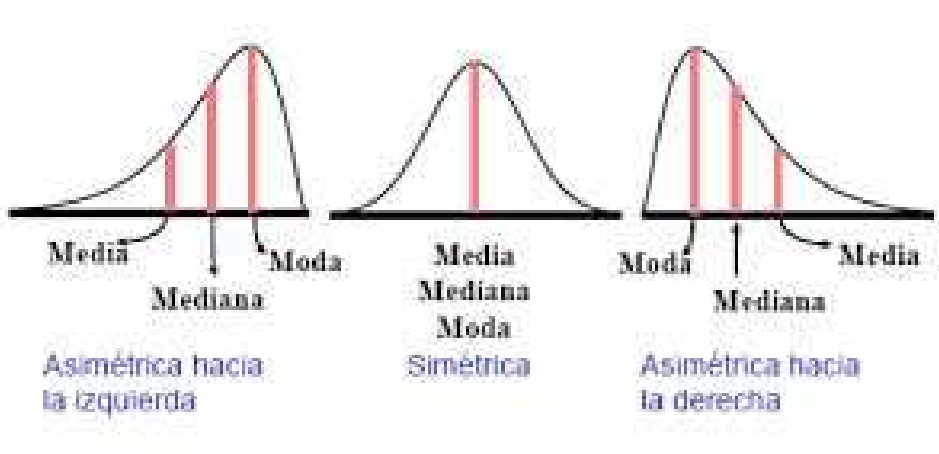
De donde, las medidas de dispersión vienen dadas por:

- El rango:
$$R = L_6 - L_1 = 10.905 - 7.405 = 3.5$$
- La varianza para datos agrupados viene dada por:
$$s^2 = \frac{1}{n-1}\sum_{i=1}^{q}(m_i - \overline{x})^2 \times f_i$$
donde q es el número de intervalos de clase en la distribución de frecuencias, entonces:
$$\begin{aligned} s^2 &= \frac{1}{20}\sum_{i=1}^{5}(m_i - \bar{x})^2 \times f_i \\ &= \frac{1}{20}\left[(7.755 - 8.52)^2 \times 7 + (8.455 - 8.52)^2 \times 9 + \cdots + (10.555 - 8.52)^2 \times 1\right] \\ &= 0.63 \end{aligned}$$
- La desviación estándar es entonces:
$$s = \sqrt{0.63} = 0.795.$$ □

1.3.7.- MEDIDAS DE FORMA

Las medidas de forma son aquellas que nos muestran si una distribución de frecuencia tiene características especiales como simetría, asimetría, nivel de concentración de datos y nivel de apuntamiento que la clasifiquen en un tipo particular de distribución. Gráficamente se puede visualizar los comportamientos de asimetría o curtosis a partir de los histogramas de las distribuciones de frecuencia.

- Coeficiente de asimetría A_s
  - Mide el grado de asimetría de la distribución de datos entorno a la media.

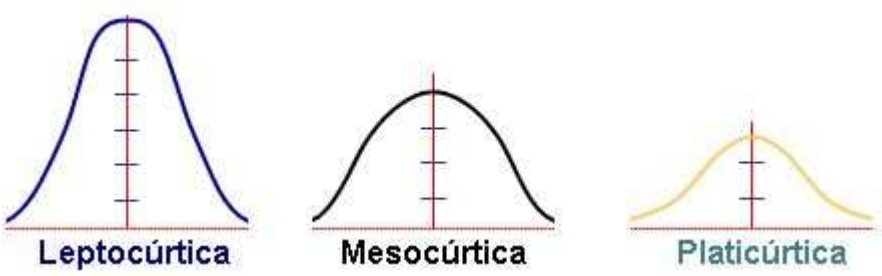


- Es adimensional y se calcula:

$$A_s = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^3/n}{s^3}$$

- Si $A_s > 0$, la distribución será asimétrica a la derecha.
- Si $A_s = 0$, la distribución será simétrica.
- Si $A_s < 0$, la distribución será asimétrica a la izquierda.

- Coeficiente de apuntamiento A_c: Este coeficiente de apuntamiento o curtosis sirve para medir el grado de concentración de los valores que toma en torno a su media.



- Se suele medir con el coeficiente de curtosis que se calcula:

$$A_c = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^4/n}{s^4}$$

$$A_p = A_c - 3$$

- LEPTOCÚRTICA: Si $A_p > 0$, más puntiaguda que la normal.
- MESOCÚRTICA: Si $A_p = 0$, es tan puntiaguda como la normal.
- PLATICÚRTICA: Si $A_p < 0$, menos puntiaguda que la normal.

