17. Tipos de distribuciones

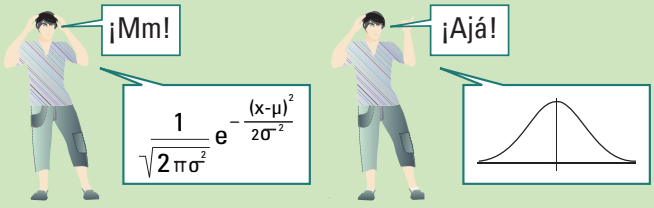
Las distribuciones Normales pueden ser las raras.

□ 17.1. Distribución Normal

Los histogramas y los diagramas tallo-hoja permiten visualizar cómo se distribuyen los valores de una variable numérica. **Muchas veces estos gráficos tienen la forma de una campana,** con una zona central en la cual los valores de la variable son más frecuentes. A medida que nos alejamos de esa zona central las frecuencias disminuyen simétricamente.

Figura 17.1. *Conjunto de datos con distribución en forma de campana, denominada Distribución Normal*

Esta forma de campana también es llamada **campana de Gauss.** Fue descubierta por Abraham de Moivre en 1720. En 1809, Carl Friedrich Gauss, la utilizó para describir los errores de observación cometidos por los astrónomos, al tomar medidas en forma repetida. Fue denominada **curva de error.**



Su fórmula es:

$$\frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

En esta expresión matemática, aparecen e (un número aproximadamente igual a 2,718), π (el ya famoso 3,1415), μ (el centro de la campana) y σ (que permite variar su ancho).

Johann Carl Friedrich Gauss (1777–1855), físico y matemático alemán. Fue un niño prodigio.

Cuentan que en la escuela, para que los alumnos se queden tranquilos por un rato, el maestro les dio la tarea de sumar los números del 1 al 100. Inmediatamente Gauss respondió 5.050. Se había dado cuenta que la suma de los extremos, y a medida que avanzaba, siempre daba 101:

1+ 100 = 101
2+ 99 = 101
3+ 98 = 101

....hasta llegar a la mitad, 50. Sumar todos es 50 veces 101, o sea 50 x 101 = 5.050

La campana de Gauss se obtiene graficando los pares (x, f (x)) en el plano:

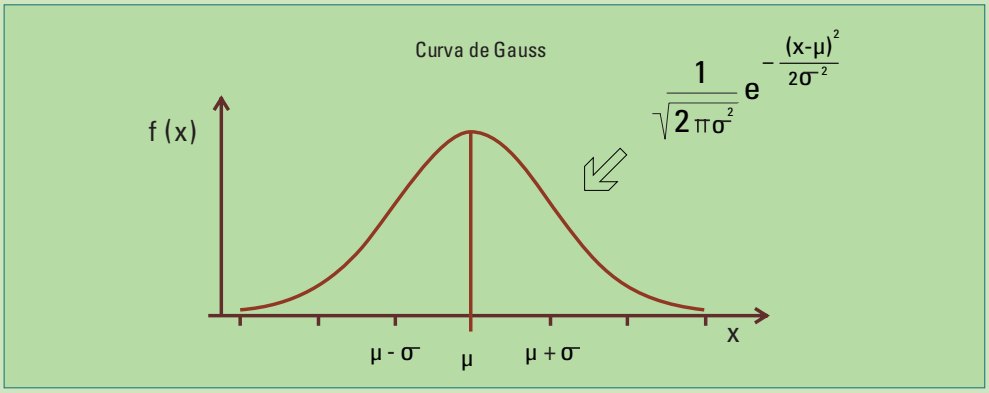


Figura 17.2. *Curva de Gauss junto con su expresión matemática.*

Sólo utilizaremos la forma de la curva y no su expresión.

En 1836 el astrónomo, meteorólogo, estadístico y sociólogo belga Adolphe **Quetelet** extendió la aplicación de la curva, y la utilizó para describir las variaciones de ciertas variables antropomórficas (medidas del cuerpo humano: peso, altura, etc.) entre individuos.

A partir de Quetelet, Francis **Galton** (primo de Charles Darwin y pionero en estudios de genética y de los mecanismos de la herencia) se enteró de la existencia de la curva y **se enamoró de ella.** Dicen que exclamó: "¡Si los griegos la hubieran conocido la habrían deificado!". Galton la llamó curva Normal por primera vez en 1889.

Cuando los datos se distribuyen en forma de campana decimos que tienen distribución Normal o Gaussiana. En la práctica, **los datos rara vez serán "perfectamente Normales**" pero muchas veces la **campana de Gauss** es una muy buena **aproximación al histograma** de un conjunto de datos.

17.1.1. Curva Normal estándar.

Si $= 0$ y $=1$ la curva Normal es:

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

Consideremos un conjunto de datos cuyo histograma puede aproximarse por la curva Normal con parámetros μ y σ . Si a cada dato se le resta μ y se lo divide por σ, entonces el histograma del nuevo conjunto de datos podrá aproximarse por la curva Normal Estándar (figura 17.3).

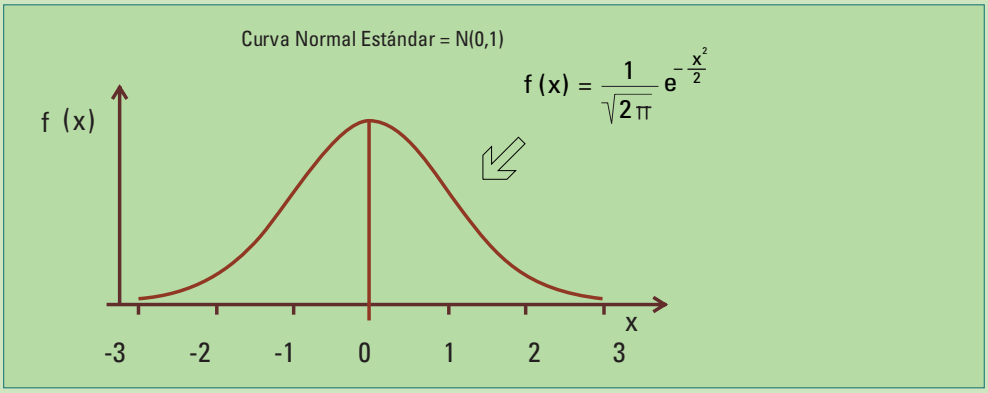


Figura 17.3. *Curva Normal Estándar junto con su expresión matemática.*

17.1.2. ¿Cuándo se obtienen datos con variabilidad Normal y cuando no?

Insistimos, un conjunto de datos rara vez podrá tener una distribución que se ajuste perfectamente a la curva Normal. Sin embargo, en muchas situaciones, esta curva provee una excelente aproximación a los histogramas de los datos. A continuación, presentamos algunos ejemplos en los cuales la aproximación puede ser buena y algunos de cuando no puede serlo.

Ejemplo 17.1. Variabilidad Normal entre unidades muestrales.

Piezas metálicas producidas con la misma máquina, por el mismo operador y en el mismo turno, podrán parecer iguales, pero al medir su **dureza** con cuidado se encontrarán **diferencias.** Cuando estas variaciones se producen en **condiciones normales** (ahora con ene minúscula) – con esto queremos decir: la máquina está funcionando como habitualmente, la materia prima es la de siempre, las herramientas están como todos los días, los operarios descansados y con el ánimo de siempre - entonces las piezas serán parecidas. Las variaciones, respecto de alguna variable (dureza, longitud, peso, elasticidad), darán muchos valores en el centro y pocos en los extremos. A esto lo denominamos "variabilidad Normal" (aquí con ene mayúscula). En cambio, si el producto se fabrica con materias primas defectuosas o los operarios estaban distraídos y siguieron operando cuando la herramienta estaba dañada, la distribución de las variables examinemos ya no será Normal.

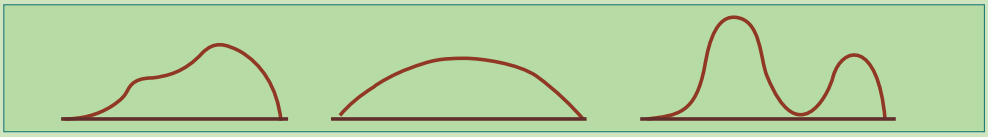


Figura 17.4. *Variabilidad no normal.*

Una aclaración: "normal" con ene minúscula es sinónimo de "habitual"; "Normal" con ene mayúscula se refiere a una distribución de datos en forma de campana

Ejemplo 17.2. Variabilidad debido al error de medición.

Si evaluamos varias veces la **dureza de un mismo** producto, los valores no serán idénticos, aunque la dureza sea siempre la misma. Habrá muchos valores cercanos al **valor verdadero** de la dureza y la frecuencia de los valores disminuirá al alejarse. Si estas mediciones son realizadas con mucho cuidado, por el mismo operador, realizando desde el comienzo los mismos pasos hasta obtener el resultado, su histograma tendrá la forma de la curva Normal. En cambio, si hubo algún descuido al realizar las mediciones, cambió el operador o alguna condición del proceso de medición, estas no tendrán un histograma que pueda ser representado mediante la curva de Gauss.

Ejemplo 17.3. Variabilidad biológica normal.

Si registramos las estaturas de las niñas de una división, encontraremos unas pocas son muy bajitas, otras pocas muy altas y la mayoría con alturas intermedias. Los datos, las mediciones, tendrán una distribución aproximadamente Normal. En cambio, si consideramos las alturas de todos los alumnos (varones y mujeres), los datos provienen de muestras no homogéneas y no tendremos una "variabilidad Normal". Pero no todas las variables antropomórficas tendrán una buena aproximación por la distribución gaussiana como creía Quetelet, por ejemplo, el peso de las personas de una edad y género determinados no tiene distribución simétrica, tampoco los niveles de los triglicéridos en sangre.

La **Normalidad estadística** no implica la normalidad biológica, social o económica. Muchas veces las distribuciones Normales son las raras.

La distribución de los salarios de una población, el caudal de un río de montaña, la precipitación diaria en cierta ciudad, son ejemplos de distribuciones asimétricas.

□ 17.2. Formas que describen diferentes tipos de distribuciones. Curvas de densidad.

La figura 17.5 muestra un histograma de 400 datos de una variable continua y una curva que describe la forma con la que se distribuyen los datos a lo largo de sus valores. Vemos que las mayores frecuencias se encuentran entre cero y cuatro. Para valores mayores que cuatro se reduce constantemente la frecuencia. Podríamos pensar que la curva se obtiene en dos pasos:

- dibujando el borde superior de cada rectángulo de clase y luego.
- suavizando los escalones.

