Breve repaso de Estadística Descriptiva

$Profesor$: René Iral Palomino
$Oficina$: $43 - 320$
$Correo$: $riral@unal.edu.co$

¿Por qué estudiar Estadística?

El estudio de la Estadística permite, entre otras cosas

- Aprender las reglas y métodos usados en el tratamiento de información
- Evaluar y cuantificar la importancia de los resultados estadísticos obtenidos
- Entender mejor algunos fenómenos de interés (Sociales, Económicos, Biológicos, Educacionales, etc.)
- Dar una visión más clara acerca de la información proveniente de diversas fuentes.

Algunos aspectos estadísticos manejados en la información obtenida de la radio, la televisión u otro medio, influencian fuertemente a gran cantidad de personas pero a veces no proporcionan una descripción cabal de lo que pretenden mostrar.

Como una de las tareas de la Estadística es el estudio de fenómenos aleatorios, esto hace muy pertinente el tratar de explicar la manera como se comportan (Variabilidad).

Entre otras cosas la Estadística se ocupa del manejo de la información que pueda ser cuantificada. Implica esto la descripción de conjuntos de datos y la inferencia a partir de la información recolectada de un fenómeno de interés.

Entre las principales funciones de la Estadística se destacan:

Resumir Simplificar Comparar Relacionar Proyectar.

Entre las tareas que debe enfrentar un estudio estadístico están:

- Delimitar con precisión la población de referencia o el conjunto de datos en estudio, las unidades que deben ser observadas, las características o variables que serán medidas u observadas.
- Estrategias de Observación: Censo, Muestreo, Diseño de Experimental.
- Recolección y Registro de la información.
- Depuración de la información.
- Producción de resúmenes estadísticos (gráficos y/o numéricos).
- Interpretación de los resultados.

Algunos tópicos fundamentales de la Estadística se presentan brevemente, los cuales se usan frecuentemente en investigación. Durante el transcurso del curso se estudiarán algunos de ellos.

- Diseño de experimentos. Esta relacionado con la etapa de obtención de información. Permite la determinación del tipo de datos a incluir en el estudio, la cantidad de datos. La determinación de cuantas unidades se deben incluir en el estudio es crucial ya que con esto se ahorra tiempo y dinero.
- Estadística descriptiva. Permite obtener un resumen de la información contenida en los datos por medio de funciones especícas llamadas estadísticos muestrales las cuales sirven para obtener valores numéricos que representan características sobresalientes que pudieran estar presentes. También permite la construcción de gráficos que permiten mirar en conjunto la totalidad de los datos y detectar comportamientos interesantes de ellos.
- Inferencia estadística. Permite evaluar la información de manera que se puedan obtener conclusiones generales del fenómeno bajo estudio.

- Estadística no parámetrica. Permite realizar pruebas estadísticas e implementar modelos donde no es posible asumir algunos supuestos previos.
- Elementos de regresión. Sirven para explorar la posible relación entre variables de respuesta y variables explicativas.

Niveles de medición y tipos de variables

Los siguientes ejemplos servirán para introducir algunas definiones importantes.

- Un investigador está interesado en determinar el caudal promedio de un río; para esto decide medir y registar tal caudal durante 30 días.
- Un investigador está interesado en determinar la proporción de personas que están a favor de una cierta ley de impuestos; para esto decide elaborar un cuestionario, selecciona adecuadamente una muestra al azar y registra la respuesta de los individuos que puede ser **SI**, **NO**, No sabe No responde **(NS/NR)**, las cuales pueden ser codificadas así: SI=1, NO=2, NS/NR=3.
- Un ingeniero esta interesado en determinar el número promedio de artículos defectuosos de una linea de producción; para esto decide contar y registrar diariamente y durante 30 días el numero de defectuosos.

Los tres experimentos expuestos tiene en común tres características:

1. Cada uno de ellos generan datos.
2. Cada uno de ellos tiene un factor de incertidumbre, pues en el momento de realizar cualquiera de ellos el investigador no sabe que resultado va a obtener.
3. Cada uno de ellos tiene un factor de variabilidad ya que en repeticiones sucesivas del experimento se pueden presentar resultados diferentes.

De los tres experimentos se puede observar que el primero de ellos (el de la medición de caudales) genera datos que son producto de mediciones. El segundo de ellos (el de la ley de impuestos) genera datos que representan categorías de respuesta y el tercero (el de la línea de producción) genera datos que son producto de conteos. Con lo anterior, podemos ahora dar algunas deficiones.

Variable. Es una característica que varía de un objeto o individuo a otro (por ejemplo la estatura, la dureza o el tiempo de duración de un componente) o en el mismo individuo (por ejemplo, la presión sanguinea). En estadística, los tipos más comunes de variables son Continuas, Discretas y Categóricas.

- **Variables continuas**. Son aquellas que provienen de procesos que involucran mediciones. Por ejemplo las estaturas de los estudiantes de primer año en una universidad.
- **Variables discretas**. Son aquellas que provienen de procesos que involucran conteos. Por ejemplo el número de vehículos que llegan a un semáforo en un intervalo de tiempo.
- **Variables categóricas**. Son aquellas que provienen de procesos que involucran clasifiaciones. Por ejemplo la variable sexo o estrato socioeconómico.

Observe que la variable que se genera en un experimento de medición de presión sanguinea es de naturaleza diferente a la de clasificar personas por su sexo. La primera se registra en milímetros de mercurio y además valores grandes dan la idea de mayor presión sanguníea mientra que la segunda se mide por medio de valores que representan la pertenencia a una categoría, por ejemplo 1=Masculino, 2=femenino, pero el 2 no indica una categoría mayor a la que representa el 1.
La diferencia en la información obtenida permite identificar cuatro niveles básicos de medición que son:

1. **Nominal**. Este nivel se utiliza cuando los valores en los que se mide la variable son códigos que representan la pertenencia a una categoría.

Por ejemplo, en un estudio de una cierta enfermedad, el 1 puede representar su presencia y el 0 su ausencia. Otro ejemplo puede ser estado civil, 1=Casado, 2=Soltero, 3=Unión libre. Observe que no se puede decir que 3 ¿2. Las variables de tipo nominal no admiten medidas básicas de resumen.

2. **Ordinal**. Se usa cuando los valores de una variable informan acerca de un orden o jerarquía. Por ejemplo, se pueden usar los valores 1, 2 y 3 para representar distintas quemaduras, es decir, 1=leve, 2=severa, 3=muy severa. Con este tipo de variables ya tiene sentido establecer una relación de orden y afirmar que $3 > 2 > 1$.

