18. Medidas resumen

Media, mediana, rango, desvío estándar, distancia intercuartil

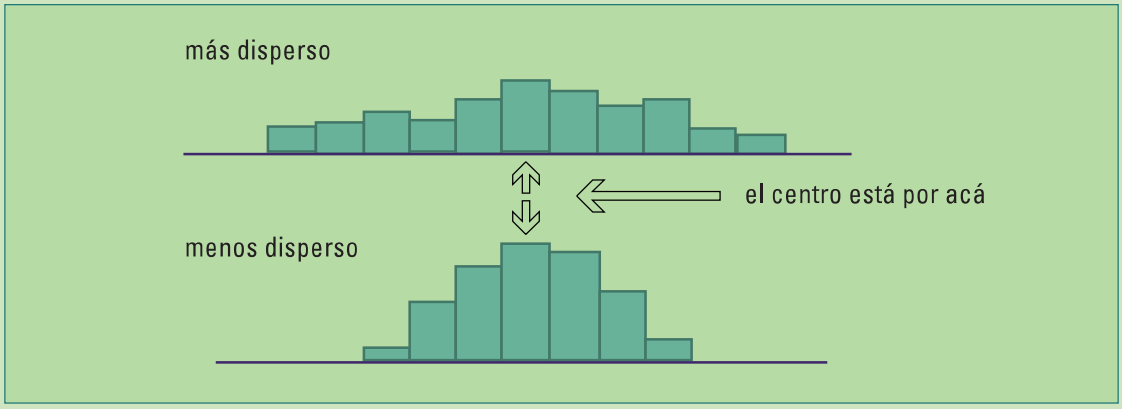
La mente humana puede captar la información que aportan diez números, cien es difícil y con mil, estamos perdidos. Por esa razón, es muy importante contar con pocos valores (medidas resumen), que de alguna manera deben describir las características más sobresalientes del conjunto que se está analizando.

Una medida resumen es un número. Se obtiene a partir de una muestra y, en cierta forma, la caracteriza. Es el valor de un estadístico. Por ejemplo, un porcentaje o una proporción son medidas resumen. Se utilizan con datos categóricos o con datos numéricos categorizados previamente. **Las medidas resumen permiten tener una idea rápida de cómo son los datos.** Pero, un estadístico mal utilizado puede dar una idea equivocada respecto de las características generales que interesa mostrar.

El cálculo de medidas resumen es el primer paso; se realiza cuando se recolectan los datos en un estudio para tener una idea de qué está pasando. Posteriormente, los investigadores pondrán a prueba sus hipótesis respecto a algún parámetro poblacional, estimarán características de la población y estudiarán posibles relaciones entre las variables. Cuando presentan sus conclusiones al público en general, las medidas resumen muestran los resultados en forma concisa y clara, volviendo a tener importancia.

En principio, se pueden obtener muchísimas formas de resumir los valores de un conjunto de datos numéricos. Es importante que sean fáciles de interpretar.

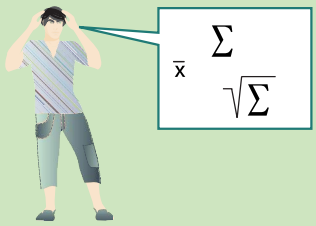
Cualquier conjunto de datos tiene **dos propiedades importantes:** un **valor central** y la **dispersión** alrededor de ese valor. Vemos esta idea en los siguientes histogramas hipotéticos:



Describiremos en este capítulo medidas de la posición del centro, la dispersión y otras medidas de posición. Veremos:

- Cómo se utilizan, en forma correcta o errónea.
- Qué significan.
- Qué dicen y qué no dicen estas medidas resumen.
- Cómo dependen de la distribución general de los datos.

Pero, a partir de ahora, además de gráficos necesitamos fórmulas.



Supongamos que tenemos un conjunto con n observaciones (datos), los representamos así:

$$x_1, x_2, x_3, \ldots, x_n$$

Se leen equis uno, equis dos, ..., equis ene y se pueden representar en una tabla:

(Número de) Observación	1	2	3	....	n
Valor	x_1	x_2	x_3	....	x_n

Ejemplo 18.1: Le preguntamos a 5 personas (n =5) cuántas cuadras camina por día y obtenemos.

Observación	1	2	3	4	5
Valor	4	15	8	31	17

Luego $x_1 = 4$, $x_2 = 15$, $x_3 = 8$, $x_4 = 31$, $x_5 = 17$

¿Cuál es el centro de estos datos? Respondemos esta pregunta en la siguiente sección.

□ 18.1. Posición del centro de los datos

El **promedio** define el valor característico o central de un conjunto de números. Existen varios métodos para calcular el promedio. El método utilizado puede influir en las conclusiones. Cuando vemos un anuncio con la palabra promedio, debemos alertarnos porque quien lo ha escrito, probablemente eligió el método de cálculo para producir el resultado que le interesa marcar.

Veremos con detalle las dos formas principales para obtener un valor central o promedio:

- **La media:** Se obtiene sumando todos los valores del conjunto de datos y dividiendo la suma por la cantidad de datos en ese conjunto.
- **La mediana:** Es el valor central del conjunto de datos ordenados.