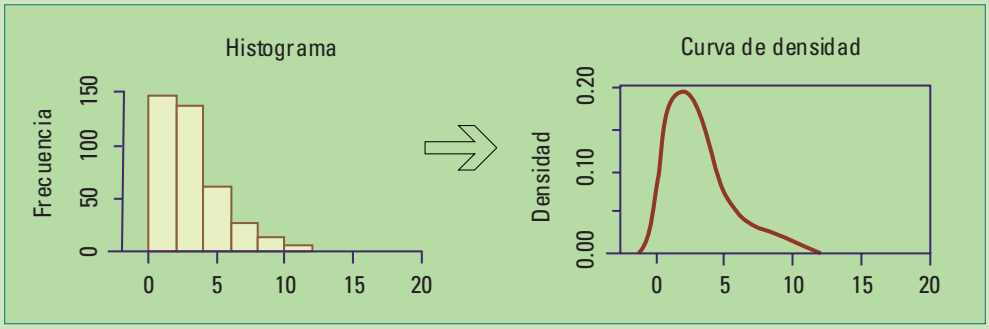
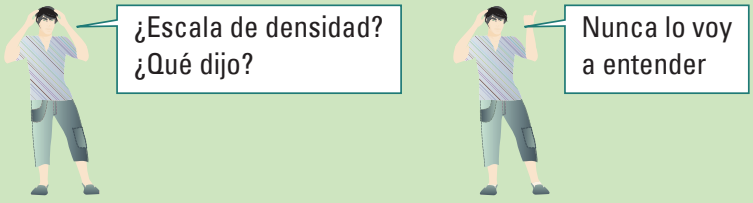


Figura 17.5. *Un histograma y su curva de densidad.*

Existe una diferencia para resaltar entre histogramas y curvas de densidad. Las curvas de densidad se grafican en escala de densidad y los histogramas en escala de frecuencias o frecuencias relativas.

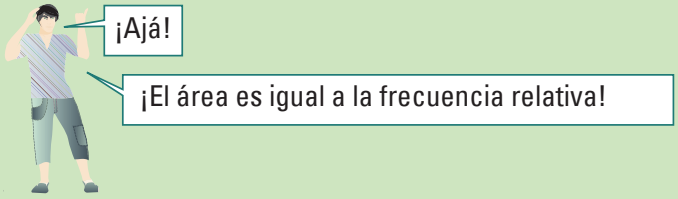


La mayoría de los histogramas muestran la cantidad (frecuencia) o proporción (frecuencia relativa) de observaciones de cada intervalo de clase mediante la altura del rectángulo. De esta manera, el **área de cada rectángulo es proporcional a la frecuencia relativa.** En una **escala de densidad,** el **área de cada rectángulo** es **IGUAL a la frecuencia relativa,** y se obtiene graficando en el eje vertical la frecuencia relativa dividida la longitud del intervalo de clase. En escala de densidad, el área total de los rectángulos del histograma es 1.

Para los datos de la figura 17.5 la longitud de los intervalos es 2 y tenemos:

Escala	Frecuencias		Frecuencias relativas		Densidad	
	Altura	Área	Altura	Área	Altura	Área
	147	294	0,3675	0,735	0,18375	0,3675
	138	276	0,3450	0,690	0,17250	0,3450
	62	124	0,1550	0,100	0,07750	0,1550
	29	58	0,0725	0,145	0,03625	0,0725
	16	32	0,0400	0,080	0,02000	0,0400
	6	12	0,0150	0,030	0,00750	0,0150
	1	2	0,0025	0,005	0,00125	0,0025
	0	0	0,0000	0,000	0,00000	0,0000
	1	2	0,0025	0,005	0,00125	0,0025
Total	400	800	1,0000	2,000	0,5000	1,0000

En **escala de densidad** el **área del rectángulo** de clase es igual a la **frecuencia relativa** y la suma de las áreas es 1.



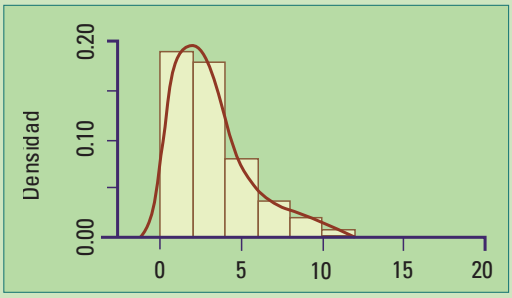
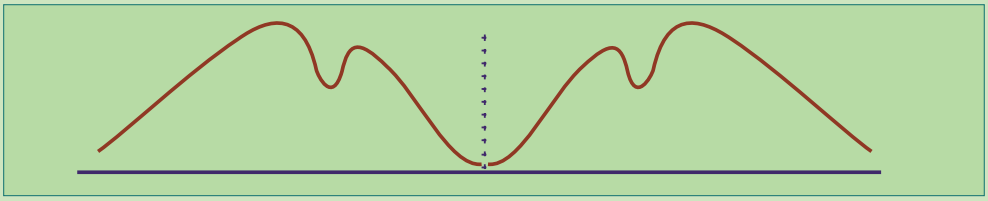


Figura 17.6. *Superposición de un histograma y una curva de densidad.*

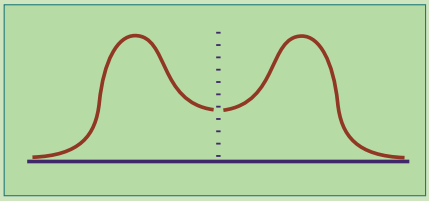
Los histogramas pueden tener distintas formas. Mostraremos algunos patrones especiales en forma simplificada mediante **curvas,** también llamadas **curvas de densidad.** La campana de Gauss es una de ellas.

17.2.1. Simétrica

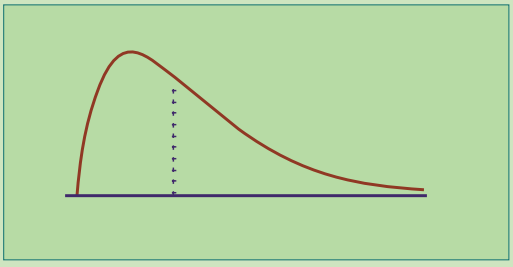
Una **distribución es simétrica** cuando sus dos mitades son imágenes especulares una de la otra.



Por ejemplo, un histograma de las alturas de los mayores de 18 años de un pueblo tendrá dos zonas más altas en espejo, una para los varones y otro para las mujeres, mientras haya la misma cantidad de varones y mujeres. Esto se debe a la superposición de dos curvas simétricas con distinto centro e igual ancho.



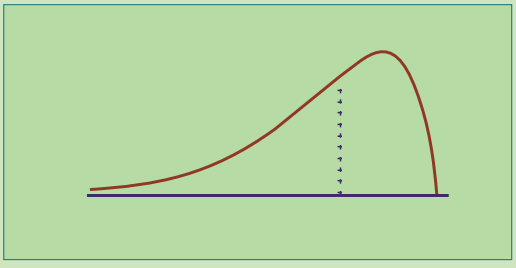
17.2.2. Asimétrica a derecha



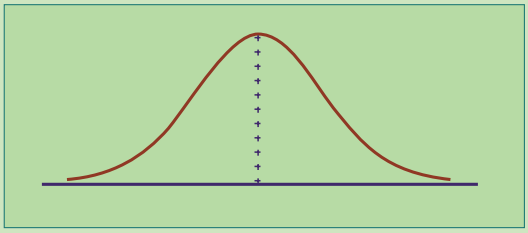
Una **distribución es asimétrica a derecha** cuando la mitad derecha es más finita y más larga. Por ejemplo, la distancia de los domicilios de los alumnos a la escuela mostrará muchos valores pequeños, en la mitad izquierda del histograma, son las de los alumnos que viven cerca y habrá pocos valores grandes de los alumnos que viven lejos.

17.2.3. Asimétrica a izquierda

Una **distribución es asimétrica a izquierda** cuando la mitad izquierda es más finita y más larga. En un **examen fácil,** la mayoría de las notas serán altas y estarán amontonadas del lado derecho, con unas pocas notas bajas (las del lado izquierdo).

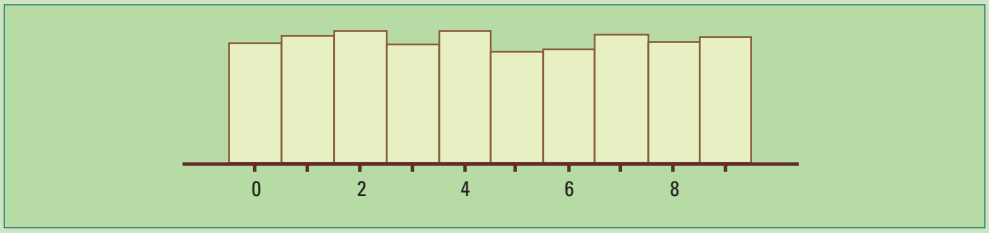


□ 17.2.4. Con forma de campana

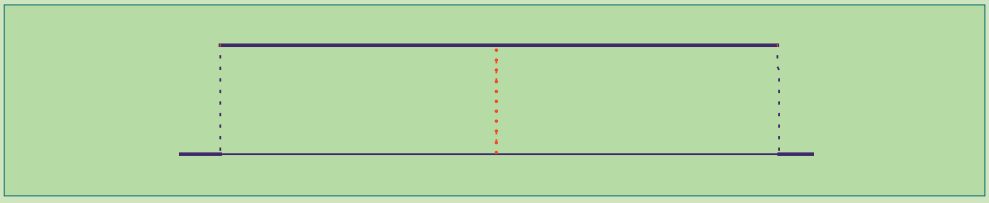


Una **distribución con forma de campana** es **simétrica** con un montículo en el centro y dos caídas como toboganes hacia los costados. Es una de las distribuciones de datos que tal vez aparezca con más frecuencia y es la más estudiada.

□ 17.2.5. Uniforme



Las frecuencias de la última cifra de los resultados de una lotería muestran una distribución pareja sobre todos los dígitos de 0 a 9. Si el mecanismo que genera los números de la lotería funciona correctamente, ninguno de los dígitos tiene más chances de aparecer. Este tipo de distribuciones se llama uniforme y se representa mediante una recta:



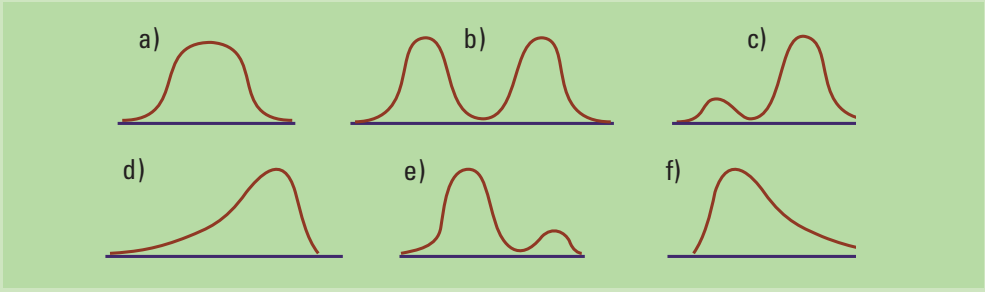
17.3. Actividades y ejercicios

1. Se enumeran las edades de los miembros de 15 familias, casados por 3 años como máximo. En cada fila de cada casilla tenemos las edades de los miembros de una familia.

20, 19	23, 23	23, 25, 2	26, 30, 1, 2	28,31, 2
25, 18, 2, 3	18, 26	38, 34	32, 32	17,19
30, 35, 1	34, 29	21, 19,1	24,26	21,27

Construya un diagrama tallo-hojas y un histograma de las edades. Describa las formas que tienen.