3. **Intervalo**. Se usa para mediciones de naturaleza cuantitativa que se hacen con escalas que tienen como base un valor de cero arbitrario. Por ejemplo un registro de $0 \pm C$ no indica la ausencia de temperatura.

4. **Razón**. Se usa para mediciones de naturaleza cuantitativa que se hacen con escalas que tienen como base un valor de cero absoluto. Por ejemplo, longitud del brazo, estatura, tiempo de duración, número de artículos defectuosos en una linea de producción, presión sanguínea.

Conceptos básicos

Un aspecto importante en Estadística está relacionado con a manera como la información es presentada y analizada. De este análisis previo pueden desprenderse diferentes formas de abordar la solución a determinada pregunta de investigación. Una primera parte consiste en realizar un adecuado resumen de la información disponible y presentarla en términos de algunas medidas puntuales o de gráficos.

Aspectos principales a tener en cuenta en la descripción de un conjunto de datos

a) Resumen y descripción de diferentes patrones en los datos por medio de:

- Presentación de tablas y graficos.

- Examinar de todas las formas posibles los gráficos en busca de características de interés.
- Buscar en los datos graficados observaciones inusuales, que se alejan del grueso de observaciones graficadas.

b) Cálculo de medidas numéricas.

- Valores típicos o representativos que den idea de centralidad o localización.
- La variabilidad presente en los datos.

Descripción de Datos por tablas o gráficos

Distribuciones de frecuencia

Cuando se tiene un número considerable de datos, una manera de representarlos es a través de un agrupamiento en clases. Si los datos son de tipo discreto o categórico, las clases estarán determinadas por las escalas de medición de la variable de interés. Sinembargo, si el número de valores que asume la variable es muy grande, es necesario agrupar dichos valores en clases. En el caso de variables continuas, es imperativo realizar un agrupamiento de los datos, considerando observaciones cercanas, en clases o intervalos. El resultado de este agrupamiento es resumido en una tabla, que usualmente se denomina *Tabla de Frecuencias*. El procedimiento a seguir para este proceso es como sigue:

a) Encuentre el mínimo y máximo de los valores registrados.

b) Escoja un número de subintervalos o clases de igual longitud, de manera que cubran el rango de los datos, sin traslaparsen (aunque es posible construir clases o intervalos de longitudes variables). Estos intervalos son llamados *Intervalos de Clase*.

c) Cuente cuantas observaciones están en cada subintervalo. Este conteo es llamado *Frecuencia de Clase*.

d) Calcule, para cada clase, la frecuencia relativa. Esta se calcula como:

$$FR = \frac{\text{Frecuencia de clase}}{\text{Número total de observaciones}} .$$

La elección del número de clases o intervalos, constituye un proceso de ensayo y error. Algunas propuestas empíricas se han planteado, buscando una selección más o menos adecuada del número de clases. No se puede establecer que una es superior a otra, sólo pueden utilizarse como puntos de referencia.

- Sturges (1926), establece que el número de clases es K puede obtenerse como $K = 1 + 3.33\, log_{10}(n)$, siendo n el número de datos. Esta propuesta subestima el número de intervalos.
- Velleman (1976), $K = 2\sqrt{n}$, recomendada cuando n es pequeño ($n \leq 50$)
- Dixon y Kronmal (1965), $K = 10 log_{10}(n)$, para n grande ($n > 50$).

En general, se sugiere que entre 5 y 25 clases es un número adecuado para agrupar los datos. Sinembargo, debe tenerse especial cuidado en esta selección. Es importante anotar que al agrupar los datos en clases, se sacrifica la información relacionada con cómo se distribuyen los datos en cada clase, y se reemplaza por la frecuencia en dicha clase. Si se tienen pocas clases, la pérdida de información es muy grande. Si se tienen pocos datos y muchas clases, no se evidenciará ningún tipo de comportamiento de interés en los datos.

Cuando se tiene un número considerable de datos, es importante establecer algún tipo de técnica para identificar datos en cada clase (una puede ser organizando los datos de menor a mayor). Los avances computacionales han permitido obviar este aspecto.

Ejemplo
Se tiene información de un grupo de estudiantes de un curso de primer semestre, donde se registraron, para cada sujeto, la Estatura(en cms), Masa(en Kg), Edad (en días), Género (HOMBRE o MUJER), Estrato socioeconómico, Gasto semanal promedio (en pesos), Longitud abdominal (en