18.1.1. La media

La media se representa por $\bar{x}$ (equis raya o equis barra). Se obtiene sumando todos los datos y dividiendo por la cantidad total n de observaciones,

$$\bar{x} = \frac{\text{SUMA DE LOS DATOS}}{n}$$

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

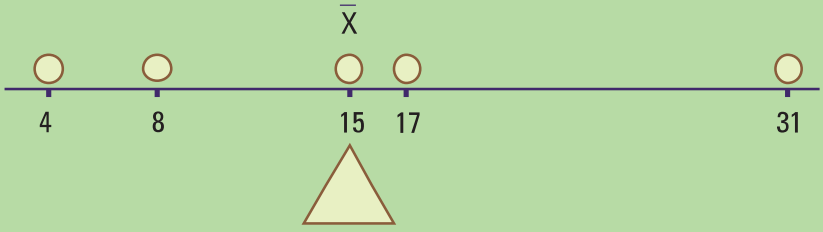
En el ejemplo anterior, la media de las cuadras caminadas por día es 15:

$$\bar{x} = \frac{4 + 15 + 8 + 31 + 17}{5}$$

$$\bar{x} = \frac{75}{5}$$

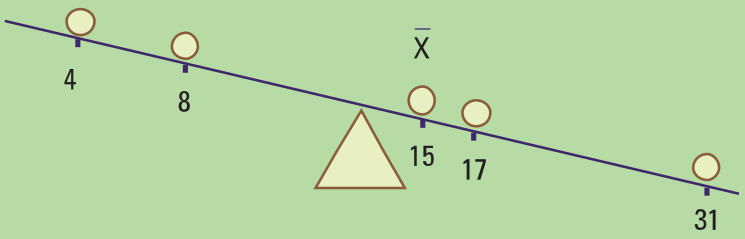
$$\bar{x} = 15 \text{ CUADRAS}$$

Si sobre una vara numerada sin peso, se colocan pesos idénticos sobre el valor de cada dato, la **vara queda en equilibrio** cuando se la apoya en el punto correspondiente a la media.



La vara está en equilibrio

La vara no queda en equilibro si se la apoya en cualquier otro punto.



La vara está no en equilibrio

Existe una abreviatura para la suma $x_1+x_2+...+x_n$. Se trata de la letra griega **sigma mayúscula** (comúnmente llamada **sumatoria**): $\sum$

En vez de la suma $x_1+x_2+...+x_n$ escribimos $\sum_{i=1}^{n} x_i$

y lo leemos como: "la suma de equis i, con i variando desde 1 hasta n".

Repito diez veces

$\sum_{i=1}^{n} x_i$ "La suma de x_i, con **i** variando desde 1 hasta **n**"

Así, la media de un conjunto de datos x_i es:

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

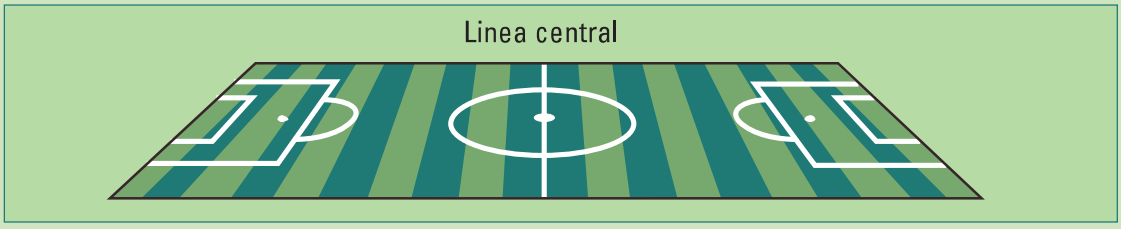
En el ejemplo 16.5, para los pesos de los 101 alumnos de 3 divisiones de 4to. Año, el peso medio es 58,90 kg:

$$\sum_{i=1}^{101} \frac{x_i}{101} = \frac{5949}{101}$$

$$= 58,90 \text{ kg}$$

18.1.2. La mediana

La **mediana** es otro tipo de centro. Es el punto central de los datos, como la línea central que divide el campo de juego de fútbol en dos partes iguales.



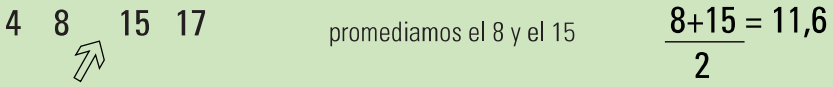
La mediana deja la misma cantidad de datos a cada lado.

Para hallar la mediana del conjunto de datos (4, 15, 8, 31, 17) del ejemplo 18.1:

- Primero los ordenamos de menor a mayor (4, 8, 15, 17, 31).
- Luego, la mediana es el valor central (15).

Para las cuadras que caminan por día las cinco personas elegidas al azar, el valor central, la mediana, es15. Quedan dos datos a cada lado de la mediana. En este ejemplo, la media coincide con la mediana, pero puede no ocurrir.

Si la cantidad de datos es **par** (4, 15, 8, 17) no hay una observación central, sino **un par** de **observaciones centrales** (8 y 15). La mediana (11,6) es el promedio de estos dos valores.



La regla general para calcular la mediana de n datos ordenados es:

- Si la **cantidad de datos** es **impar,** la mediana es el valor del centro, se encuentra en la posición (n+1)/2.
- Si la **cantidad de datos** es **par,** la mediana es el promedio de los dos valores centrales, se encuentran en las posiciones n/2 y (n/2)+1

Para los datos de los pesos de los 101 alumnos (ejemplo 16.5) la mediana es 58 kg. Como ya hemos construido el diagrama tallo hoja ordenado, la obtenemos directamente contando desde el dato más pequeño hasta el dato en la posición 51 (51=101+1)/2):

3 |
3 | 78
4 | 2334
4 | 566788888
5 | 000000111112222222223444444
5 | 5556677788**8**99 ← Aquí se encuentra la mediana
6 | 011223333444
6 | 55566666777777788899999
7 | 00112234
7 | 99
8 | 1
8 | 5

Pruebe contar 51 desde el dato más grande hacia los más chicos; la mediana también da 58.

18.1.3 ¿Por qué utilizamos más de una medida de posición del centro de los datos?

Cada una de las dos medidas presentadas tiene ventajas y desventajas.

La media utiliza todos los datos para su cálculo. Si los datos presentan un histograma simétrico calcular **la media es lo mejor** para obtener el centro de los datos, en este caso la mediana será muy parecida.

Siguiendo con el ejemplo 18.1 (cuadras que caminan por día 5 personas) la media y la mediana coinciden.

4 8 (15) 17 31 mediana = media

La mediana no se verá afectada si los datos presentan algún **valor atípico (316),** es decir, un dato alejado del patrón general (también llamado **outlier** en inglés), mientras que la media sí.