2. ¿Cuál de las siguientes figuras puede representar histogramas de las edades de todos los miembros de familias constituidas a lo sumo hace 2 años?



3. ¿Puede el año de emisión de las monedas decirnos algo más? Para hacer entre todos

- Cada alumno obtiene 10 monedas de 10 centavos y las agrupa en pilas de acuerdo con su año de emisión.
- Cuenta cuantas monedas tiene para cada año.
- Cada alumno/a indica cuantas monedas tiene por cada año y completa la tabla.
- Se obtiene un histograma de la distribución de las fechas.
- ¿Qué forma tiene?
- ¿Qué forma, le parece, debería tener si se perdiera una proporción constante de monedas cada año y a su vez se emitiera una misma cantidad de monedas cada año?
- ¿Puede hallar alguna explicación a la forma del histograma correspondiente a las monedas verdaderas?

Año	Frecuencia
1992	
1993	
.....	
2008	
.....	
	Completar

4. Para estudiar las longitudes de las palabras, seleccione un artículo de una revista de deportes y otro de una de divulgación científica. Para cada uno de los artículos obtenga:

   - la distribución de frecuencias
   - la distribución de frecuencias relativas
   - el histograma

de la variable "cantidad de letras" que tiene cada palabra. Compare las distribuciones obtenidas.

Observación: Diferentes idiomas tienen diferentes distribuciones de las longitudes de las palabras.

18. Medidas resumen

Media, mediana, rango, desvío estándar, distancia intercuartil

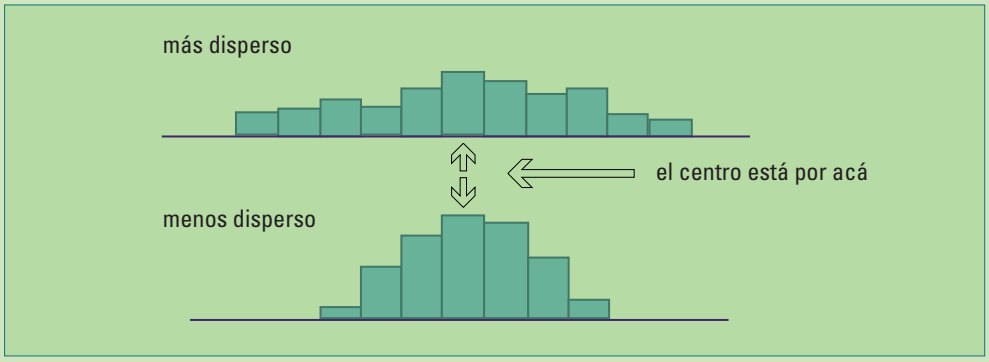
La mente humana puede captar la información que aportan diez números, cien es difícil y con mil, estamos perdidos. Por esa razón, es muy importante contar con pocos valores (medidas resumen), que de alguna manera deben describir las características más sobresalientes del conjunto que se está analizando.

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.

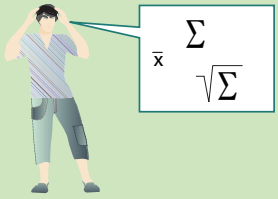
Cualquier conjunto de datos tiene **dos propiedades importantes:** un **valor central** y la **dispersión** alrededor de ese valor. Vemos esta idea en los siguientes histogramas hipotéticos:



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,

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

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

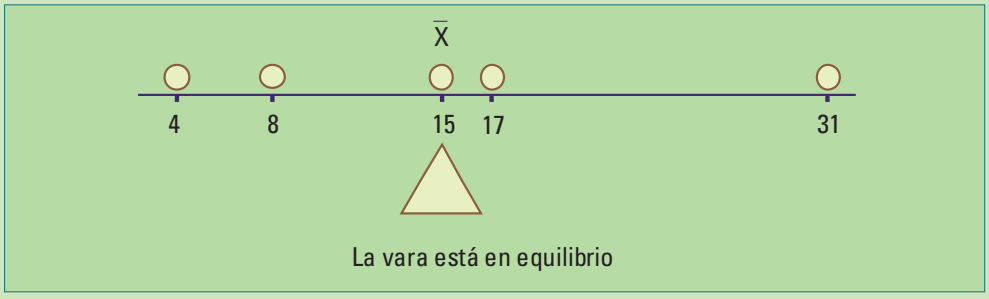
En el ejemplo anterior, la media de las cuadras caminadas por día es 15:

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

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

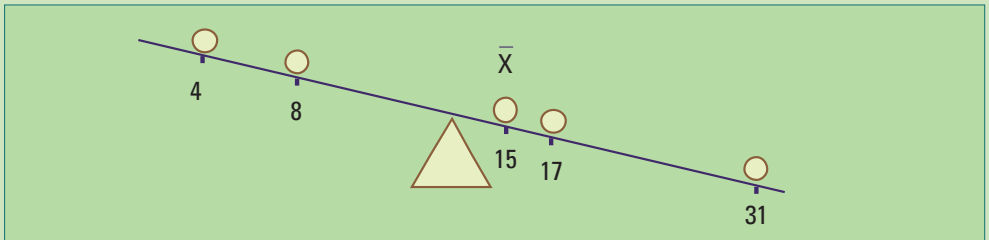
$$\bar{x} = 15 \text{ CUADRAS}$$

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:

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

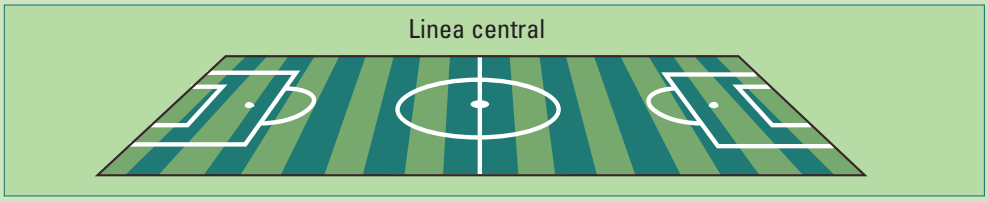
En el ejemplo 16.5, para los pesos de los 101 alumnos de 3 divisiones de 4to. Año, el peso medio es 58,90 kg:

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

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

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.



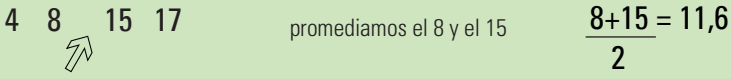
La mediana deja la misma cantidad de datos a cada lado.

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 | 5556677788899 ← 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?

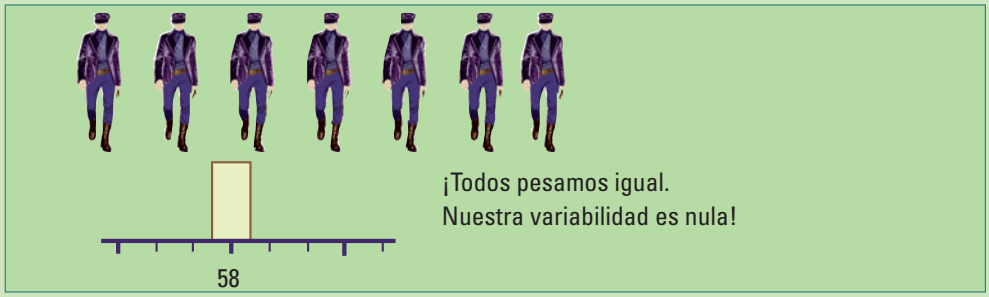
4 8 15 17 316

mediana ⇗ media = 72 ⇗

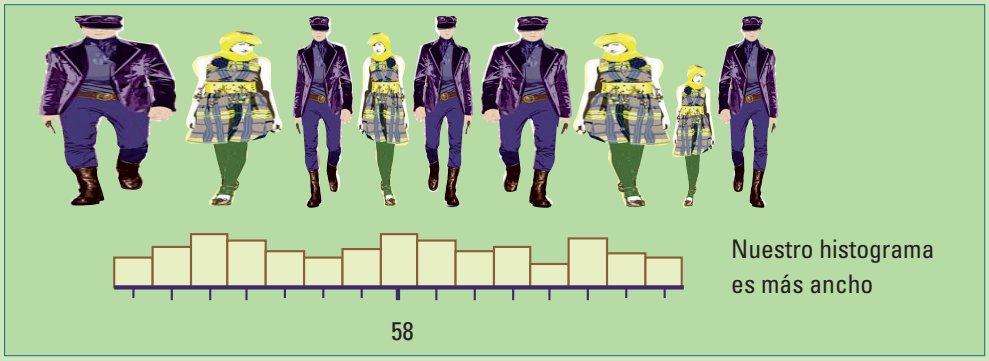
La media (72) ya no representa a la mayoría de los datos, por eso, decimos que **la media es sensible ante la presencia de valores atípicos (outliers).**

□ 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.

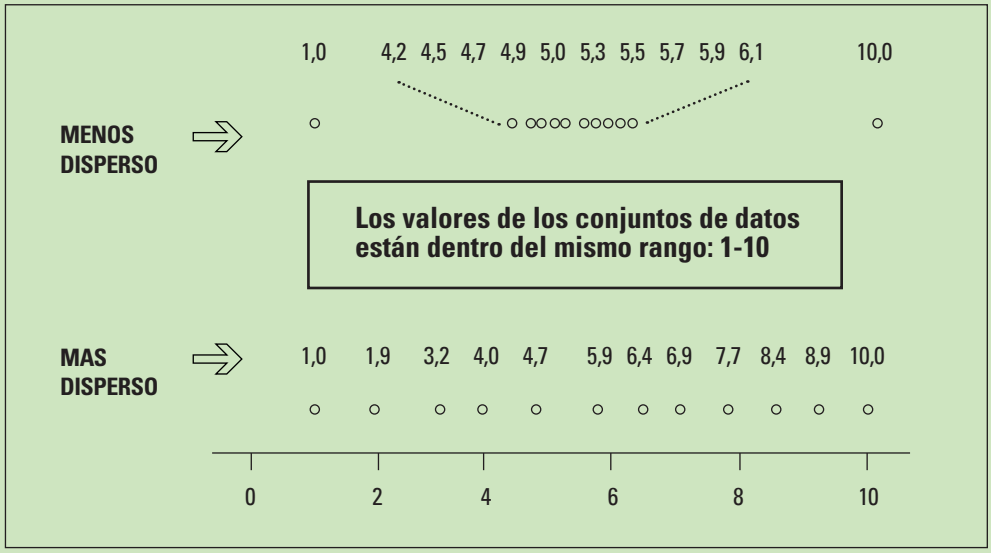
18.2.1. Rangos y distancia intercuartil

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).