cms), Tipo de colegio del cual se graduó, Horas dedicadas semanalmente a estudiar (sin incluir las horas de clase), Tiempo empleado para llegar a la Universidad (en min), Longitud promedio de los dos pies (en cms) y longitud promedio de las manos (en cms). Los datos se muestran a continuación:

```
MASA ESTATURA EDAD GENERO ESTRATO GASTO LONG_ABDO TIPO_COLE HORAS TIEMPO FUMA LONG_PIE LONG_MAN
76 181 6279 HOMBRE 2 80000 83 PUBLICO 21 60 NO 26.3 19.6
55 160 8760 MUJER 2 49167 50.0 PUBLICO 2 30 NO 20.0 17.6
64 166 6393 MUJER 4 100000 73 PUBLICO 20 80 SI 18.0 17.6
60 157 6205 HOMBRE 1 10000 67.3 PUBLICO 7 20 NO 27.0 16.8
68 166 8481 HOMBRE 3 45000 65 PRIVADO 10 15 NO 27.0 17.9
70 170 8605 HOMBRE 1 60000 90 PUBLICO 20 30 NO 28.0 18.3
60 150 6406 MUJER 3 40000 86 PUBLICO 21 90 NO 21.0 16
52 157.4 6938 HOMBRE 2 30000 70 PUBLICO 10 40 NO 25.0 16.0
65 159 6574.5 MUJER 5 60357 72.4 PRIVADO 12 15 NO 22.0 16
45.4 160 6371 MUJER 2 50000 60 PUBLICO 15 90 NO 22.0 16
52 167 6205 HOMBRE 2 20000 70 PUBLICO 10 40 NO 25.0 16.0
63 170 6265 MUJER 3 75000 73 PUBLICO 8 60 NO 25.0 16
50 153 6442 MUJER 2 50000 70 PUBLICO 8 65 NO 21.3 16.5
48 153 6484 MUJER 2 49167 64 PUBLICO 32 30 NO 23.6 16.7
43 152 6380 MUJER 3 45000 67 PUBLICO 12 60 NO 24.5 16.75
46 155 6060 MUJER 1 15000 65 PUBLICO 10 30 NO 23.0 17
45 156 7227 MUJER 4 60357 68 PUBLICO 8 15 NO 23.0 17
64 158 6603 MUJER 2 30000 78 PUBLICO 10 30 NO 23.0 17
51 163 7875 MUJER 5 70000 71 PUBLICO 20 30 SI 22.0 17
70 164 6178 MUJER 2 35000 80 PUBLICO 12 15 NO 24.0 17
50 165 6974 HOMBRE 3 60000 70 PUBLICO 28 60 SI 23.0 17.0
65 165 6918 HOMBRE 3 60000 73 PUBLICO 25 70 NO 23.3 17.0
60 171 6318 HOMBRE 3 20000 67.3 PRIVADO 0 20 NO 24.0 17.0
60 171 6316 HOMBRE 3 20000 95 PRIVADO 6 20 NO 24.0 17.0
54 170 5850 MUJER 2 49167 64 PUBLICO 20 25 NO 23.0 17.5
64 170 6438 MUJER 2 50000 68 PUBLICO 12 120 NO 25.0 17.5
62 172 6626 HOMBRE 4 47200 88 PRIVADO 18 30 NO 24.0 17.5
90 177 7323 HOMBRE 4 40000 96 PRIVADO 9 20 NO 27.5 17.5
65 170 6224 HOMBRE 4 29533 71.2 PUBLICO 15 20 NO 24.5 17.8
63 161 6261 MUJER 3 55000 77 PUBLICO 5 180 NO 23.5 18
61 163 7050 HOMBRE 1 99640 82 PUBLICO 10 50 SI 25.0 18.0
61 163.5 6938 HOMBRE 1 99640 85 PUBLICO 10 50 SI 24.0 18.0
68 167 8122 HOMBRE 3 60000 97 PRIVADO 10 15 NO 30.0 18.0
68 169 10585 HOMBRE 3 60000 72 PUBLICO 14 20 SI 23.0 18.0
56 173 7115 HOMBRE 3 35000 64.2 PUBLICO 25 20 NO 25.0 18.0
71 175 6545 HOMBRE 4 10000 92 PRIVADO 10 45 NO 25.0 18.0
63 178 6248 MUJER 2 80000 70 PUBLICO 28 90 NO 23.0 18
70 185 6244 HOMBRE 4 10000 81 PRIVADO 4 15 SI 27.0 18.0
70 185 6251 HOMBRE 4 30000 81 PRIVADO 21 15 SI 28.0 18.0
57 178 6378 HOMBRE 4 40000 65.0 PRIVADO 14 40 NO 25.1 18.3
68 170 8986 HOMBRE 3 160000 83 PUBLICO 32 50 NO 26.0 18.5
52.1 170 6205 MUJER 2 49167 62 PUBLICO 10 50 NO 24.5 18.5
90 171 8015 HOMBRE 1 230000 95 PUBLICO 40 15 SI 25.0 18.5
51 172 6160 HOMBRE 2 95000 72 PUBLICO 24 30 NO 24.5 18.5
68 179 6136 HOMBRE 5 27000 73 PRIVADO 9 25 NO 26.8 18.5
73 169 7572 MUJER 4 150000 97 PRIVADO 18 30 NO 27.0 19
60 175 6244 HOMBRE 3 45000 67.3 PUBLICO 15 20 NO 25.0 19.0
57 176 8006 HOMBRE 3 25000 65.0 PUBLICO 10 30 NO 25.5 19.0
60 178 6240 HOMBRE 3 30000 75 PRIVADO 8 30 NO 25.0 19.0
80 180 6654 HOMBRE 2 50000 83.0 PRIVADO 10 80 NO 26.0 19.0
99 182 7192 HOMBRE 3 40000 110 PUBLICO 10 40 SI 27.0 19.0
70 185 6345 HOMBRE 3 27000 70 PRIVADO 20 40 NO 24.0 19.0
70 185 6345 HOMBRE 3 50000 75.2 PRIVADO 7 40 NO 24.0 19.0
81 183 9302 HOMBRE 2 71875 82 PUBLICO 8 30 NO 25.1 19.0
84 178 6408 HOMBRE 3 51333 85 PRIVADO 5 50 NO 27.0 19.5
78 178 6348 HOMBRE 3 7000 85 PUBLICO 35 10 NO 27.7 19.5
59.3 172 6570 HOMBRE 2 71875 81 PUBLICO 15 90 NO 26.0 20.0
```

```
57 172 6504 HOMBRE 2 30000 73 R_ESPECIAL 9 25 NO 27.0 20.0
62 174 9685 HOMBRE 3 150000 80 PUBLICO 20 20 NO 26.0 20.0
88 180 7342 HOMBRE 3 30000 98 PRIVADO 4 10 SI 26.0 20.0
88 183 6309 HOMBRE 1 45000 86 PUBLICO 6 110 NO 27.0 20.0
60 185 6244 HOMBRE 2 71875 77 PUBLICO 13 45 NO 27.8 20.0
76 188 6500 HOMBRE 1 153200 79.8864 PUBLICO 3 50 NO 28.0 20.0
74 187 6251 HOMBRE 2 70000 82 PRIVADO 19 35 NO 28.0 21.0
80 170 6883 HOMBRE 2 200000 83.028 PUBLICO 20 20 NO 30.0 25.0
40 154 6075 MUJER 4 60357 57 PUBLICO 35 20 NO 33.0 26.0
82.5 182 7088 HOMBRE 5 10000 94 PRIVADO 34 15 NO 26.3 29.0
```

Para la variable GENERO. Se elabora una tabla de frecuencias, indicando el número de hombres y mujeres y el porcentaje que representa cada categoría.

GENERO	Frecuencia	Porcentaje
HOMBRE	46	68.66
MUJER	21	31.34

Para la variable Edad. Observe que esta variable aparece en una escala de razón, pero con valores enteros. Sin agrupar la información se tiene el siguente diagrama de barras en R:

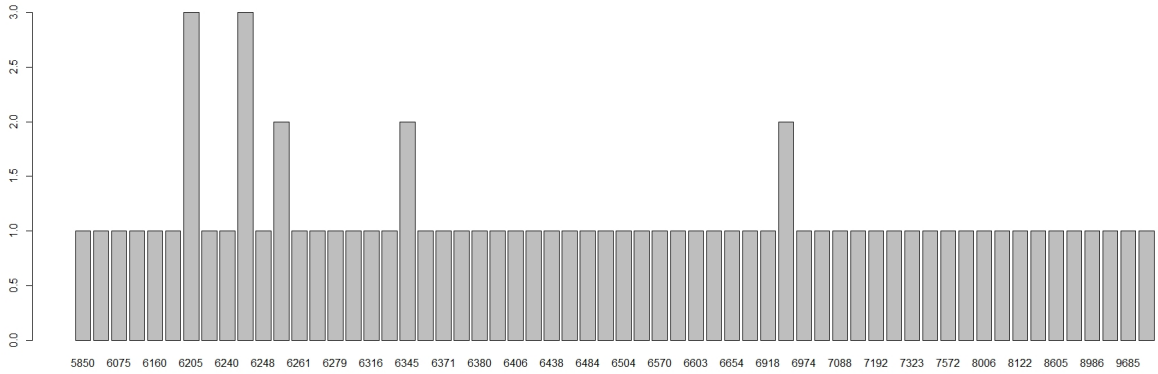

Fig. 1: Diagrama de Barras para Edad

Observe que claramente, este gráfico no permite evidenciar algún comportamiento de interés para la Variable. En este caso, se prefiere agrupar la información. Una manera es creando categorías o rangos.