Ejemplo 11. (Cont. Ejemplo 2) Se tiene que el número de hijos en la muestra de 20 familias son los siguientes:

2	1	1	3	1	2	5	1	2	3
4	2	3	2	1	4	2	3	2	1

1. Determinar los coeficientes de asimetría y curtosis de la distribución del número de hijos.

Solución. Vamos a determinar el coeficiente de asimetría y el de curtosis:

- Primero, determinemos el coeficiente de asimetría, como sabemos el coeficiente de asimetría sirve para determinar si una distribución de observaciones es simétrica o asimétrica, para esto utilizaremos el método definido para un conjunto de datos individuales, por lo tanto:

$$\begin{aligned} A_s &= \frac{1}{ns^3}\sum_{i=1}^{n}(x_i-\overline{x})^3 \\ &= \frac{1}{20(1.16)^3}\sum_{i=1}^{20}(x_i-\overline{x})^3 \\ &= \frac{1}{20(1.16)^3}\left[(2-2.25)^3+(1-2.25)^3+\cdots+(1-2.25)^3\right] \\ &\approx 0.68. \end{aligned}$$

  En este caso, el coeficiente de asimetría es positivo por lo tanto la distribución del número de hijos presenta un sesgo positivo o es sesgada a la derecha, esto lo podemos comprobar gráficamente en el diagrama de barras respectivo:

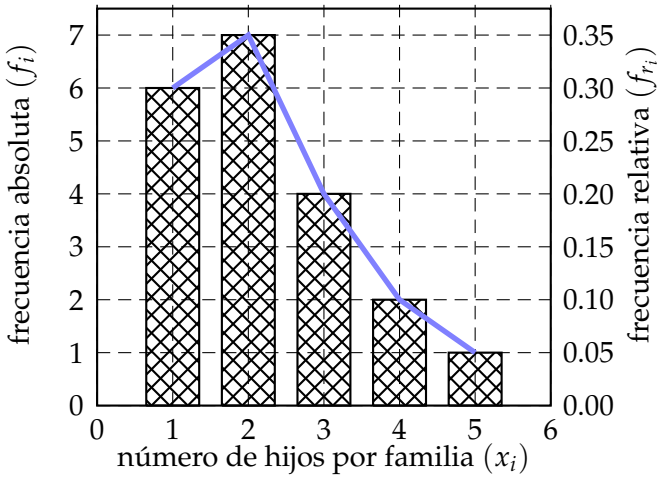


Figura 1.11: Diagrama de barras del número de hijos en las familias.

  En el diagrama de barras podemos ver gráficamente la asimetría hacia la derecha de la distribución del número de hijos en las familias.

- Segundo, determinemos el coeficiente de curtosis, para un conjunto de datos individuales tenemos que este coeficiente viene dado por:

$$\begin{aligned} A_c &= \frac{1}{ns^4}\sum_{i=1}^{n}(x_i-\overline{x})^4 \\ &= \frac{1}{20(1.16)^4}\sum_{i=1}^{20}(x_i-\overline{x})^4 \end{aligned}$$

$$= \frac{1}{20(1.16)^4}\left[(2-2.25)^4 + (1-2.25)^4 + \cdots + (1-2.25)^4\right]$$
$$\approx 2.5.$$

El coeficiente de curtosis es menor a 3 por lo tanto podemos decir que hay una escasa agrupación de datos alrededor de la media o a su vez que la distribución es platicúrtica.

Ahora, utilizaremos datos agrupados, la tabla de frecuencias del número de hijos es:

x_i	f_i	f_{r_i}	F_i	F_{r_i}
1	6	0,3	6	0,3
2	7	0,35	13	0,65
3	4	0,2	17	0,85
4	2	0,1	19	0,95
5	1	0,05	20	1

- Primero, determinemos el coeficiente de asimetría, para un conjunto de datos agrupados por tablas de frecuencia este coeficiente viene dado por:

$$A_s = \frac{1}{ns^3}\sum_{i=1}^{q}(x_i - \overline{x})^3 \cdot f_i,$$

donde q es el número de clases, por lo tanto:

$$A_s = \frac{1}{20(1.16)^3}\sum_{i=1}^{5}(x_i - \overline{x})^3 \cdot f_i = \frac{1}{20(1.16)^3}\left[6(1-2.25)^3 + \ldots + 1(5-2.25)^3\right] \approx 0.68,$$

se puede observar que obtenemos el mismo resultado obtenido para datos individuales.

- Segundo, determinemos el coeficiente de curtosis, en una tabla de frecuencia se calcula mediante:

$$\begin{aligned} A_c &= \frac{1}{ns^4}\sum_{i=1}^{q}(x_i - \overline{x})^4 \cdot f_i \\ &= \frac{1}{20(1.16)^4}\sum_{i=1}^{5}(x_i - \overline{x})^4 \cdot f_i \\ &= \frac{1}{20(1.16)^4}\left[6(1-2.25)^4 + \ldots + 1(5-2.25)^4\right] \\ &\approx 2.5. \end{aligned}$$

□

Ejemplo 12. En el ejemplo 3 se determinó la siguiente distribución de frecuencias de la paralaje del sol respecto a la tierra.