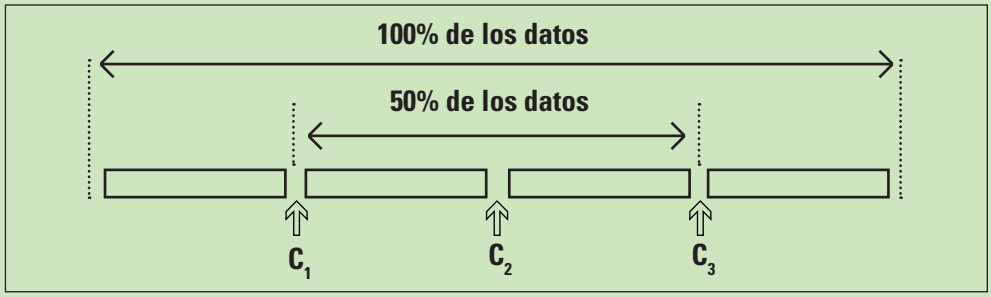
Tiene algunos inconvenientes:

- 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 | 55566777888**8**99 ← **Aquí se encuentra la mediana**
6 | 011223333444
6 | 555666667**7**7777788899999
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:

Mínimo C_1 M C_3 Máximo

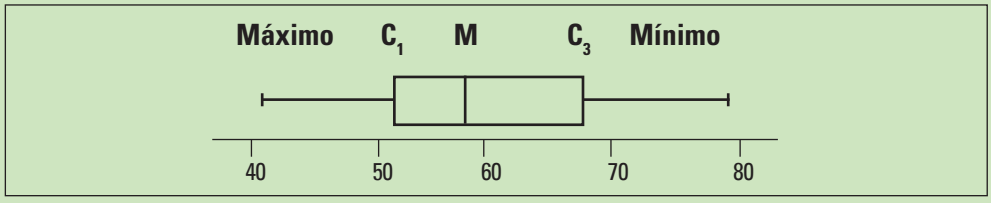
El 50% de los datos se encuentran entre el cuartil inferior y el superior.

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.

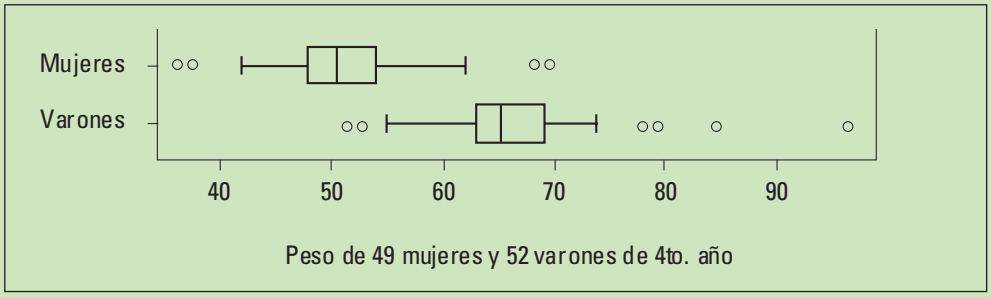
Los cinco números resumen se representan gráficamente en un Gráfico de caja (Box-plot).



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.

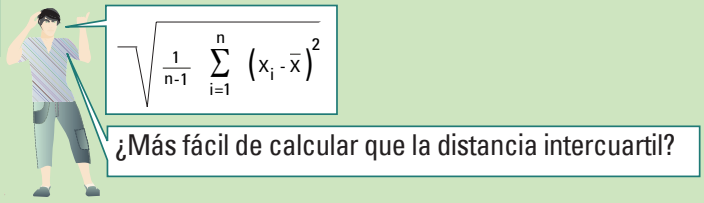
El **desvío estándar** es una **medida** de dispersión **basada en la media** y **utiliza todos los datos.** Durante muchos años la **media y el desvío estándar** fueron, y tal vez sigan siendo, las **medidas resumen más utilizadas.**

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.

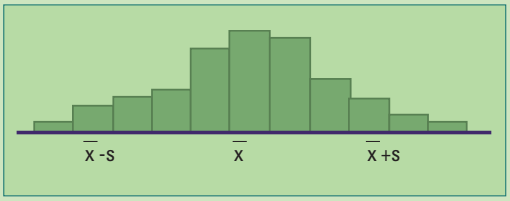
El desvío estándar es 10,37 cuadras:

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

$$s = 10,37$$

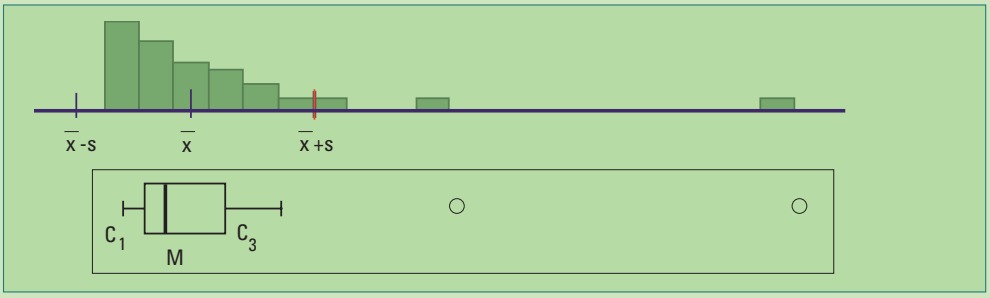
18.3. Centro y dispersión en diferentes tipos de distribuciones

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

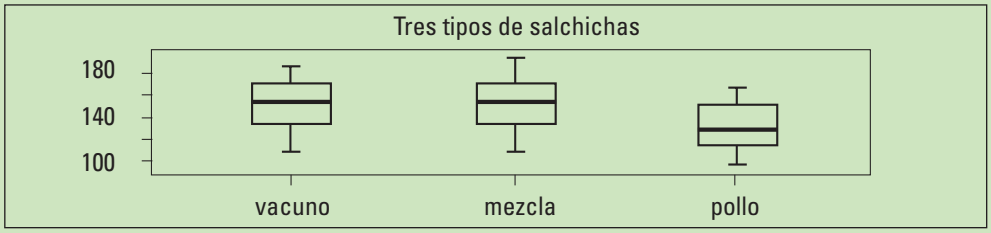


Figura 18.1. *Gráficos caja de la cantidad de sodio de tres tipos de salchichas.*

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:

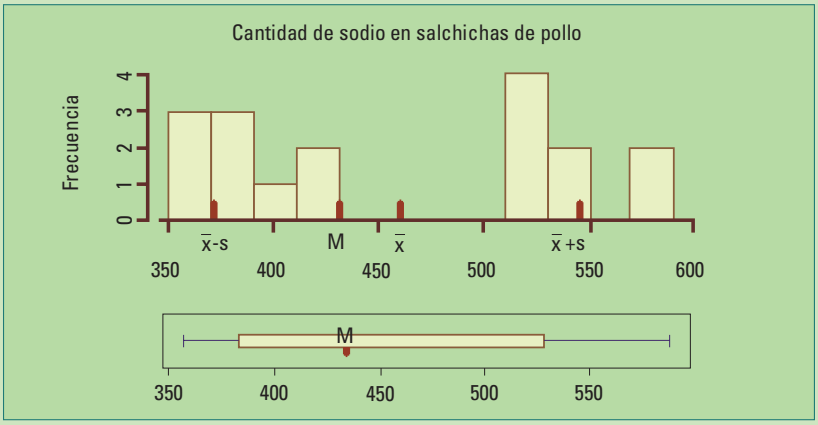


Figura18.3. *Histograma y gráfico caja de la cantidad de sodio de salchichas de pollo de diferentes marcas.*

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.

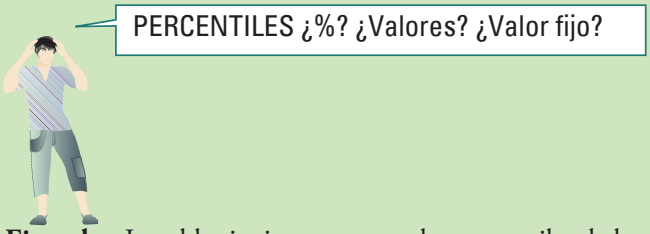
19. Otras medidas de posición: los percentiles

El cuartil inferior es el percentil 25.
La mediana es el percentil 50.

Los percentiles nos permiten responder preguntas como:

- Tiene 16 años, mide 1,57 cm, es la primera de la fila de su división, ¿significa que es petisa?
- Los alumnos de 2do. año de una escuela practican deportes 6 horas por semana ¿es mucho o poco en comparación con otras escuelas?
- Tiene 25 años, se tomó la presión y obtuvo 130/70 (sistólica /diastólica). ¿Es normal?
- ¿Cuál es la longitud debajo de la cual se encuentran el 90% de los bebés recién nacidos?
- Un bebé nació con 3800 g y 50 cm de longitud corporal ¿No está un poco gordito? ¿Será petiso?
- En relación al volumen de ventas, ¿cómo está posicionada una compañía azucarera en comparación con la competencia?

Los percentiles se utilizan para determinar la posición relativa, en porcentaje, de la posición que ocupa un valor dado, de una variable, en relación a todos los valores de la misma en un grupo o en una población.



Ejemplo: La tabla siguiente muestra los percentiles de la altura (m) de mujeres y varones de 16 años (Gráfico Nº 6 Guías para la Evaluación del Crecimiento. Sociedad Argentina de Pediatría. 2001).

PERCENTILES DE LA ALTURA (m) DE MUJERES Y VARONES DE 16 AÑOS. TABLA 19.1

Percentil	**3**	10	**25**	50	75	90
Mujer	**1,49**	1,53	**1,56**	1,60	1,64	1,68
Varón	1,56	1,60	1,65	1,70	1,74	1,79

Fuente. *Guías para la Evaluación del Crecimiento. Sociedad Argentina de Pediatría. 2001.*

El 3% de las jóvenes de 16 años miden menos o igual que 1,49 m. El **percentil 3 de la altura** de las jóvenes de 16 años es de **1,49 m.**

El 25% de las jóvenes de 16 años miden menos o igual que 1,56 m. El **percentil 25 de la altura** de las jóvenes de 16 años es de **1,56 m.**

El 50% de las jóvenes de 16 años miden menos o igual que 1,60 m. El **percentil 50 de la altura** de las jóvenes de 16 años es de **1,60 m.**

En general hablaremos del **percentil K**. Si decimos percentil 25, eso significa K = 25. También se lo denomina **percentil del K** %, se dice "percentil del 25%" en lugar de "percentil 25".

- ¿A qué población corresponden los percentiles de la tabla 18.1? A mujeres y varones de 16 años de la Argentina.
- ¿Cuál es el percentil 10 de los varones de 16 años? El percentil 10 de la altura de los varones es 1,60 m ¿Qué significa ese valor? Significa que el 10 % de los jóvenes de 16 años miden menos o igual que 1,60 m.
- ¿Cuál es el percentil 10 de las mujeres de 16 años? El percentil 10 de la altura de los varones es 1,53 m ¿Qué significa ese valor? Significa que el 10 % de las jóvenes de 16 años miden menos o igual que 1,53 m.
- ¿Son iguales los valores anteriores? ¿Por qué? Los percentiles 10 de varones y mujeres difieren porque las distribuciones de las alturas son diferentes.

¡Ajá!

Percentil 25 ¡Es el cuartil inferior!
Percentil 50 ¡Es la mediana!
Percentil 75 ¡Es el cuartil superior! y hay muchos más!!!

En general: Una proporción p de jóvenes de 16 años tiene una altura por debajo del percentil 100 x p de la altura de las jóvenes de 16 años.