4 8 15 17 316 ⇐ valor atípico

El outlier puede ocurrir si una de las personas entrevistadas tiene hábitos diferentes a lo habitual (316 en lugar de 31), o si cometimos un error. La mediana seguirá siendo 15, pero la media será 72. ¿Es razonable decir que 72 cuadras por día en promedio representan las distancias caminadas por la mayoría de las personas?

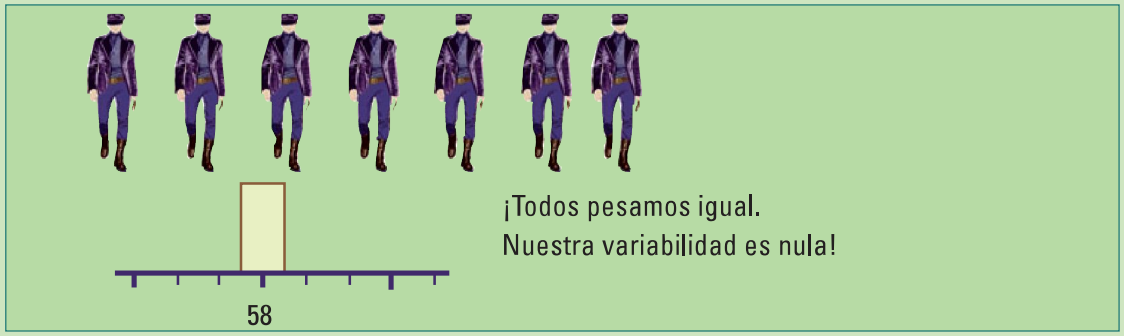
4 8 15 17 316

mediana ⇗ media = 72 ⇗

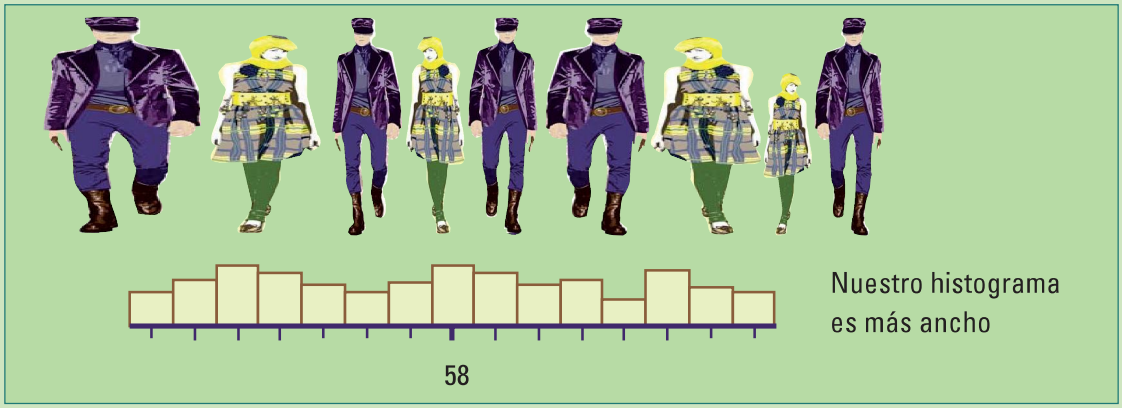
La media (72) ya no representa a la mayoría de los datos, por eso, decimos que **la media es sensible ante la presencia de valores atípicos (outliers).**

□ 18.2. Medidas de dispersión o variabilidad

Si todos los alumnos pesaran 58 kg, tendríamos un conjunto de datos iguales.



Otro conjunto de alumnos con mediana igual a 58kg podría tener pesos diferentes y los datos estarían más dispersos.



Además de conocer el punto central de un conjunto de datos, también nos interesa describir su dispersión, es decir cuán lejos tienden a estar los datos de su centro.

La variabilidad está presente en todos los conjuntos de datos. Sea cual fuere la característica, es casi imposible que dos mediciones sean idénticas. Esto se debe a que:

- Diferentes individuos tienen diferentes características (peso, altura, inteligencia, glóbulos rojos en sangre), al cuantificarlas resultan en valores diferentes de las variables correspondientes.
- Diferentes mediciones de una misma característica dan como resultado diferentes valores debido al inevitable error de medición.

Los métodos estadísticos son imprescindibles para analizar los datos debido a su variabilidad. El truco consiste en tener medidas que la capten de la mejor manera posible.

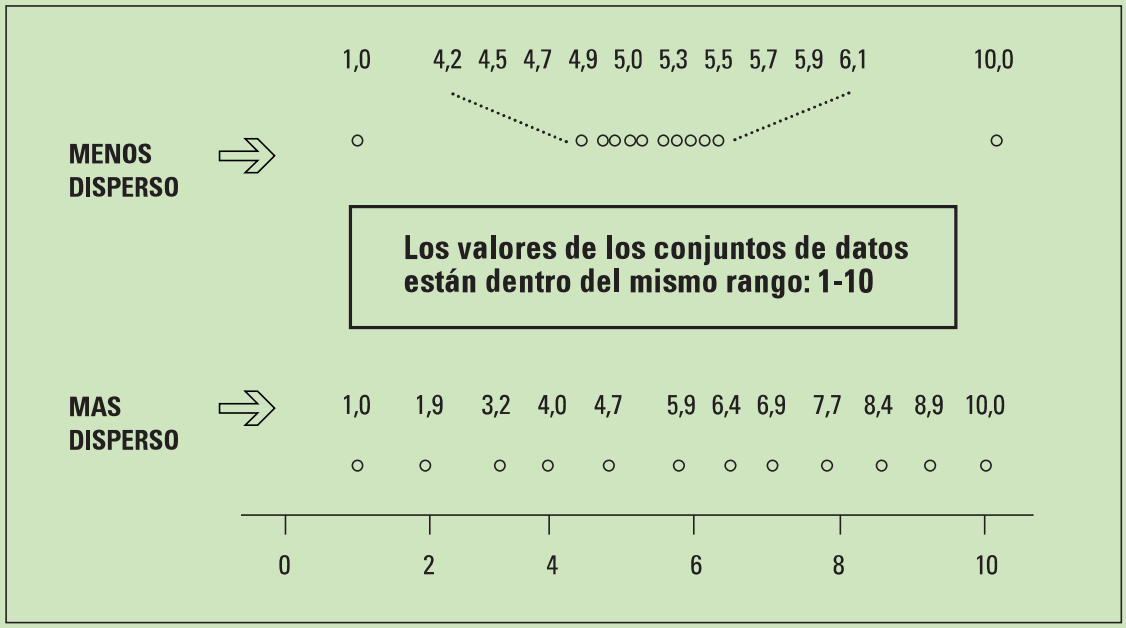
18.2.1. Rangos y distancia intercuartil

El rango de valores donde se encuentran los datos permite apreciar su variabilidad o dispersión (cuán desparramados están).