$L_i - L_{i+1}$	m_i	f_i	f_{r_i}	F_i	F_{r_i}
7.47 - 8.17	7.82	9	0.429	9	0.429
8.17 - 8.87	8.52	7	0.333	16	0.762
8.87 - 9.57	9.22	2	0.095	18	0.857
9.57 - 10.27	9.92	2	0.095	20	0.952
10.27 - 10.97	10.62	1	0.048	21	1
Total		21	1		

1. Determinar los coeficientes de asimetría y curtosis de la distribución de la paralaje solar.

Solución. ▪ Primero, determinemos el coeficiente de asimetría, del histograma (figura 1.12) podemos concluir que esta distribución tiene un sesgo positivo.

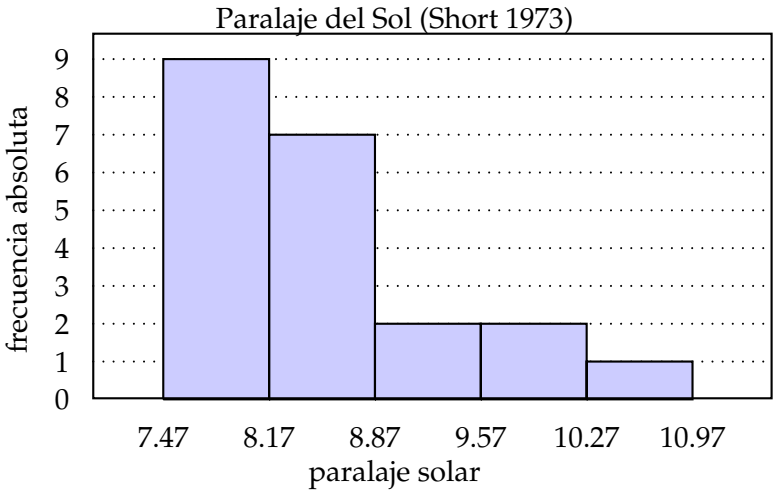


Figura 1.12: Histograma de las medidas de la paralaje del Sol.

El coeficiente de asimetría en una distribución de frecuencias viene dado por

$$A_s = \frac{1}{ns^3}\sum_{i=1}^{q}(m_i - \overline{x})^3 \cdot f_i,$$

donde q es el número de intervalos de clase, por lo tanto:

$$A_s = \frac{1}{21(0.83)^3}\sum_{i=1}^{5}(m_i - \overline{x})^3 \cdot f_i = \frac{1}{21(0.83)^3}\left[9(7.82 - 8.52)^3 + \ldots + 1(10.62 - 8.52)^3\right] \approx 1.03,$$

que es un valor positivo y verifica el sesgo hacia la derecha de la distribución, otra forma gráfica de verificar la asimetría de esta distribución es mediante un diagrama de caja o bigotes, como sabemos este gráfico permite ver como está distribuido un conjunto de datos a través de los cuartiles, el diagrama de caja para el paralaje solar es el siguiente:

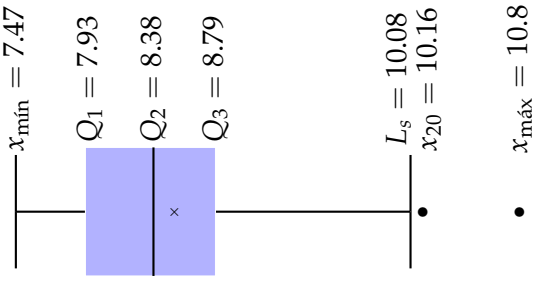


Figura 1.13: Diagrama de caja de la variable paralaje del Sol respecto a la Tierra.

De la misma forma, se puede notar la asimetría hacia la derecha de la distribución observando que el bigote de la derecha ($x_{\text{máx}} - Q_3$) es más largo que el bigote de la izquierda ($Q_1 - x_{\text{mín}}$) y que la cola de la derecha ($x_{\text{máx}} - Q_2$) es mas alargada que la cola de la izquierda ($Q_2 - x_{\text{mín}}$), por lo tanto podemos decir que el conjunto de datos tiene una mayor concentración de observaciones en la cola izquierda que en la cola derecha de la distribución y se concluye que la distribución tiene un sesgo positivo, si la distribución fuese simétrica tendríamos que todas las distancias entre cuartiles consecutivos serían semejantes con sus opuestos respectivos, esto gráficamente se tendría cuando el bigote de la derecha es igual de largo que el de la izquierda así como también las respectivas colas de la distribución observándose un rectángulo centrado justo en la mitad del gráfico, en el cual la mediana coincidiría con la media.