Más en general: Una proporción p de observaciones de una variable está por debajo del percentil 100 x p de dicha variable.

"Mido 1,57 m. ¡No soy petisa! El 25 % de las chicas de mi edad miden menos que yo."

19.1. ¿Cómo se calcula un percentil en un conjunto de datos?

19.1.1. Cuando los datos no están agrupados

Veamos ahora una **forma general para hallar el percentil K para cualquier conjunto de datos:**

Paso 1. Ordene los datos de menor a mayor
Paso 2. Calcule **K/100** y multiplíquelo por la cantidad total de datos **n.**
Paso 3. Redondee n K/100 al entero más cercano.
Paso 4. Cuente desde el dato más chico hacia el más grande tantos lugares como el número hallado en el paso 3.

Ejemplo: Ahora usaremos **peso,** no altura.

Retomemos nuevamente los datos del ejemplo 16.5. Consideremos los **pesos** (en kg) de las 49 alumnas de 4to. año y hallemos el percentil 40.

Paso 1. Ordenamos los datos:
37 38 42 43 43 44 45 46 46 47 48 48 48 48 48 50 50 50 50 50 50 51 51 51 51
52 52 52 52 52 52 52 52 54 54 54 54 54 54 55 55 56 56 57 58 60 62 68 70

Paso 2. Calculamos **20/100** = 0,2 y lo multiplicamos por la cantidad total de datos 49. Esto da como resultado 9,8

Paso 3. Redondeamos 9,8 al número entero más cercano, o sea 10.

Paso 4. Contamos desde el dato más chico hacia el más grande 10 lugares:

37 38 42 43 43 44 45 46 **46** 47 48 48 48 48 48 50 50 50 50 50 50 51 51 51 51
52 52 52 52 52 52 52 52 54 54 54 54 54 54 55 55 56 56 57 58 60 62 68 70

No confundir el valor del percentil con el porcentaje.

Por lo tanto, 46 kg es el percentil 20 para los datos de los pesos (en kg) de las 49 alumnas de 4to. año; 46 es el valor y 20 es el porcentaje.

19.1.2. Cuando los datos están agrupados

Cuando los datos tienen muchos **valores repetidos** es más conveniente utilizar una tabla de frecuencias para calcular los percentiles. Utilizaremos nuevamente las 49 alumnas de 4to. año. para calcular el percentil 20. Si consideramos **todos los valores desde el más chico hasta 46** (tabla 19.2) se acumulan **aproximadamente** el 20 % de los pesos.

Ejercicios utilizando la tabla 19.2:

- Halle el peso correspondiente al percentil 90. Si no lo encuentra exactamente, obtenga el más cercano.
- Una alumna pesa 52 kg, ¿en qué percentil se encuentra? ¿Es un percentil respecto de las 49 alumnas o respecto a toda la población?

Solución

- El porcentaje acumulado más cercano a 90 es 89,80 y le corresponde un peso de 57 kg. Podemos decir que aproximadamente el 90% de las alumnas del curso pesa a lo sumo 57 kg.
- Se encuentra en el percentil del 67,35 %. Se trata de un percentil respecto a las 49 alumnas del curso.

FRECUENCIAS DEL PESO DE LAS 49 ALUMNAS DE 4TO. AÑO. TABLA 19.2

Peso	Frecuencia	Frec. acumulada	Frec. relativa (en %)	Porcentaje acumulado
37	1	1	2,04	2,04
38	1	2	2,04	4,08
39	0	2	0,00	4,08
40	0	2	0,00	4,08
41	0	2	0,00	4,08
42	1	3	2,04	6,12
43	3	6	6,12	12,24
44	1	7	2,04	14,29
45	1	8	2,04	16,33
46	**2**	**10**	**4,08**	**20,41**
47	1	11	2,04	22,45
48	5	16	10,20	32,65
49	0	16	0,00	32,65
50	5	21	10,20	42,86
51	4	25	8,16	51,02
52	8	33	16,33	67,35
53	0	33	0,00	67,35
54	6	39	12,24	79,59
55	2	41	4,08	83,67
56	2	43	4,08	87,76
57	1	44	2,04	89,80
58	1	45	2,04	91,84
59	0	45	0,00	91,84
60	1	46	2,04	93,88
61	0	46	0,00	93,88
62	1	47	2,04	95,92
63	0	47	0,00	95,92
64	0	47	0,00	95,92
65	0	47	0,00	95,92
66	0	47	0,00	95,92
67	0	47	0,00	95,92
68	1	48	2,04	97,96
69	0	48	0,00	97,96
70	1	49	2,04	100,00

19.2. Percentiles poblacionales de peso y talla en niños

Los percentiles utilizados habitualmente para evaluar el crecimiento de un niño, son estimaciones de los verdaderos percentiles poblacionales. Suelen obtenerse para el peso, la talla, el perímetro cefálico y el índice de masa corporal. Pueden hallarse las tablas actualizadas para Argentina en: http://www.garrahan.gov.ar/docs/2270/rgenerales.html

Los percentiles del **peso y talla** son los más utilizados. Se representan con curvas mostrando los percentiles 3, 10, 25, 50, 75, 90 y 97 en función de la edad, correspondientes a valores de niños normales, sanos.

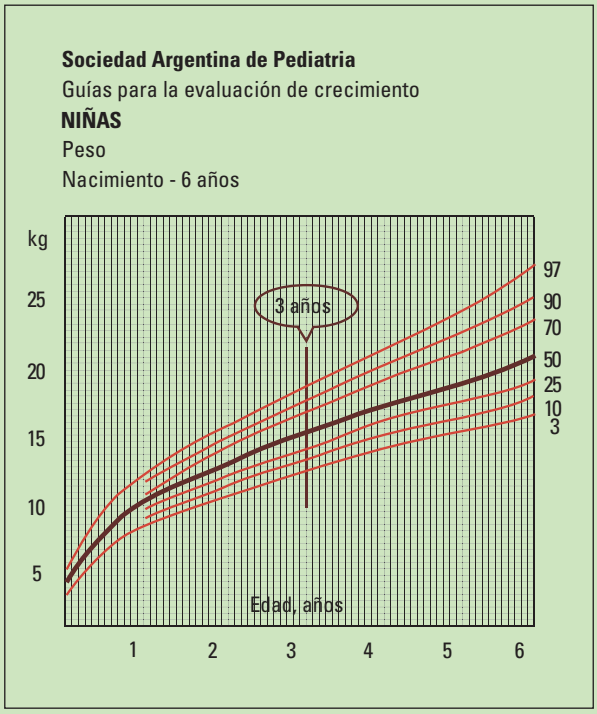


Figura19.1. *Fuente: Sociedad Argentina de Pediatría, Guías para la Evaluación del Crecimiento, 2001.*

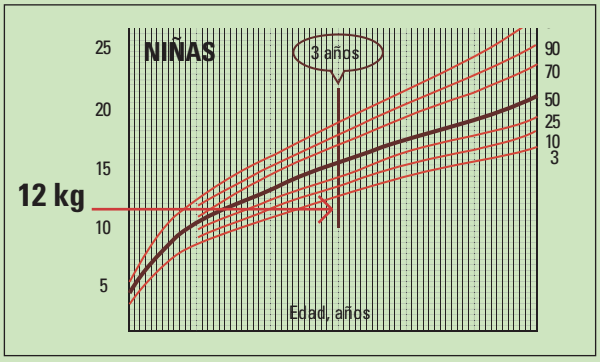


Figura19.2. *Peso de una niña de 3 años que se encuentra en el percentil 10.*

Ejemplo1:
La figura 19.1 muestra los percentiles de peso, desde el nacimiento hasta 6 años de edad.

Las curvas muestran los percentiles 3, 10, 25, 50, 75, 90 y 97 del peso de niñas sanas con edades entre 0 y 6 años. Se destacan los valores para 3 años de edad.

Si el médico dice que una niña de 3 años está en el percentil 10 respecto al peso significa que el 10 % de las niñas sanas de 3 años pesan a lo sumo como ella. Pero, ¿cuánto pesa? Como está en el percentil 10 y tiene 3 años podemos hallar su peso. Lo obtenemos (12 kg) trazando una línea horizontal, en el gráfico de los percentiles del peso, a la altura en que el percentil 10 corta la línea de 3 años (figura 19.2)

Ejemplo2:
La figura 19.3 (Sociedad Argentina de Pediatría, Guías para la Evaluación del Crecimiento, 2001) muestra los percentiles 3, 10, 25, 50, 75, 90 y 97 del peso de niñas sanas con edades entre 0 y 19 años. Este gráfico permite hallar el rango de valores de los pesos del 94% de las niñas sanas para cada edad. Solamente un 6% de niñas sanas tendrán pesos fuera de ese rango con un 3% por debajo y un 3% por encima.

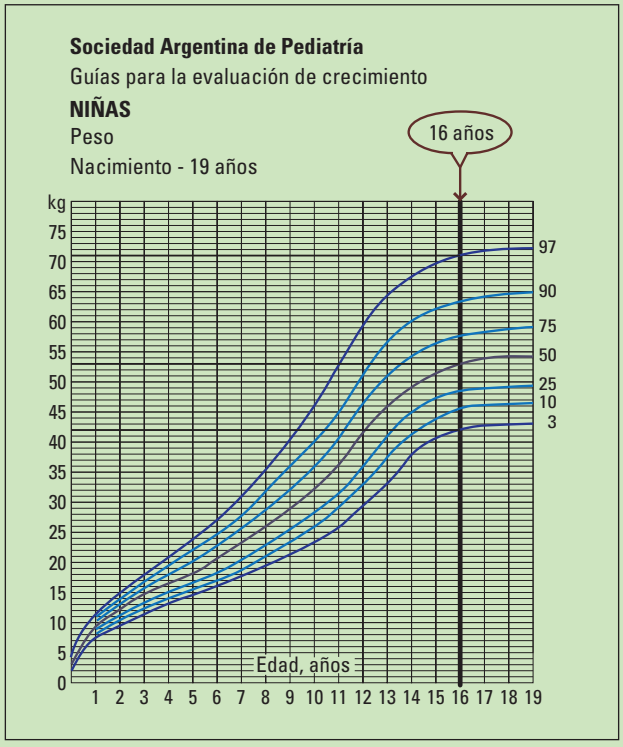


Figura19.3. *Las curvas muestran los percentiles 3, 10, 25, 50, 75, 90 y 97 del peso de niñas sanas con edades entre 0 y 19 años. Se destacan los valores para 16 años de edad.*

Percentiles del peso (kg) de las mujeres de 16 años. TABLA 19.3

Percentil	Peso
3	42,0
10	45,5
25	48,5
50	53,0
75	57,5
90	63,5
97	71,0

Fijemos nuestra atención en edad =16 años (figura 19.3). Se observa una línea vertical para esa edad. Los pesos que se obtienen, de los puntos donde la línea vertical corta a cada una de las curvas, se muestran en la tabla 19.3 y en la figura 19.4.

El rango de valores para el 94 % central de los datos se encuentra entre los pesos correspondientes a los puntos donde esa línea corta los percentiles 3 y 97 respectivamente. Este rango va desde 42 kg (percentil 3) hasta 71 kg (percentil 97). Solamente el 6% tiene su peso fuera de ese rango de valores, se trata de las extremadamente livianas y las extremadamente pesadas. La mediana es de 53 kg.