La medida natural para evaluar dicha dispersión es la distancia entre el valor mínimo y el valor máximo de los datos (máximo-mínimo).

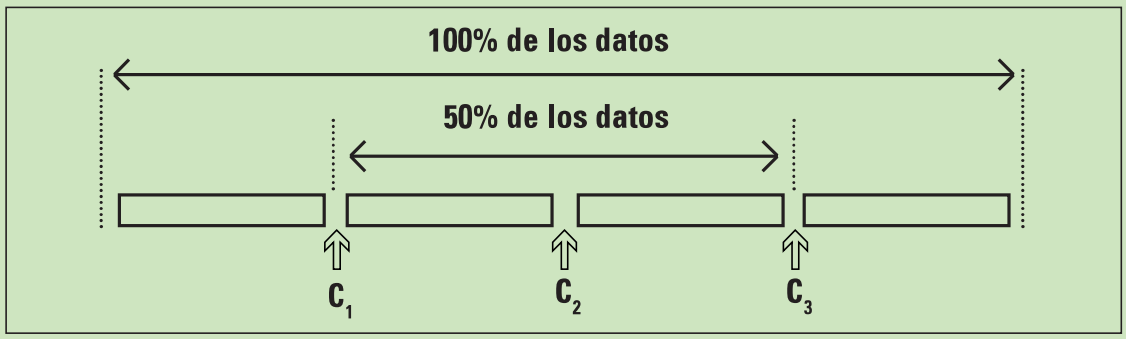
Tiene algunos inconvenientes:

- Es muy sensible a la presencia de valores atípicos.
- Como utiliza sólo dos datos, no puede distinguir dos conjuntos con máximos y mínimos coincidentes, pero uno tendrá la mayoría de sus valores mucho más concentrados que el otro.



La figura representa a los siguientes conjuntos de datos {1,0 4,2 4,5 4,7 4,9 5,0 5,3 5,5 5,7 5,9 6,110,0 } y {1,0 2,9 3,5 4,0 4,7 5,9 6,4 6,9 7,7 8,4 8,9 10,0}. La mayoría de los valores del primer conjunto están más concentrados que la mayoría del segundo conjunto pero tienen el mismo rango. El rango en este caso no distingue dos conjuntos de datos con diferentes dispersiones.

Para corregir los problemas se utiliza la distancia entre el valor mínimo y el valor máximo del 50% central de los datos, llamada distancia intercuartil.



¿Cómo se calcula la distancia intercuartil?:

1. Se ordenan los datos.
2. Se calcula la mediana (C_2), que los divide en 2 partes con igual cantidad de datos de cada lado.
3. Se calcula la mediana de la mitad más baja (grupo inferior), es el cuartil inferior (C_1)
4. Se calcula la mediana de la mitad más alta (grupo superior), es el cuartil superior (C_3)
5. La distancia intercuartil (DIC) es la diferencia entre el cuartil superior y el cuartil inferior: **$DIC = C_3 - C_1$**

Cuando la mediana coincide con uno de los datos se la puede considerar parte de los dos grupos, el superior y el inferior (esta regla es arbitraria y algunos autores no la cuentan en ninguno de los dos).

¿Qué mide la distancia intercuartil?
Como medida de dispersión, la distancia intercuartil mide la longitud del intervalo en el cual se encuentra el 50% central de los datos. Cuanto más dispersos estén los datos, mayor será la distancia intercuartil.

Nuevamente, consideremos los pesos de los 101 alumnos (ejemplo 16.5). La mediana está en la posición 51 y vale 58 kg. Para hallar el cuartil inferior calculamos la mediana de los 51 valores más chicos. Se encuentra en la posición (51+1)/2=26. Contamos 26 lugares desde los más chicos y obtenemos el valor 51 kg del cuartil inferior.

> **No confundir la posición 51** (donde se encuentra la mediana) **con 51 kg,** el valor del cuartil inferior que se encuentra en la posición 26.

Contando 26 lugares desde los **valores más altos** obtenemos el valor 67 kg del cuartil superior.

La distancia intercuartil se obtiene como la diferencia entre el cuartil superior y el cuartil inferior (DIC = 67 kg - 51 kg = 16), es la diferencia entre la mediana de los alumnos más pesados y la mediana de los más livianos. El 50% de los pesos difieren a lo sumo en 16 kg. El 50% de los pesos están entre 51 kg y 67 kg.

```
3 |
3 | 78
4 | 2334          Cuartil inferior
4 | 566788888  ↓
5 | 000000111112222222223444444
5 | 55566777888999  ← Aquí se encuentra la mediana
6 | 011223333444
6 | 55566666777777788899999
7 | 00112234   ↑
7 | 99            Cuartil superior
8 | 1
8 | 5
```

La mediana esta en la posición 51 y tiene un valor de 58 kg.
El cuatil inferior se encuentra en la posición 26 y tiene un valor de 51 kg.

No confundir la posición de un dato con el valor de un dato.