1. Se agrupa la información por rangos (8 clases), elegidos a conveniencia.

Categoría	Frecuencia	Frec Rela
≥ 6300	20	0.299
(6300,6900]	24	0.358
(6900, 7500]	11	0.164
(7500, 8100]	4	0.060
(8100, 8700]	3	0.045
(8700, 9300]	2	0.030
(9300, 10000]	2	0.030
> 10000	1	0.015

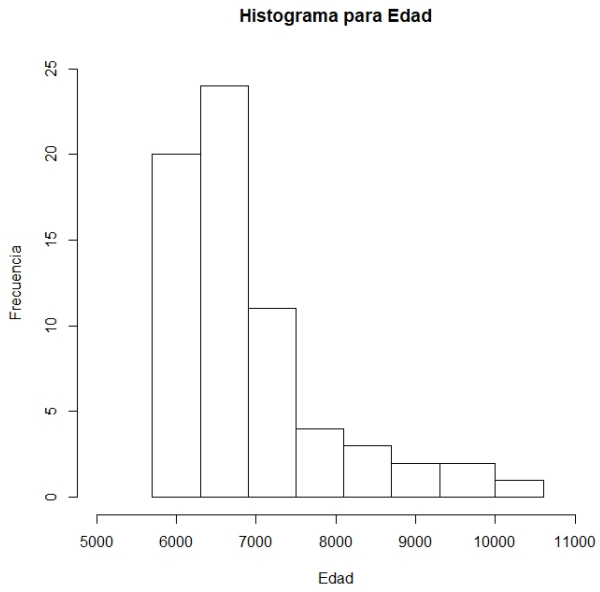

Fig. 2

2. Se agrupa la información por rangos con una regla específica.

Para la variable estatura. Usando la regla de Sturges se obtiene

$$K = 1 + 3.33\, log_{10}(n) = 1 + 3.33 * \ log_{10}(67) = 7.08 \approx 8 \ .$$

Así, se consideran 8 clases o intervalos. La mínima Estatura es 150 cms y la máxima es de 188 cms. El rango de las estaturas es $R = 38$. Si se asumen intervalos o clases de igual longitud, la amplitud para cada intervalo estará dada por:

$$A = \frac{Rango}{K} = \frac{38}{8} = 4.75 \ .$$

Para efectos de manejar un valor más simple como amplitud esta es redondeada a 5. Con esto se tiene que el rango a sido ampliado en 2 cms. (N Rango $= 8 \times 5 = 40$). La pregunta es como repartir este excedente. La mayoría de usuarios propone que se haga de manera equitativa, es decir, restar al mínimo la misma cantidad que se le suma al máximo. La figura 3 ilustra lo que se propone.

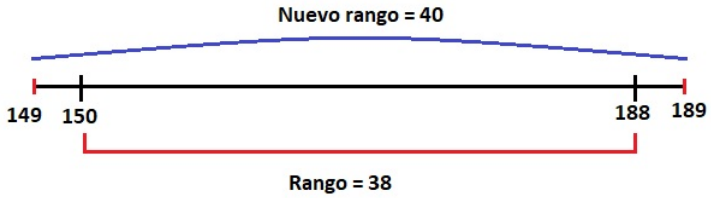

Fig. 3: Rango Ampliado para Estatura

Los intervalos de clase que se conforman son:

(149, 154], (154, 159], (159, 164], (164, 169], (169, 174], (174, 179], (179, 184], (184, 189] .

La respectiva tabla de frecuencias está dada por:

Intervalo	Frecuencia	Frec Rela	Marca
(149, 154]	5	0.075	151.5
(154, 159]	6	0.090	156.5
(159, 164]	7	0.104	161.5
(164, 169]	8	0.119	166.5
(169, 174]	17	0.254	171.5
(174, 179]	10	0.149	176.5
(179, 184]	7	0.104	181.5
(184, 189]	7	0.104	186.5

El gráfico resultante se ilustra en la figura 4. Los valores en medio de cada barra son los puntos medios de cada intervalo de clase o Marcas de Clase.

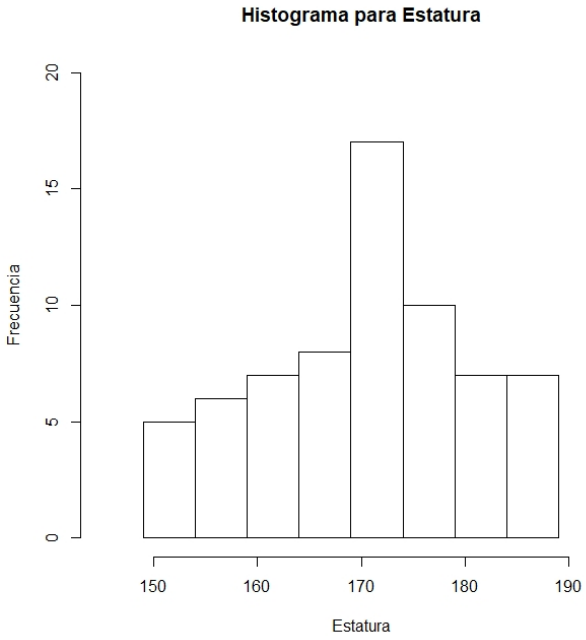

Fig. 4

Cuando solo se desea tener una idea gráfica del comportamiento de una variable, a veces no es tan necesario intervenir tanto en la construcción de un histograma. La mayoría de software estadístico, tienen reglas muy similares para la elección del número de clases o intervalos. Usando el paquete R se muestran diferentes gráficos para estas variables en la figura 5.

Aunque no es una regla general una tabla de frecuencias deberia poseer las siguientes caracteristicas:

1. UNIFORMIDAD: Clases de igual amplitud o de amplitud variable que dependen del tipo de datos.
2. UNICIDAD: Clases no traslapadas.
3. COMPLETEZ: Cada dato pertenece a una y sólo una clase.