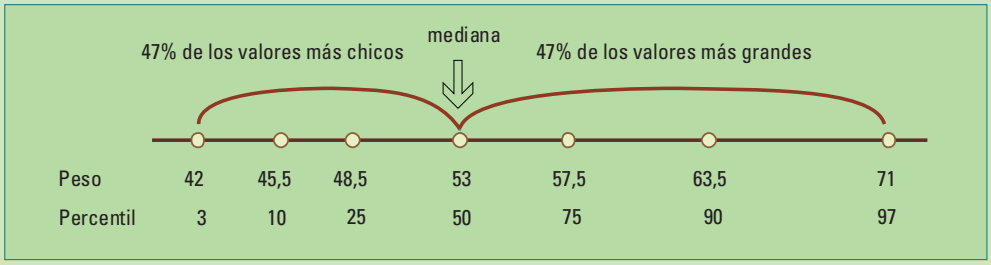


Figura 19.4. *Diagrama de puntos de los pesos correspondientes a los percentiles 3, 10, 25, 50, 75, 90 y 97 de las mujeres de 16 años.*

En la figura 19.4 se aprecia la mayor concentración de los valores más chicos, por debajo de la mediana, en comparación con los valores más grandes. Por lo tanto el **peso es una variable no simétrica,** con leve asimetría hacia la derecha.

Ejemplo 3:
Pesa 46 kg y tiene 16 años. En relación con sus compañeras está en el percentil 20 (tabla 19.2), ¿y respecto de la población? Puede obtenerse esa respuesta utilizando los percentiles poblacionales de la figura 19.5. Un poco más del 10 % de las chicas de su edad pesan menos que 46 kg; está levemente por encima del percentil 10 y por debajo del percentil 25 respecto al peso.

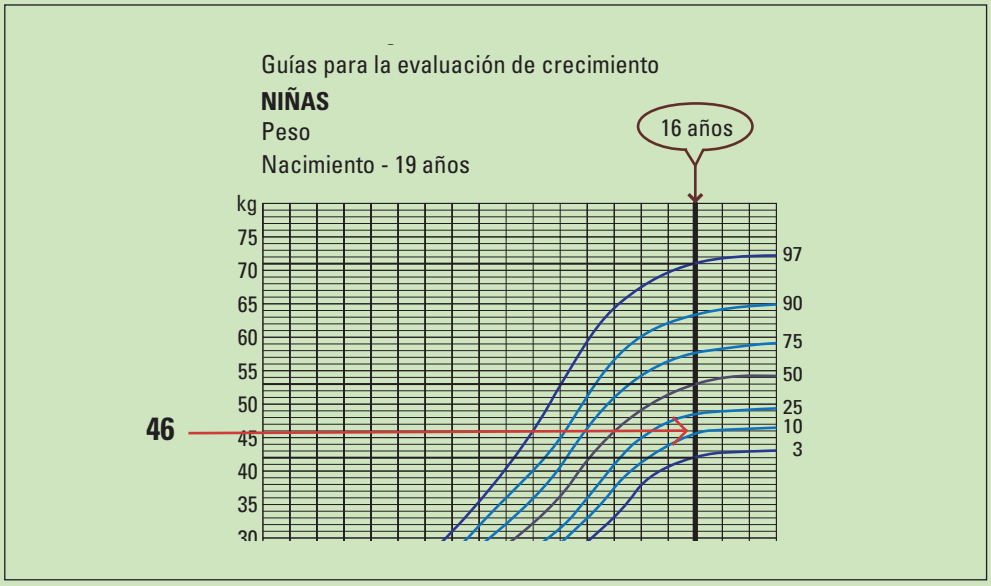


Figura 19.5. *Un peso de 46 kg, es un percentil entre el percentil 10 y el percentil 25.*

19.3. Actividades y ejercicios

1. La mediana deja la mitad de los datos ordenados a cada lado.

   - ¿Por qué en la figura 19.4 aparecen solamente el 47% de los valores más chicos a la izquierda de la mediana, el 47 % a la derecha y no el 50% a cada lado?
   - ¿Faltan datos? ¿Cuál es el porcentaje de datos faltantes? ¿Por donde estarán?

2. Complete la tabla de frecuencias siguiente utilizando la información de la figura 19.4

Intervalo de peso (kg)	Longitud del Intervalo	Frecuencia Relativa en %	Frecuencia Relativa en % / Longitud del Intervalo
[42 ; 45,5)	3,5	10- 3 = 7	7/3,5 = 2
[45,5 ; 48,5)	3,0	25-10 = 15	15/3,0 = 5
[48,5 ; 53)	4,5	50-25 = 25	
[53 ; 57,5)	4,5	75-50 =	
[57,5 ; 63,5)			
[63,5 ; 71)			

3. Construya un histograma para el 94% central de los valores de la variable peso de las mujeres de 16 años utilizando la tabla anterior. Grafique en el eje horizontal los intervalos y utilice la escala densidad para el vertical (la frecuencia relativa en %) / (la longitud del intervalo). Es necesario utilizar la escala densidad porque los intervalos de clase del histograma tienen distinta longitud. Indique donde se encuentra el percentil 50 y observe que los datos presentan una leve asimetría a derecha.

4. Construya un histograma para el peso de las jóvenes de 13 años siguiendo los siguientes pasos:

a) Trace una línea vertical en la figura 19.3 en edad = 13 años.
b) Halle los pesos que corresponden a los puntos donde la línea vertical corta a cada una de las curvas. Esos pesos son los percentiles 3, 10, 25, 50, 75, 90 y 97 del peso de chicas de 13 años.
c) Obtenga una tabla de frecuencias similar a la presentada en 2. pero en este caso con edad=13 años.
d) Construya un histograma para el 94% central de los valores de la variable peso de las chicas de 13 en forma similar a lo realizado en 3.

5. Utilice los pesos de los alumnos y alumnas de su división (1. de 18.4 Actividades y Ejercicios) para obtener tablas de porcentajes acumulados de los pesos de varones y mujeres por separado. Construya diagramas de tallo y hojas de los pesos de alumnos y alumnas por separado. Obtenga los percentiles del 3, 10, 25, 50, 75, 90 y 97 %.

6. Utilice los pesos de los alumnos y alumnas de los otros dos años (2. de 18.4 Actividades y Ejercicios) para obtener tablas de porcentajes acumulados de los pesos de varones y mujeres por separado. Construya diagramas de tallo y hojas de los pesos de alumnos y alumnas por separado. Obtenga los percentiles del 3, 10, 25, 50, 75, 90 y 97 %.

7. Grafique los percentiles 3, 10, 25, 50, 75, 90 y 97 del peso para las mujeres obtenidos para cada una de las divisiones en un único gráfico, en función de la edad, de la siguiente manera:

   En el eje horizontal la mediana de la edad de las mujeres de cada año.
   En el eje vertical los percentiles del peso.

   Una los tres puntos de cada percentil.

8. Grafique los percentiles 3, 10, 25, 50, 75, 90 y 97 del peso para los varones obtenidos para cada una de los años en un único gráfico, en función de la edad, en forma similar al punto anterior. Una los tres puntos de cada percentil.

20. Curvas de densidad

Poderosa herramienta para describir la distribución de los datos.

Hemos desarrollado un conjunto de herramientas para describir la distribución de los datos: tablas de frecuencias, histogramas, diagramas tallo-hoja, cálculo de medidas resumen (media, mediana, desvío estándar, distancia intercuartil, percentiles), gráfico de caja y brazos. Algunas veces estas herramientas tienen inconvenientes:

- un diagrama tallo-hoja no es práctico para conjuntos con muchos datos.
- las tablas de frecuencias, así como sus representaciones gráficas (los histogramas), eliminan los detalles y dependen de la longitud de los intervalos de clase.
- las medidas resumen (media, mediana, desvío estándar, distancia intercuartil, percentiles) muestran aspectos parciales de los datos.

¿Es posible describir la distribución de los datos en forma completa mediante una única expresión?

La respuesta es: ¡Depende!

¿De qué depende?

Si estamos dispuestos a **describir el patrón general de los datos,** omitiendo los atípicos, **la respuesta es sí.**

Esa respuesta la provee la expresión de una curva - **un modelo matemático, curva de densidad** - para la distribución de los datos.

En la sección 17.2 presentamos algunos patrones especiales que pueden presentar los histogramas mediante curvas. Las expresiones de dichas curvas son precisamente los modelos que necesitamos. Se trata de **descripciones matemáticas idealizadas;** constituyen poderosas herramientas para describir la distribución de los datos. Son especialmente útiles cuando de trata de describir una cantidad muy grande de observaciones.

Podemos establecer un paralelo con la física del movimiento de los cuerpos. La ecuación de la recta describe un movimiento rectilíneo uniforme; pero, ningún desplazamiento real será perfectamente rectilíneo y uniforme. Si graficamos distancia en función del tiempo, con valores medidos de un desplazamiento real, los puntos no caerán exactamente sobre una recta, pero la recta es una buena descripción del movimiento cuando la velocidad es pareja y el desplazamiento es en una única dirección.

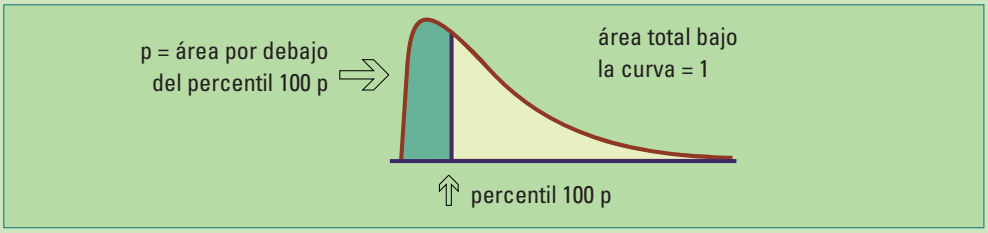
De la misma manera, como la recta es una de las muchas curvas requeridas para describir el desplazamiento de un objeto en función del tiempo, la Curva de Gauss o curva Normal es uno de los tipos de curvas que pueden utilizarse para describir los diferentes tipos de variabilidad de los datos.

□ 20.1. Medias resumen en curvas de densidad

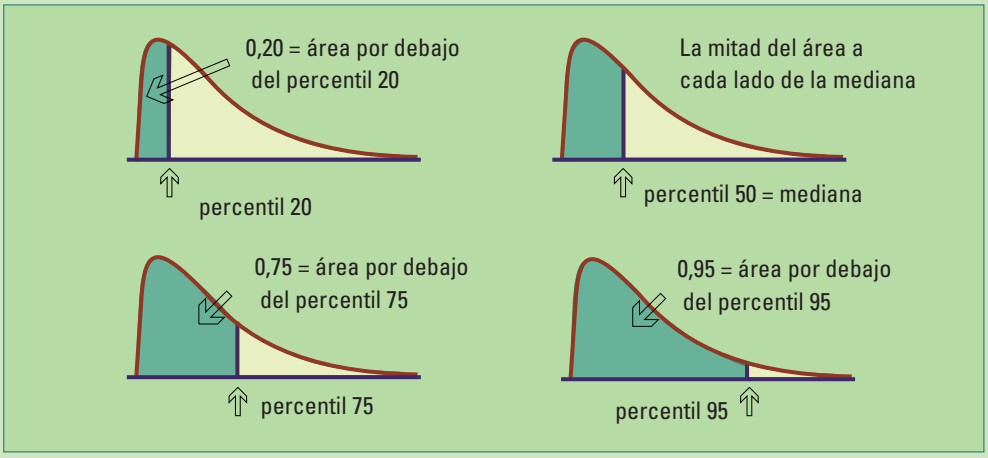
Las curvas de densidad son histogramas idealizados. Las medidas de posición y dispersión se aplican tanto a curvas de densidad como a conjuntos de datos.