18.2.2. Los cinco números resumen y el gráfico de caja y brazos

El mínimo, el cuartil inferior, la mediana, el cuartil superior y el máximo son cinco números. Dan una idea de cómo está distribuido un conjunto de datos. Se los llama los cinco números resumen y se los representa por:

$$\text{Mínimo } C_1 \;\; M \;\; C_3 \text{ Máximo}$$

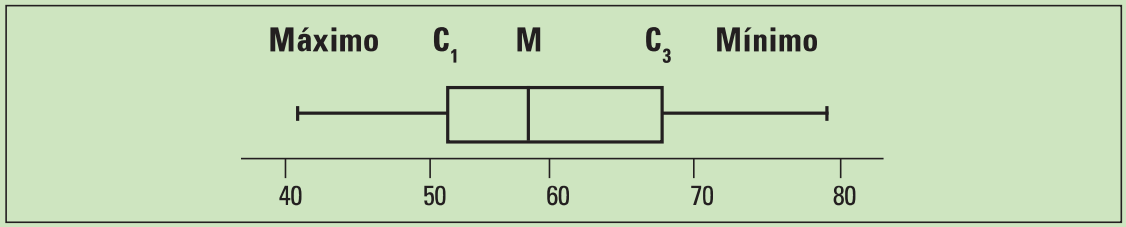
El 50% de los datos se encuentran entre el cuartil inferior y el superior.

Los cinco números resumen de los pesos de los alumnos de 4to. año son:

Mínimo	C_1	M	C_3	Máximo
37	51	58	67	85

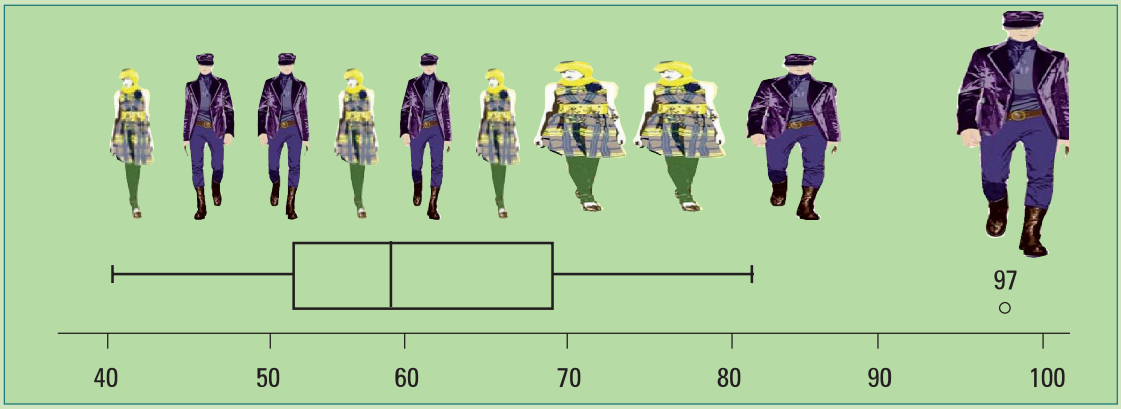
El 50% de los alumnos tiene un peso entre 51 y 67 kg.

Los cinco números resumen se representan gráficamente en un Gráfico de caja (Box-plot).

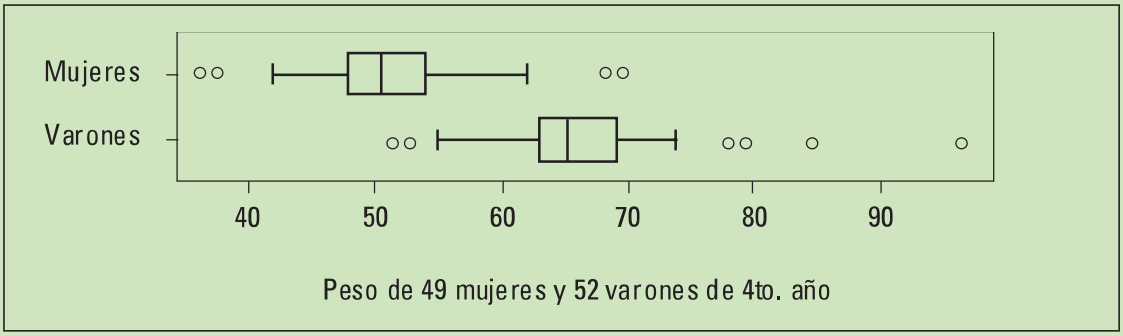


Los cuartiles forman los bordes de la caja y la mediana está dentro de la caja. Dos líneas - los brazos- se extienden, una desde cada borde de la caja, hasta el dato con valor máximo y mínimo respectivamente, mientras no sean valores atípicos (es decir, se encuentren dentro de 1,5 DIC).

Si agregamos un peso de 97 kg a los datos de los pesos, el boxplot muestra un valor atípico.



Los gráficos caja sirven especialmente cuando queremos comparar varios conjuntos de datos. En el ejemplo de los pesos, comparemos los de varones y de mujeres.



Peso de 49 mujeres y 52 varones de 4to. año

Entre las mujeres hay 2 que pesan menos que la mayoría y otras 2 más (por fuera de los brazos). Entre los varones se detectan 2 en los valores menores y 4 en los valores mayores. El 75% de las mujeres son más livianas que los hombres (excluyendo los 2 valores atípicos bajos de los hombres). Los **cinco números resumen** muestran los detalles:

Peso de Mujeres

Mínimo	Cuartil inferior - C_1	Mediana	Cuartil superior - C_3	Máximo
37	48	51	54	70

Peso de Varones

Mínimo	Cuartil inferior - C_1	Mediana	Cuartil superior - C_3	Máximo
51	63	66	69	97

18.2.3. Desvío estándar

La descripción de una distribución mediante medidas resumen es utilizada desde hace muchísimos años. Pero, la propuesta de utilizar los 5 números resumen es relativamente nueva. Fue hecha por John Tukey por los años 70, cuando comenzaban a utilizarse las computadoras.

La mediana y los cuartiles son muy sencillos de calcular a mano cuando la cantidad de datos es relativamente pequeña. Cuando se tienen muchos datos, la dificultad se encuentra en ordenarlos. Por esa razón, aunque la mediana era conocida casi no se utilizaba antes del advenimiento de las computadoras.