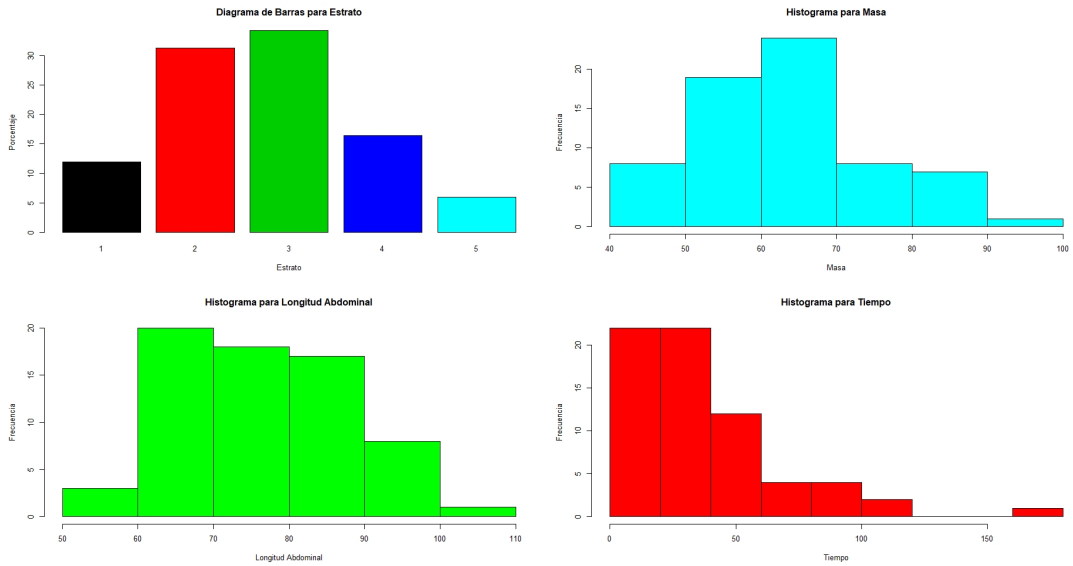

Fig. 5: Diagramas de Barras e Histogramas

Box-Plot o Diagrama de Cajas y Bigotes

Los diagramas de caja y bigotes son herramientas gráficas muy útiles para describir características importantes en un conjunto de datos, como son centro, simetría o asimetría, valores atípicos(raros), etc. La construcción de este diagrama emplea medidas descriptivas que son poco sensibles a datos extremos y por lo tanto presentan una descripción más clara de la información. Básicamente empleamos para su construcción los tres cuarteles, los valores mínimos y máximos y la media Muestral solo como medida de localización en el gráfico. Una observación se dice Atípica o Inusual si está a más de 1.5 veces el rango intercualtil de alguno de los cuarteles Q1 o Q3. Una observación se dice Atípica Extrema

si está a más de 3 veces el rango Intercuartil de alguno de los cuartiles $Q1$ o $Q3$. El diagrama está conformado por una caja la cual se construye con ayuda del primer y tercer cuartil. La mediana es dibujada en el interior de la caja al igual que la media muestral. Los bigotes se extienden desde los cuartiles a la derecha y a la izquierda. Su longitud depende de si hay o no datos atípicos. En la figura 6, se muestran dos tipos de boxplot.

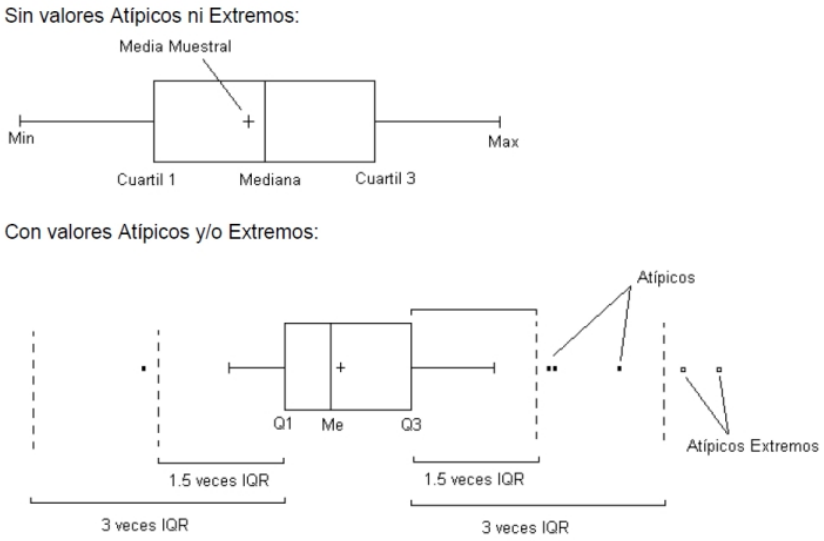

Fig. 6: Construcción de un Box-Plot

Los Box-plot, para las variables Estatura, Masa, Edad y Gasto, se muestran en la figura 7.

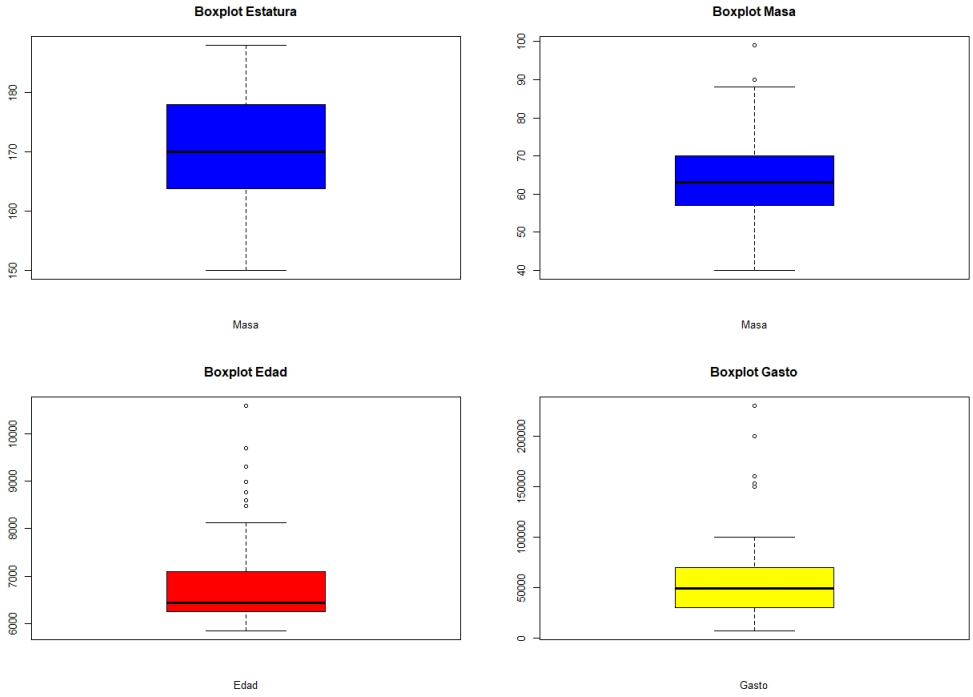

Fig. 7: Box-Plot para Estatura, Masa, Edad y Gasto

Análisis descriptivo de un conjunto de n datos

Suponga que se tienen n datos recopilados sobre una variable y que pueden representar los tiempos de duración de una batería para vehículo. Estos datos pueden ser los que aparecen a continuación:

2.2 3.4 2.5 3.3 4.7 4.1 1.6 4.3 3.1 3.8 3.5 3.1 3.4 3.7 3.2 4.5 3.3 3.6 4.4 2.6
3.2 3.8 2.9 3.2 3.9 3.7 3.1 3.3 4.1 3.0 3.0 4.7 3.9 1.9 4.2 2.6 3.7 3.1 3.4 3.5

Este conjunto de datos por si solo no muestra ninguna faceta interesante. A simple vista se puede apreciar un valor mínimo y un valor máximo y que hay algunos valores que se repiten. Por lo tanto es supremamente difícil tratar de determinar alguna característica de interés de la población de la cual provienen; si el número de datos aumenta es todavía más difícil detectar características importantes. Existen técnicas estadísticas que permiten extraer información que puede resultar de algún modo importante para tomar decisiones en un determinado momento.
Hay dos maneras de analizar estos datos:

1. **Datos agrupados**: Consiste básicamente en la conformación de clases de una cierta longitud donde la pertenencia de un dato a cada clase estará determinada por su valor. Con esta técnica es posible experimentar perdida de información.

2. **Datos sin agrupar**: Consiste en manipular los datos tal y como fueron recopilados.

Medidas numéricas en datos agrupados

Las medidas numéricas descriptivas se dividen en dos: **Medidas de localización** y **Medidas de dispersión**. En las medidas de localización se circunscriben las medidas de tendencia central.

Medidas de localización y de tendencia central

Estas medidas permiten cuantificar numéricamente, características de la población de la cual fueron tomados los datos. Entre las más comunes se encuentran:

La media muestral para datos agrupados

Es un valor que trata de representar el comportamiento promedio del conjunto de datos. Corresponde a una estimación de la media poblacional. En el caso de datos agrupados se define como:

$$\overline{X}_a = \frac{\sum_{i=1}^{k} m_i\, f_i}{n} = \frac{\sum_{i=1}^{\#celdas} \text{marca de clase} \times \text{Frecuencia de clase}}{\text{Total Frecuencias}}\,.$$

Como ejemplo, considere los datos de las Estaturas de los estudiantes del curso de primer semestre. A esta tabla le adicionamos una columna que contiene la frecuencias relativas acumuladas. La tabla resultante se muestra a continuación:

Intervalo	Frecuencia	Frec Rela	F Rel Acum	Marca
(149, 154]	5	0.075	0.075	151.5
(154, 159]	6	0.090	0.164	156.5
(159, 164]	7	0.104	0.269	161.5
(164, 169]	8	0.119	0.388	166.5
(169, 174]	17	0.254	0.642	171.5
(174, 179]	10	0.149	0.791	176.5
(179, 184]	7	0.104	0.896	181.5
(184, 189]	7	0.104	1.000	186.5

$$\begin{aligned}\overline{X}_a &= \frac{\sum_{i=1}^{8} m_i\, f_i}{67} \\ &= \frac{(151.5\times 5)+(156.5\times 6)+(161.5\times 7)+(166.5\times 8)+(171.5\times 17)+(176.5\times 10)+(181.5\times 7)+(186.5\times 79}{67} \\ &= 170.38\ .\end{aligned}$$

La moda muestral para datos agrupados

Es el valor que que presenta mayor frecuencia. Se define como la marca de clase del intervalo con mayor frecuencia absoluta. En el ejemplo anterior, se tiene que, $moda = 171.5$ cms.

Percentiles muestrales para datos agrupados

Los percentiles son aquellos valores abajo y arriba de los cuales se encuentra una cierta proporción de datos del conjunto. Por ejemplo, el percentil 10 es aquel valor tal que al menos el 10 % de los datos son inferiores a el y al menos el 90 % de los datos son superiores a el. Si la característica de interés está asociada a una variable X, el percentil $100p\,\%$, para $0 < p < 1$, suele denotarse por x_p. Otra manera de denotar un percentil, es a través del porcentaje que representa. Por ejemplo, el percentil 25, suele denotarse como P_{25}.

Para calcularlo se requiere la columna de frecuencias relativas acumuladas, que se obtiene de la tabla de frecuencias, usando la siguiente fórmula:

$$x_p = L + \frac{(p-a)\times h}{f}\ ,$$

donde:

- L: Límite inferior de la clase que contiene el percentil.
- n: Número de datos.

- f: Frecuencia relativa de la clase que contiene el percentil.
- a: Frecuencia relativa acumulada del intervalo anterior al del percentil.
- h: Longitud de la clase del percentil.

Para identificar la clase del percentil se identifica cual clase tiene una frecuencia relativa acumulada igual o superior a p.

Ejemplo
Usando los datos de estaturas calcule el P_{50}.
Se observa que en la columna de frecuencias relativas acumuladas el intervalo de clase donde esta frecuencia supera a 0.5 es el quinto intervalo, donde la frecuencia acumulada es 0.642, la cual excede a 0.5. Por lo tanto la clase del percentil será (169, 174]. Usando este intervalo se tiene que: $L = 169,\ n = 67,\ f = 0.254,\ a = 0.388,\ p = 0.5,\ h = 5$. Así:

$$P_{50} = 169 + \frac{(0.5 - 0.388) \times 5}{0.254} = 171.2 \ .$$

El 50 % de los estudiantes del curso tienen estaturas inferiores o iguales a 171.2 cms.

Los percentiles P_{25}, P_{50} y P_{75}, dividen los datos en cuatro partes porcentualmente iguales. Estos percentiles son llamados *Cuartiles* y se denotan Q_1, Q_2 y Q_3, respectivamente.

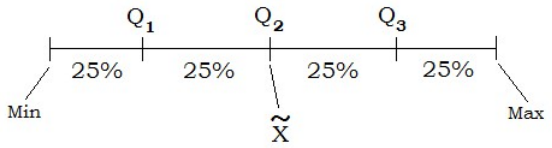

Fig. 8: Cuartiles

Por ejemplo, para el percentil 25 observe que el intervalo que contiene a $P25$ es (159, 164]. Así $L = 159$, $p = 0.25$, $a = 0.164$, $h = 5$ y $f = 0.104$. Con esto se tiene que:

$$P25 = 159 + \frac{(0.25 - 0.164) * 5}{0.104} = 163.1 \ .$$

El 25 % de los estudiantes del curso tienen estaturas inferiores o iguales a 163.1 cms.