- Segundo, determinemos el coeficiente de curtosis:

$$\begin{aligned} A_c &= \frac{1}{ns^4}\sum_{i=1}^{q}(m_i - \overline{x})^4 \cdot f_i \\ &= \frac{1}{ns^4}\sum_{i=1}^{5}(m_i - \overline{x})^4 \cdot f_i \\ &= \frac{1}{21(0.83)^4}\left[9(7.82 - 8.52)^4 + \ldots + 1(10.62 - 8.52)^4\right] \\ &\approx 3.01, \end{aligned}$$

ya que su valor es mayor que 3 entonces podemos decir que la distribución es leptocúrtica. □

Ejemplo 13. La siguiente tabla muestra la distribución de sueldos de 120 trabajadores de una empresa.

Sueldo	Trabajadores
600-700	10
700-800	20
800-900	60
900-1000	20
1000-1100	10

1. Determinar los coeficientes de asimetría y curtosis de la distribución de sueldos.

Solución. Primero, debemos determinar los puntos medios de cada intervalo de clases, entonces:

Sueldo	m_i	f_i
600-700	650	10
700-800	750	20
800-900	850	60
900-1000	950	20
1000-1100	1050	10

La media y desviación estándar de la distribución de sueldos son \$850 y \$100, respectivamente. Ahora, podemos determinar los coeficientes de asimetría y curtosis.

- Primero, determinemos el coeficiente de asimetría:

$$A_s = \frac{1}{ns^3}\sum_{i=1}^{q}(m_i - \overline{x})^3 \cdot f_i,$$

entonces:

$$A_s = \frac{1}{120(100)^3}\sum_{i=1}^{5}(m_i - \overline{x})^3 \cdot f_i = \frac{1}{120(100)^3}\left[10(650-850)^3 + \ldots + 10(1050-850)^3\right] = 0,$$

en este caso el coeficiente de asimetría es 0, por lo tanto la distribución es perfectamente simétrica, veamos gráficamente la distribución de los sueldos mediante su histograma: Del histo-

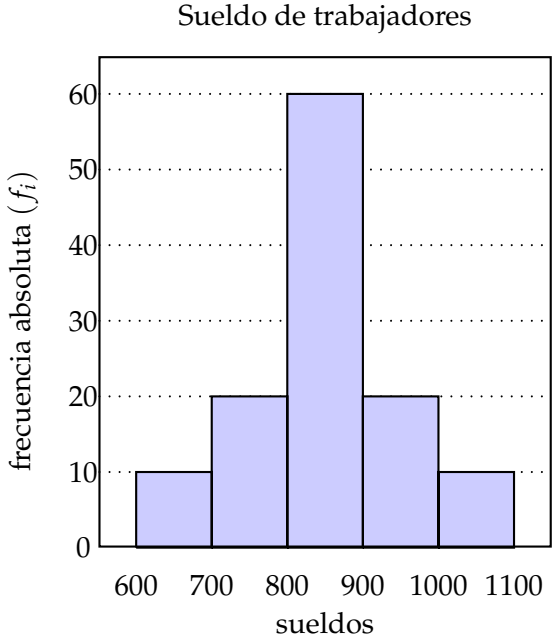


Figura 1.14: Histograma de los sueldos en la empresa.

grama, podemos ver la simetría perfecta de la distribución alrededor de la media (\$ 850), lo que corrobora el valor de su coeficiente de asimetría, el diagrama de caja respectivo es el que consta a continuación:

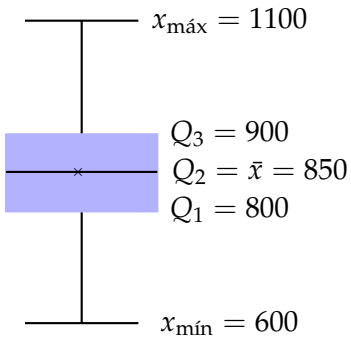


Figura 1.15: Diagrama de caja de los salarios en la empresa.

Igualmente, podemos notar del diagrama de caja la simetría perfecta de la distribución, los bigotes tienen la misma longitud, vemos un rectángulo centrado en la mitad del gráfico en el cual coinciden la media, la mediana y la moda.

- Segundo, determinemos el coeficiente de curtosis:

$$\begin{aligned} A_c &= \frac{1}{ns^4}\sum_{i=1}^{q}(m_i - \overline{x})^4 \cdot f_i \\ &= \frac{1}{120(100)^4}\sum_{i=1}^{5}(x_i - \overline{x})^4 \cdot f_i \\ &= \frac{1}{120(100)^4}\left[10(650-850)^4 + \ldots + 10(5-2.25)^4\right] \\ &= 3, \end{aligned}$$

como el valor del coeficiente es 3 podemos decir que la distribución de los sueldos es mesocúrtica. □