La media, es mucho más **fácil de calcular a mano** cuando hay muchos datos. Sólo requiere del uso de operaciones aritméticas, para hallar un número representativo de la mayoría de los datos.

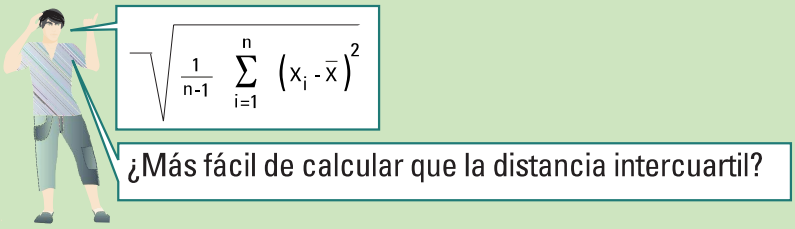
El **desvío estándar** es una **medida** de dispersión **basada en la media** y **utiliza todos los datos.** Durante muchos años la **media y el desvío estándar** fueron, y tal vez sigan siendo, las **medidas resumen más utilizadas.**

El desvío estándar representa una distancia típica de cualquier punto del conjunto de datos a su centro (medido por la media). Es una distancia promedio de cada observación a la media.

El desvío estándar de los datos de toda una población (desvío estándar poblacional) se denota con la letra griega σ (sigma minúscula). Pero la mayoría de las veces los parámetros poblacionales son desconocidos. ¿Qué se hace? Se calcula un estimador (s, desvío estándar muestral) utilizando una muestra.

La distinción entre el desvío estándar poblacional y el desvío estándar muestral vale para todos los estadísticos descriptos (media, mediana, cuartiles, distancia intercuartil, etc.). Tal como vimos en los capítulos 9 y 10, si el cálculo de un estadístico se realiza utilizando una muestra para estimar un parámetro, el resultado tendrá un error de muestreo.

¡Desvío estándar!



El desvío estándar se calcula promediando la diferencia entre cada dato y la media, elevadas al cuadrado. Como este resultado tiene las unidades al cuadrado, luego se saca la raíz cuadrada.

Para un conjunto de n datos:

1. Se calcula la distancia de cada dato a la media: $x_i - \bar{x}$
2. Se eleva al cuadrado: $(x_i - \bar{x})^2$
3. Se promedie dividiendo por n-1 y, así, se obtiene la varianza muestral

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

4. Por último se calcula la raíz cuadrada

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

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

Para el conjunto de datos de las cuadras que 5 personas caminan por día, del ejemplo 18.1 (x_1=4, x_2=15, x_3=8, x_4=31, x_5=17, n=5 y $\bar{x}$=5), la varianza muestral es 107,5 cuadras^2:

$$s^2 = \frac{(4-15)^2 + (15-15)^2 + (8-15)^2 + (31-15)^2 + (17-15)^2}{(5-1)}$$

$$s^2 = \frac{121 + 0 + 49 + 256 + 4}{4}$$

$$s^2 = 107,5$$

Cuanto más grande es la varianza muestral, más dispersos están los datos. Una medida de dispersión debe tener las mismas unidades que los datos.

La varianza muestral, en nuestro ejemplo está en cuadras al cuadrado, entonces por supuesto, debemos sacar la raíz cuadrada.

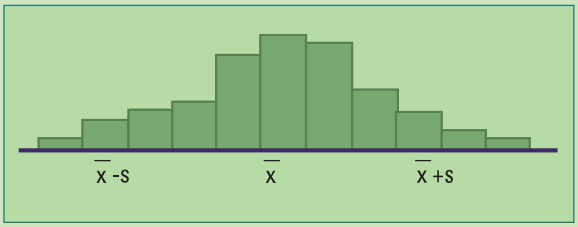
El desvío estándar es 10,37 cuadras:

$$s = \sqrt{107,5}$$

$$s = 10,37$$

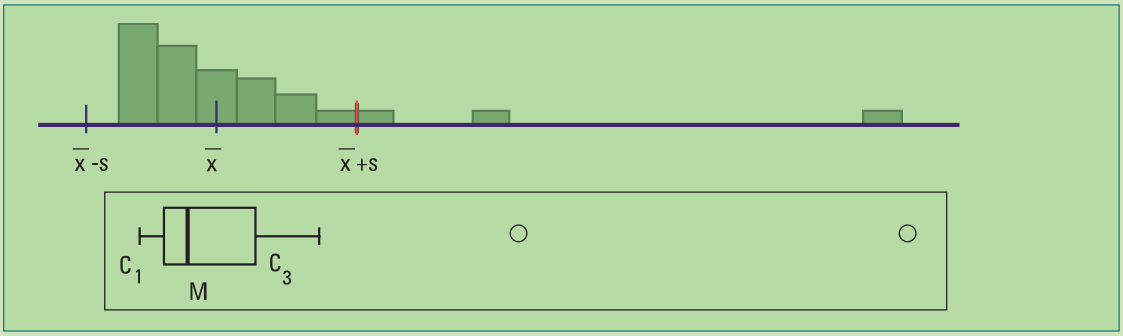
18.3. Centro y dispersión en diferentes tipos de distribuciones

La media y el desvío estándar son muy buenos para resumir datos con histogramas razonablemente simétricos y sin valores atípicos.



Sin embargo, la media y el desvío estándar no son una buena representación de la distribución de datos cuando tienen **valores atípicos** o sus **histogramas son asimétricos.**

La figura siguiente muestra el histograma de un conjunto de datos con distribución **asimétrica a derecha.** En este caso, **es mejor utilizar** la mediana y la distancia intercuartil, y mejor aún, **los 5 números resumen.**



El intervalo con extremos en $\bar{x}$-s y $\bar{x}$+s **no es una buena representación de los datos:** $\bar{x}$-s se encuentra fuera del rango de los valores observados (está a la izquierda del valor más pequeño) y quedan valores a la derecha de $\bar{x}$+s. El gráfico caja (boxplot) describe más precisamente el rango donde se encuentran los datos. El rango intercuartil que forma la caja contiene el 50% de los datos y los brazos se extienden hasta el último dato de cada lado. Se distinguen dos datos atípicos (en inglés: outliers, significa: yacen fuera).