La **Mediana** corresponde al percentil 50. Es usualmente denotada $\tilde{X}$. Su cálculo se realiza con el mismo procedimiento utilizado en la obtención de los percentiles.

Medidas de dispersión

Estas medidas permiten cuantificar numéricamente, que tan dispersos se encuentran los datos ya sea con respecto a la media o con respecto a las unidades de medición. Entre las más comunes se encuentran:

La varianza muestral para datos agrupados

Esta medida indica que tanto se alejan los datos respecto de la media muestral. Se denota S^2_{agrup}. Se calcula por medio de la siguiente fórmula:

$$S^2_a = \frac{\sum_{i=1}^{k} \left(m_i - \overline{X}_a\right)^2 \times f_i}{n-1} \, .$$

Donde m_i representa la marca de clase del intervalo i, f_i la frecuencia del intervalo i, k el número de intervalos de clase.

Rango intercuartil

Es la diferencia entre el percentil 75 y el percentil 25. Valores grandes quiere decir que el 50 % de los datos más centrales se encuentran muy dispersos.

$$\textbf{QRANGE} = Q_3 - Q_1 = P_{75} - P_{25}$$

Para los datos de estaturas se tiene: $S^2_a = 101.4$, con esto $S_a = 10.07$, los percentiles 25 y 75 son $P25 = 163.1$, $P75 = 177.6$. El rango intercuartil es QRANGE $= 177.6 - 163.1 = 14.5$.

Cálculo de medidas numéricas para datos no agrupados

Para el cálculo de estas medidas se consideran todos y cada uno de los datos, por lo cual la pérdida de información contenida en la muestra se reduce. También se dividen en dos: **Medidas de localización** y **Medidas de dispersión**. En las medidas de localización se circunscriben las medidas de tendencia central.

Medidas de localización y de tendencia central

Media muestral

Se define como la suma de todos los elementos de la muestra dividido por el tamaño de la muestra. Cuando la distribución de la cual provienen los datos es simétrica y no hay presencia de valores extremos, la media muestral es un buen representante del conjunto de datos. La media no necesariamente es un valor del conjunto de observaciones. Se denota con el símbolo $\overline{X}$. Se calcula con la siguiente fórmula,

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}, \quad i = 1, 2, \cdots, n$$

La media muestral puede verse como el punto de equilibrio de los datos.

Ejemplo
(La media es sensible a valores extremos). Considere los ingresos mensuales en pesos de 8 empleados públicos:
$500000, 750000, 600000, 550000, 700000, 550000, 550000, 600000$.
Calcule el ingreso mensual medio.
Solución

$$\overline{x} = \frac{500000 + \cdots + 550000 + 600000}{8} = 600000\ .$$

El ingreso de los empleados muestreados está alrededeor de los 600000 pesos. Suponga que un nuevo empleado es adiccionado a la lista y su ingreso es de 2000000. El ingreso promedio será $\bar{X} = 755555.6$. Observe que este valor es superior a la mayoría de las cifras del conjunto de datos. Esto se debe a que uno de los ingresos es muy grande en comparacion con los otros ingresos. Este valor no representa en gran medida al grueso de los ingresos.

Ejemplo
Se registra el número de tasas de café consumidas por un empleado de oficina en un período de 20 dias:

$$4\ \ 5\ \ 3\ \ 6\ \ 7\ \ 1\ \ 2\ \ 3\ \ 0\ \ 5\ \ 6\ \ 5\ \ 8\ \ 4\ \ 0\ \ 2\ \ 3\ \ 7\ \ 5\ \ 6$$

Calcule el número promedio de tasas de café.
Solución
$\bar{X} = \frac{82}{20} = 4.1$. Un empleado consume en promedio alrededor de 4 tazas por dia.

Ejemplo
Se registran las edades de 15 personas en un grupo. Estas son:
$18, 20, 19, 19, 21, 22, 20, 23, 21, 24, 19, 20, 22, 21, 24$ (en años). Calcule la edad promedio de las 15 personas.
Solución
La edad promedio de este grupo es : $\bar{X} = \frac{313}{15} = 20.86 \approx 20.9$. Si resumimos esta información en una tabla de frecuencia

Edad	18	19	20	21	22	23	24
frecuencia	1	3	3	3	2	1	2

La edad promedio se puede calcular como:

$$\bar{X} \ = \ \frac{1}{n} \sum_{i\,=\,1}^{n} f_i\, X_i\ .$$

Es un cálculo similar al caso de datos agrupados. Usando esta tabla se obtiene:

$$\bar{X} = \frac{18(1) + 19(3) + 20(3) + 21(3) + 22(2) + 23(1) + 24(2)}{15}$$
$$\bar{X} = \frac{313}{15} = 20.9$$
$$\bar{X} = \frac{\sum x_i f_i}{15} = \frac{\sum x_i f_i}{\sum f_i}$$

En el ejemplo anterior, suponga que otra persona adicional tiene una edad de 35 años. El cálculo de la edad promedio es: $\bar{X} = \frac{313+35}{16} = \frac{348}{16} = 21.8$.

Si la edad adiccional fuera 45 años, entonces $\bar{X} = \frac{313+45}{16} = 22.4$.

Si la edad fuera 55 años, entonces $\bar{X} = \frac{313+55}{16} = 23$.

Observe que la media muestral tiende a acercarse al valor extremo.

La moda en datos sin agrupar

Se define como el dato que presenta mayor frecuencia en la muestra. Para calcularla se recomienda ordenar las observaciones de menor a mayor. Es posible que un conjunto de datos no tenga moda o que tenga varias modas.

Ejemplo
Considere los siguientes datos ordenados de menor a mayor:
$500, 550, 550, 600, 700, 750, 750, 800, 900, 950$. Para esta muestra calcule la moda.
Solución
Se puede observar que el conjunto de datos tiene dos modas que son respectivamente: 550 y 750.

Percentiles para datos sin agrupar

Una aproximación a los valores de los percentiles se puede obtener por medio del siguiente algoritmo que muestra como se calcula el percentil de orden p, con $0 < p < 1$.

1. Ordene la muestra de menor a mayor
2. Calcule el percentil $100p\,\%$, x_p como

$$x_p = \begin{cases} \frac{X_{(n\,p)} + X_{(n\,p+1)}}{2} & \text{Si } n\,p \text{ es un número natural} \\ \\ X_{([[n\,p]]+1)} & \text{Si } n\,p \text{ no es un número natural} \end{cases},$$

El símbolo [[ ]] representa la función *Mayor Entero*.