Consideremos los percentiles en primer término.

Sabemos que una proporción p de observaciones está por debajo del percentil 100 p.



El percentil 100 x p de una curva de densidad es el punto sobre el eje horizontal para el cual queda a su izquierda el 100 x p % del área bajo la curva, o una proporción p.



En una curva de **densidad simétrica** es fácil **ver "a ojo" donde se encuentra la mediana,** el punto que divide al área en dos partes iguales.

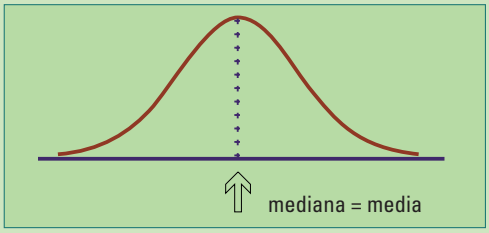


Figura 20.3. *Media y mediana en una curva de densidad simétrica.*

Para una **curva simétrica, la media, el punto de equilibrio coincide con la mediana** que divide el área en dos partes iguales (figura 20.3.).

Como parte de la idealización inherente a un modelo matemático, las curvas de densidad simétricas son "perfectamente simétricas" aunque los datos reales rara vez presenten una simetría perfecta.

Para cualquier curva general **no es fácil hallar a ojo** la mediana, la media y los percentiles. Pero es posible utilizar integrales para obtenerlos. Las integrales son herramientas de análisis matemático que permiten obtener el área por debajo de una curva cuando se conoce la expresión de la misma. No lo haremos aquí.

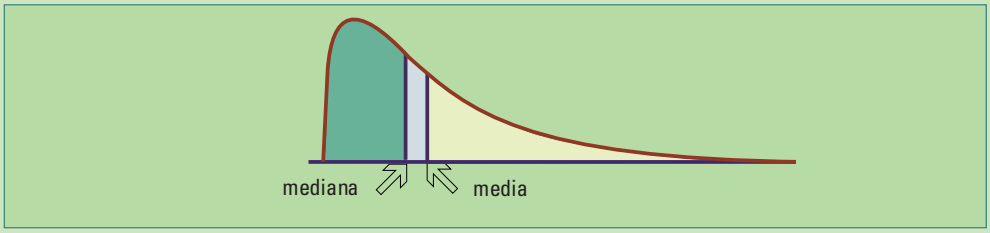


Figura 20.4. *Media y mediana en una curva de densidad asimétrica a derecha.*

La media es el punto de equilibrio de una vara sin peso sobre la que se colocan en cada punto correspondiente al valor de cada dato, pesos idénticos (sección 18.1.1.). La **vara no queda en equilibro** si se apoya en cualquier otro punto. Esta interpretación se extiende a curvas de densidad.

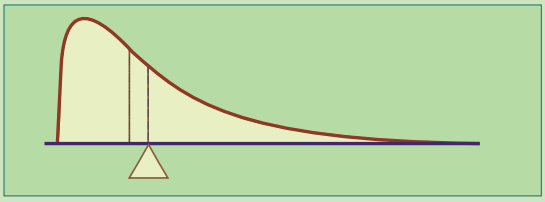


Figura 20.5. *La media es el punto donde la curva de densidad quedaría en equilibrio.*

En una curva asimétrica la media (el punto de equilibrio) es arrastrado hacia la cola larga de la distribución más que la mediana (figuras 20.4 y 20.5). Hallar a ojo la media en una curva asimétrica es más difícil que la mediana, pero la podemos obtener mediante integrales (no lo haremos aquí).

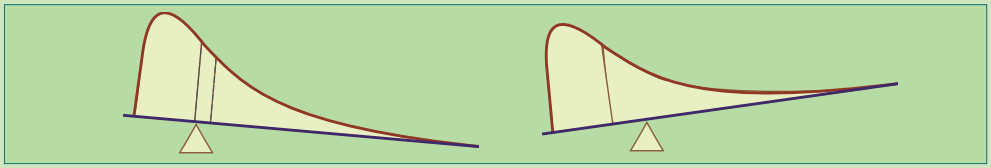


Figura 20.6. *La curva no queda en equilibrio cuando se apoya en un punto diferente de la media.*

La parte inferior de la figura 20.6 ilustra que **la curva no queda en equilibrio cuando se apoya en la mediana.** El área bajo la curva del lado derecho de la mediana "pesa más". Decimos que la distribución tiene **cola pesada a derecha.**

Media y mediana de una curva de densidad: La mediana es el punto que divide el área bajo la curva en dos partes iguales. La media es el punto de equilibrio o centro de gravedad, sobre el cual quedaría en equilibrio si se construyera con un material sólido.

Para calcular la mediana a ojo tratamos de dividir el área en dos partes iguales. Para hallar los **cuartiles,** tratamos de dividir el área por debajo de la curva de densidad en 4 partes iguales (figura 20.7).

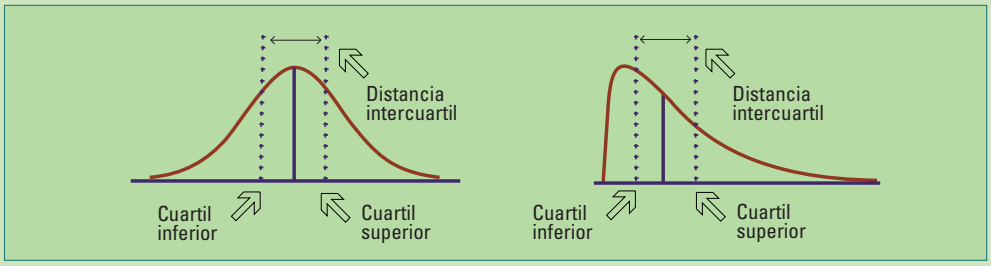


Figura 20.7. *Los cuartiles, la mediana y la distancia intercuartil en una curva simétrica y en una curva asimétrica a derecha.*

La distancia intercuartil es la diferencia entre el cuartil superior y el inferior (también llamados tercer y primer cuartil).

Los cuartiles, por lo tanto, la mediana y la distancia intercuartil, pueden calcularse en forma aproximada a ojo para cualquier curva de densidad. Esto no ocurre con el desvío estándar (18.2.3), que no es una medida natural para la mayoría de las distribuciones. Cuando es necesario, el desvío estándar correspondiente a una curva de densidad, también (como dijimos para los percentiles), puede calcularse utilizando integrales. No se desarrollará esta forma de calcular en este libro.

La curva de densidad es una descripción idealizada de la distribución de los datos, por eso distinguimos la **media** y el **desvío estándar** de una **curva de densidad** de los números

y s (media muestral y desvío estándar muestral respectivamente) y se obtienen a partir de un conjunto de datos. La forma habitual de indicar la media de una distribución idealizada es mediante la letra griega "mu": **μ.** El desvío estándar se indica por **σ**, la letra griega "sigma".

□ 20.2. Ventajas de la curva Normal

¿Para qué sirve tener un conjunto de datos cuyo histograma es aproximadamente Normal? ¿Por qué se habrá enamorado Galton de la curva gaussiana? (sección 17.1)

Hemos visto que es muy útil reemplazar un conjunto de datos por unos pocos valores, las medidas resumen, para describir sus características generales.

> Cuando los datos tienen una **distribución Normal** la distribución de los mismos se puede reducir a **dos números:** la media y el desvío.

En general, es deseable tener **patrones** que representen **la forma** de la distribución de **los datos** y que permitan además representar sus características más importantes mediante **una cantidad pequeña de números.**

20.2.1. Histogramas y la curva Normal

Pensemos primero en un conjunto con **muchísimos datos.** Podemos construir histogramas con intervalos de distinta longitud y superponerle una Curva de Gauss. Como los datos son muchísimos podemos achicar la longitud de los intervalos de clase tanto como queramos.

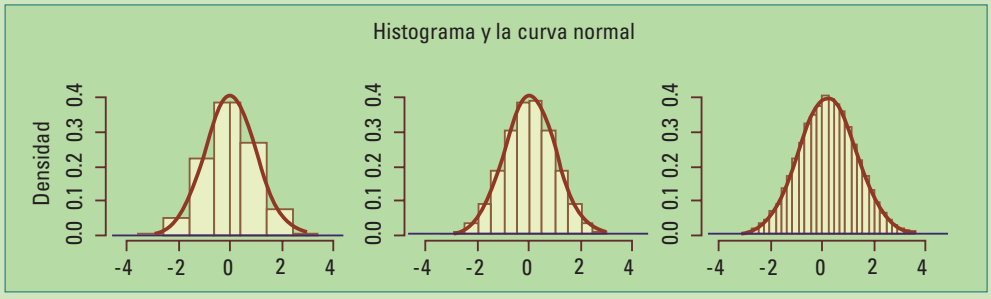


Figura 20.8. *Superposición de la curva de Gauss a histogramas con intervalos de longitud decreciente.*

A medida que se achica la longitud de los intervalos de clase mejora la aproximación de la campana de Gauss.

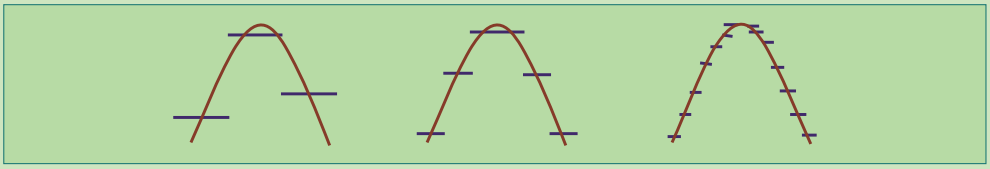


Figura 20.9. *Detalle, en un sector ampliado, de la arista superior de los rectángulos de clase de los histogramas y su aproximación creciente a la curva de Gauss.*

El primero de los histogramas muestra escalones resultantes del agrupamiento de los datos en intervalos de clase, pero estas irregularidades disminuyen al reducir la longitud de los intervalos (figura 20.9). La curva de Gauss en la figura 20.8 describe la distribución de los datos en forma más precisa que los histogramas.

Cuando un histograma se grafica utilizando las frecuencias en el eje vertical, la escala depende de la cantidad de datos. Si se utilizan frecuencias relativas o porcentajes esto es menos arbitrario y el **área del rectángulo es proporcional** a la frecuencia relativa.