En el ejemplo siguiente mostramos cómo las medidas resumen pueden contar una parte muy parcial de la historia.

Ejemplo. "Admítelo una salchicha no es una zanahoria". Así decía la revista "El Consumidor" en un comentario sobre la baja calidad nutricional de las salchichas.
(Introduction to the practice of Statistics Moore mc Cabe pág. 28).

Hay tres tipos de salchichas:
1. carne vacuna,
2. mezcla (carne porcina, vacuna y de pollo)
3. pollo.

¿Existe alguna diferencia sistemática entre estos tres tipos de salchichas, en estas dos variables? Mirar directamente los datos sirve de muy poco.

CALORÍAS Y SODIO EN SALCHICHAS POR TIPO. TABLA 18.1

Vacuno		Mezcla		Pollo	
Calorías	**Sodio**	**Calorías**	**Sodio**	**Calorías**	**Sodio**
186	495	173	458	129	430
181	477	191	506	132	375
176	425	182	473	102	396
149	322	190	545	106	383
184	482	172	496	94	387
190	587	147	360	102	542
158	370	146	387	87	359
139	322	139	386	99	357
175	479	175	507	170	528
148	375	136	393	113	513
152	330	179	405	135	426
111	300	153	372	142	513
141	386	107	344	86	358
153	401	195	511	143	581
190	645	135	405	152	588
157	440	140	428	146	522

Vacuno		Mezcla		Pollo	
Calorías	Sodio	Calorías	Sodio	Calorías	Sodio
157	440	140	428	146	522
131	317	138	339	144	545
149	319				
135	296				
132	253				

Comparemos la cantidad de calorías entre los tres tipos de salchichas utilizando gráficos caja. Recordemos que están basados en los números resumen:

	Mínimo	Cuartil inferior C_1	Mediana	Cuartil superior C_3	Máximo
Vacuno	111	140,5	152,5	178,5	190
Mezcla	107	139	153	179	195
Pollo	86	102	129	143	170

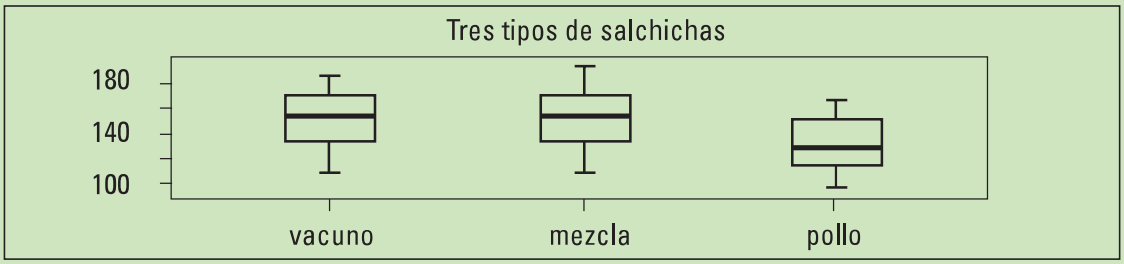


Figura 18.1. *Gráficos caja de la cantidad de sodio de tres tipos de salchichas.*

Vacuno	Mezcla	Pollo
8 |	8 |	8 | 67
9 |	9 |	9 | 49
10 |	10 | 7	10 | 226
11 | 1	11 |	11 | 3
12 |	12 |	12 | 9
13 | 1259	13 | 5689	13 | 25
14 | 1899	14 | 067	14 | 2346
15 | 2378	15 | 3	15 | 2
16 |	16 |	16 |
17 | 56	17 | 2359	17 | 0
18 | 146	18 |	18 |
19 | 00	19 |	19 |

Figura18.2. *Diagramas tallo hoja de la cantidad de sodio de tres tipos de salchichas. La coma decimal se encuentra un dígito a la derecha de la barra vertical* (|).

Vemos una tendencia general en las salchichas de pollo a presentar menor cantidad de calorías. Pero nos perdemos los detalles.

Los diagramas tallo hoja de las salchichas de carne vacuna y mezcla (figura 18.2) muestran la presencia de 2 grupos, y un valor aislado en la cola inferior. Sin embargo, como cada cuartil se encuentra aproximadamente en el centro de cada uno de los dos grupos, la distancia intercuartil refleja la distancia entre los grupos y, por lo tanto, el valor inferior no es detectado como dato atípico.

Analicemos ahora la distribución de la **cantidad de sodio** en las salchichas de **pollo** (tabla 18.1), cuyo diagrama tallo hoja tenemos a continuación

```
3 | 666.889
4 | 033
4 |
5 | 11.234
5 | 589
```

Tanto el diagrama tallo hoja como el histograma (figura 18.3) revelan la presencia de dos grupos:

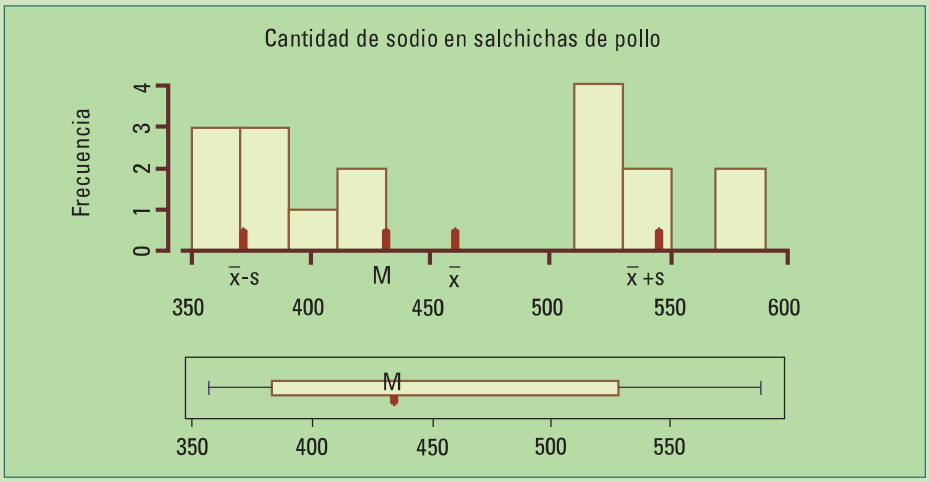


Figura18.3. *Histograma y gráfico caja de la cantidad de sodio de salchichas de pollo de diferentes marcas.*