La mediana en datos sin agrupar

La Mediana es un valor real que divide los datos en dos partes porcentualmente iguales. Obtenido dicho valor, el 50 % de los datos son inferiores o iguales a el y el restante 50 % lo supera. No es tan sensible como $\overline{X}$ a valores extremos. Se denota $\tilde{X}$. Para hallarla se deben ordenar los datos de menor a mayor. Suponga que se tiene el siguiente conjunto de datos $X_1, X_2, \ldots, X_n$, si se ordenan de menor a mayor se obtiene la siguiente sucesión $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$. $\tilde{X}$ se calcula por medio de,

$$\tilde{X} = \begin{cases} X_{\left(\frac{n+1}{2}\right)} & \text{Si } n \text{ es impar} \\ \\ \frac{X_{\left(\frac{n}{2}\right)} + X_{\left(\frac{n}{2}+1\right)}}{2} & \text{Si } n \text{ es par} \end{cases} .$$

Ejemplo
Considere los ingresos mensuales en dolares de 8 empleados públicos, 500, 750, 600, 550, 700, 2000, 55
La muestra ordenada es 500, 550, 550, 550, 600, 700, 750, 2000. Calcule la mediana.
Solución
Como n es par

$$\tilde{X} = \frac{X_{\left(\frac{n}{2}\right)} + X_{\left(\frac{n}{2}+1\right)}}{2} = \frac{X_{(4)} + X_{(5)}}{2} = \frac{550 + 600}{2} = 575 .$$

Este valor de la mediana es una medida más representativa que $\overline{X}$. El 50 % de los ingresos de los empleados son inferiores o iguales a 575 dólares.

Ejemplo
Considere los siguientes datos ordenados de menor a mayor:
$500, 550, 550, 600, 700, 750, 750, 800, 900, 950$. Halle el percentil 76 usando el método expuesto arriba.
Solución
La muestra ya esta ordenada, entonces usando el método anterior $n\,(0.76) = 7.6$ por lo tanto, $P_{76} = X_{(8)} = 800$. El 76 % de los empleados tienen ingresos inferiores o iguales a 800 dólares.

Para el ejemplo anterior, de las edades, calcular la mediana.
Solución
La mediana se calcula como: $\tilde{X} = X_{\left(\frac{15}{2}+1\right)} = X_{(8)}$. Ahora, como

18	19	19	19	20	20	20	21	21	21	22	22	23	24	24
X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}	X_{13}	X_{14}	X_{15} ,

entonces: $X_8 = 21$, es decir, $\tilde{X} = 21$. El 50 % de las personas encuestadas tienen edades inferiores o iguales a 21 años.

Se encuesta a otra persona y su edad resulta ser 30 años, la mediana en este caso es $\tilde{X} = \frac{X_{(8)} + X_{(9)}}{2} = \frac{21+21}{2} = 21$.

Si la edad de esa nueva persona es 50 años, se tiene que
$\tilde{X} = \frac{X_{(8)} + X_{(9)}}{2} = \frac{21+21}{2} = 21$. Es decir, la mediana no se vé afectada por datos muy extremos o atípicos.

Al igual que para datos agrupados, se pueden calcular los Cuartiles Q_1, Q_2, Q_3 y los percentiles.

Ejemplo
Para los datos de edades, calcule el primer cuartil y el percentil 60.
Solución
El primer cuartil es el percentil 25. Ahora $0.25\,(15) = 3.75$. El primer cuartil $Q_1 = X_{(4)} = 19$. El 25 % de las personas tienen edades inferiores o iguales a 19 años.
El percentil 60 . $(0.6)\ (15) = 9$. $P60 = \frac{X_{(9)}+X_{(10)}}{2} = 21$. El 60 % de las personas tienen edades inferiores o iguales a 21 años.

Medidas de dispersión

La varianza

La varianza permite evidenciar la dispersión de los datos con respecto a la media muestral. Valores grandes de la varianza indican una gran dispersión. Se denota por S^2. Se calcula con la siguiente fórmula:

$$S^2 = \frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}{n-1} .$$

Interpretar la varianza puede resultar un poco complicado ya que esta expresada en unidades cuadradas; por ejemplo, la varianza podría estar en minutos cuadrados o en kilogramos cuadrados. Por esta razón se acostumbra reportar la raíz cuadrada de la varianza, que recibe el nombre de *Desviación estándar*. Si por ejemplo un investigador toma mediciones de temperatura en una región durante cierto tiempo y al final reporta: 'Se observó una temperatura promedio de $28^{o}C$ con una desviación estándar de $1^{o}C$ 'quiere decir que algunas veces la temperatura puede bajar hasta $27^{o}C$ y algunas veces puede subir hasta $29^{o}C$.

Ejemplo
Para los datos de las edades, se tiene que:
$S^2 = \frac{\sum(X_i - 20.9)^2}{15-1} = 3.4095 \approx 3.41 \quad y \quad S = 1.8466 \approx 1.85$
Lo cual significa que la desviación promedio en cuanto a la media es de 1.85 años. En otras palabras, la mayoría de los estudiantes del curso tienen edades entre 19 y 23 años.

El rango intercuartil

Esta medida es la diferencia entre el percentil 75 y el 25. Mide que tan disperso está el 50 % de los datos más centrales. Se calcula así:

$$\text{RANGO INTERCUARTIL} = \text{Qrange} = Q3 - Q1 = P_{75} - P_{25} \ .$$

Coeficiente de variación

El *coeficiente de variación* que se define como : $C.V = \frac{S}{\overline{X}}$.
Es una fracción de la media muestral. Se usa para comparar la variabilidad de dos o más conjuntos de datos.

Ejemplo
Considere las siguientes medidas que se tomaron a dos poblaciones, una de hombres de 25 años y otra de niños de 11 añnos. Tales medidas son,

$$\begin{aligned} \overline{X}_{\text{adultos}} &= 66\ kgs \\ S_{adultos} &= 4.5\ kgs \\ \overline{X}_{\text{niños}} &= 36\ kgs \\ S_{\text{niños}} &= 4.5\ kgs \end{aligned}$$

Calcule el coeficiente de variación para los adultos y para los niños. Con los datos anteriores se puede observar que

$$\begin{aligned} C.V_{\text{adultos}} &= \frac{4.5}{66} = 0.0682 \\ C.V_{\text{niñnos}} &= \frac{4.5}{36} = 0.125 \end{aligned}$$

Se puede concluir que los pesos de los niños son más variables que los de los adultos.