Los valores ordenados de la cantidad de sodio en salchichas de pollo son:
357 358 359 375 383 387 396 426 430 513 522 528 542 581 588

La media (449,66) se encuentra fuera de los datos, la mediana (426) cerca del borde de uno de los dos grupos. El intervalo ($\bar{x}$-s, $\bar{x}$+s) no es una buena representación de los datos y el gráfico caja tampoco.

Recomendamos realizar gráficos caja fundamentalmente para comparar la distribución de varios conjuntos de datos. Un diagrama de tallo y hojas o un histograma son mejores para analizar la distribución de datos de una única variable. Generalmente, los detalles agregan poco, pero es importante estar preparados para las ocasiones en que sí agregan mucho.

El significado de las medidas resumen está atado a la forma de la distribución de los datos. Esto tiene especial importancia con el desvío estándar pues se utiliza muchísimo en las descripciones de los datos. Su fama se debe a la íntima conexión que tiene el desvío estándar con la curva de Gauss. Lo veremos en el capítulo 20.

El **desvío estándar no significa nada** si los datos **no son Normales** ni aproximadamente Normales.

La **media no describe el centro** si los datos **no son simétricos.**

La mediana y la distancia intercuartil pueden fallar si los datos forman grupos.

18.4. Actividades y ejercicios

En los ejercicios 1-4 indique cual es la respuesta correcta o la afirmación que completa la frase. Explique brevemente

1. ¿Cuál de las siguientes opciones da la mejor descripción de los datos cuando estos presentan intervalos vacíos y grupos?

   a) La media y el desvío estándar.
   b) La mediana y el rango intercuartil.
   c) Un gráfico caja con los 5 números resumen.
   d) La mediana y el rango.
   e) Un diagrama tallo-hoja o un histograma.
   f) Ninguno de los anteriores permite mostrar intervalos vacíos y grupos.

2. ¿Cual de las siguientes medidas de posición y variabilidad son adecuadas cuando se sospecha la presencia de datos atípicos?

   a) La media y el desvío estándar.
   b) La media y el máximo menos el mínimo.
   c) La media y la distancia intercuartil.
   d) La mediana y la distancia intercuartil.
   e) La mediana y el máximo menos el mínimo.

3. Si el desvío estándar de un conjunto de datos es cero, se puede concluir que:

   a) La media es cero.
   b) La mediana es cero.
   c) Todos los datos valen cero.
   d) Hay un error de cálculo.
   e) La media mayor que la mediana.
   f) Todos los datos son iguales.

4. Si el 20% de los datos se encuentra entre 10 y 40. Si se dividen por dos todos los valores y luego se les suma 10, también a todos, entonces:

   a) El 10% de los datos resultantes estarán entre 15 y 30.
   b) El 20% de los datos resultantes estarán entre 15 y 30.
   c) El 15% de los datos resultantes estarán entre 15 y 30.
   d) El 10% de los datos resultantes estarán entre 5 y 20.
   e) El 15% de los datos resultantes estarán entre 5 y 20.
   f) El 20% de los datos resultantes estarán entre 5 y 20.

5. Lleve una **balanza** a su división y **registre el peso y la edad** de todos los alumnos y alumnas.

   a. Describa, utilizando histogramas, cómo se distribuyen los **pesos** de todos, juntos y separados, varones y mujeres. Utilice también medidas resumen: media o mediana; distancia intercuartil desvío estándar. Indique cuales son las más adecuadas.
   b. Describa como se distribuyen las **edades** de todos juntos y separados, varones y mujeres. Utilice herramientas gráficas para comparar y también medidas resumen: media o mediana; distancia intercuartil o desvío estándar. Indique cuáles son las más adecuadas.

6. Lleve la **balanza** a **2 divisiones** de **años anteriores** y registre el peso y la edad de todos los alumnos y alumnas.

   a. Describa como se distribuyen los pesos de todos juntos y separados, varones y mujeres. Utilice herramientas gráficas para comparar y también medidas resumen: media o mediana; distancia intercuartil o desvío estándar. Indique cuáles son las más adecuadas.
   b. Describa como se distribuyen las **edades** de todos juntos, varones y mujeres. Utilice herramientas gráficas para comparar y también medidas resumen: media o mediana; distancia intercuartil o desvío estándar. Indique cuáles son las más adecuadas.

Compare los resultados de los distintos años.

7. Realice una encuesta en **su división** para averiguar la cantidad de horas que dedica cada alumno a estudiar y a mirar televisión.

   a. Pregúntele a todos y tendrá datos poblacionales para su división.
   b. ¿Le parece que esa muestra es representativa de todos los alumnos de la escuela?
   c. Elija las variables más relevantes para su encuesta. Establezca las preguntas y evalúe si estas pueden producir sesgo en las respuestas.
   d. Compare cómo se distribuyen las horas entre varones y también entre mujeres.
   e. Utilice herramientas gráficas para comparar y también medidas resumen. Media o mediana. Distancia intercuartil o desvío estándar. Indique cuales son las más adecuadas.

8. Realice una encuesta **en toda su escuela** para averiguar la cantidad de horas que dedica cada alumno a estudiar y a mirar televisión. Utilice **una muestra representativa de todos los años y de género,** especifique como la elegirá. Para esta encuesta puede utilizar las mismas variables y las preguntas que utilizó para su división o modificarlas, según se consideren adecuadas a la luz de los resultados obtenidos. Puede utilizar la colaboración de algún alumno de cada división.