Es más natural que el **área del rectángulo** sea **igual a la frecuencia relativa.** Lo logramos si en el **eje vertical** graficamos la **frecuencia relativa dividida la longitud del intervalo.** Esto se llama **escala de densidad** y permite tener la misma escala vertical aunque cambiemos la longitud de los intervalos y el área total de los rectángulos del histograma siempre 1 (ó 100 si las frecuencias relativas están expresadas como porcentajes).

La curva que describe la forma de la distribución se llama **curva de densidad** y tiene área 1. El **área bajo la curva** sobre cualquier **intervalo de valores** del eje horizontal es **la proporción de observaciones** que caen en ese intervalo.

En la figura 20.8 la escala de densidad va de 0 a 0,4 en los 3 histogramas y en la curva. Podemos calcular en forma aproximada el área total pues la figura es aproximadamente un triángulo cuya base tiene longitud aprox. 5 y la altura es aprox. 0,4. El cálculo aproximado resulta:

$$\frac{\text{long de la base x altura}}{2} = \frac{5 \times 0{,}4}{2}$$

$$\frac{\text{long de la base x altura}}{2} = 1$$

20.2.2. Media y desvío de la curva normal

Todas las curvas Normales son simétricas, tienen un único pico y forma de campana.

Sus colas caen rápidamente, por lo tanto no se esperan valores muy alejados (outliers). La media, la mediana y el pico coinciden en el centro de la curva.

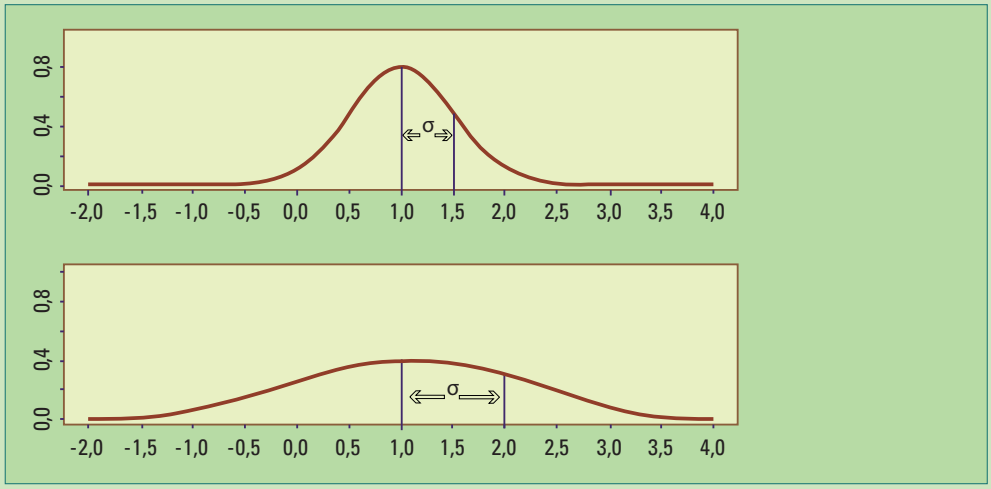
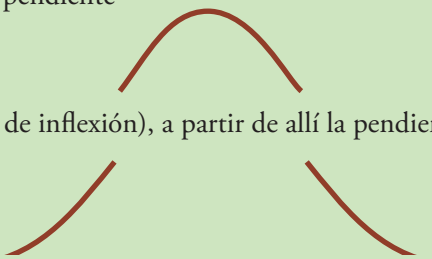


Figura 20.10. *Dos curvas normales con media 1 y distintos desvíos.*

Otra propiedad importante de la curva de densidad Normal es poder localizar el desvío estándar a ojo: a medida que nos movemos en ambas direcciones desde el centro **μ** de la curva, ésta aumenta su pendiente



hasta un punto (punto de inflexión), a partir de allí la pendiente empieza a disminuir

Los dos puntos en los cuales ocurre este cambio de curvatura están localizados a una distancia **σ** a cada lado del centro **μ**.

μ es la media
σ es el desvío

Recuerde, **μ** y **σ** solos no determinan la forma de una distribución en general. Éstas son propiedades de las distribuciones gaussianas.

Pero...

> Cuando los valores de una variable tienen distribución Normal, **sólo dos números** alcanzan para determinar la distribución de todos sus valores. Esos dos números, **μ** y **σ** son **los parámetros** de la distribución Normal.

Pero...

Pequeños alejamientos de la distribución Normal pueden llevar a que **μ** y **σ** no signifiquen nada.

Un detalle extra:

Siempre es más seguro utilizar los percentiles porque tienen el mismo significado en todo tipo de distribuciones. Cuando no hay grupos aislados, las 5 medidas resumen: mínimo, cuartil inferior, mediana, cuartil superior y máximo, son en general una buena representación de los datos.

20.2.3. Otras característcas interesantes

Si un histograma se aproxima por una curva Normal podremos decir algunas cosas más que, simplemente, caracterizar su media y su desvío.

Podremos establecer **criterios** sobre **donde se encuentra** la mayoría de los **datos.**

Los **criterios** que veremos a continuación se utilizan cuando **podemos suponer** que los **datos** tienen una distribución **aproximadamente Normal,** por la naturaleza del experimento, con **μ** y **σ** conocidos. Cuando no son conocidos se estiman mediante la media muestral (x) y el desvío estándar muestral (s), respectivamente, tal como vimos en las secciones 18.1.1 y 18.2.3:

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

donde los x_1 son los datos y n es la cantidad total ($x_1, x_2, \ldots x_n$)

Utilizaremos estas reglas en el próximo capítulo sobre control de calidad.

Si una distribución tiene una forma gaussiana, entonces vale la siguiente **regla 68-95-99,7 :**

- Aproximadamente el 68% de los valores se encuentran dentro de 1 desvío estándar (**σ**) de la media (**μ**). Es decir que bastante más de la mitad de los valores están comprendidos dentro del intervalo (μ - **σ**, μ +**σ**) ó μ ± **σ** (figura 20.10).
- Cerca del 95% de los valores están entre la media menos 2 veces el desvío estándar y la media más 2 veces el desvío estándar, o sea dentro de **σ**μ ± 2 (figura 20.11).

- El 99,7% (casi todos) de los valores se encuentran en el intervalo (**μ** - 3**σ**, **μ** + 3**σ**), o sea dentro de μ ± 3 **σ** (figura 20.12).

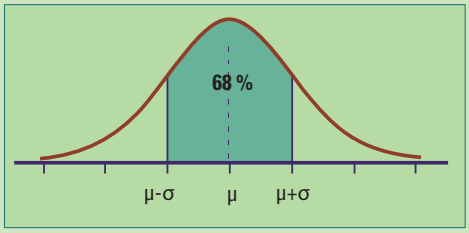


Figura 20.11.

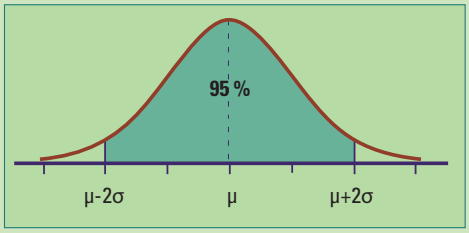


Figura 20.12.

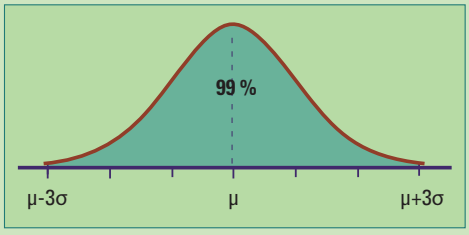


Figura 20.13.

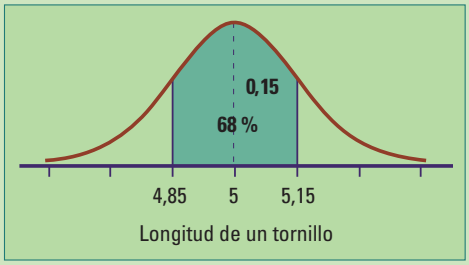


Figura 20.14

Como la mayoría de los valores en una distribución Normal se encuentran en la zona central, alrededor de la media (**μ**), el 68% de los valores están a una distancia no mayor al desvío. Al alejarnos un desvío más de la media, hacia los dos lados, agregamos más valores (un 14% a cada lado); pero, son menos porque se trata de una zona de menor concentración de datos. Obtenemos así el intervalo (**μ**-2**σ**, **μ**+2**σ**) allí se encuentra aproximadamente el 95% de los valores.

Alejándonos otro desvío más, agregamos apenas un 2% de cada lado, llegando a 99.7 % en el intervalo (**μ**-3**σ**, μ+3**σ**).

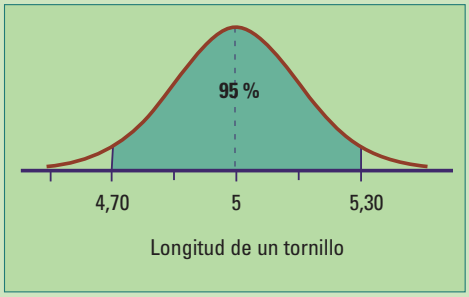
Ejemplo: Un taller metalúrgico produce remaches cuya longitud debe ser de 5 cm con una tolerancia de 0,3 cm (5 ± 0,3 cm). Por lo tanto, las longitudes aceptables están en el intervalo (4,7; 5,3). Interesa evaluar la calidad de la producción teniendo en cuenta este requerimiento.

Como suele ocurrir en esta industria, si la producción se realiza en condiciones normales tendremos muchos remaches cuya longitud esté cerca de 5 cm y pocos alejados; las longitudes tendrán una distribución gaussiana.

Supongamos que los registros históricos de la producción de estos remaches, con el mismo equipamiento, muestran que la media de las longitudes es efectivamente 5 cm con un desvío de 0,15 cm (**μ**=5 y **σ**=0,15).

Luego:

- Aproximadamente el 68% de los valores se encuentran dentro de1 desvío estándar (σ) de la media (μ). Es decir que bastante más de la mitad de los valores están comprendidos dentro del intervalo (μ-σ ,μ+σ) ó μ ± σ (figura 20.11).
- Cerca del 95% de los valores están entre la media menos 2 veces el desvío estándar y la media más 2 veces el desvío estándar, o sea dentro de μ ± 2**σ** (figura 20.12).



- El 99,7% (casi todos) de los valores se encuentran en el intervalo (μ-3**σ** ,μ+3**σ**), o sea dentro de μ ± 3**σ** (figura 20.13).

Casi el 5% de los remaches tendrá una longitud por fuera de los límites especificados. El encargado de control sabrá si este porcentaje de remaches a desechar (scrap) es admisible. Si no lo es deberán modificarse los procesos de producción, hasta que se logre un perfil de calidad adecuado.

20.2.3.1. Regla 68 - 95 - 99,7

Supongamos que un conjunto de datos (x_1, x_2 ... x_n) tiene una distribución gaussiana con media **μ** y desvío estándar **σ**. El conjunto **de datos estandarizados** (z_1, z_2 ... z_n), o "puntajes z", que se obtiene restando **μ** y dividiendo por **σ** ($z_i = \frac{x_i - \mu}{\sigma}$), tendrá una distribución Normal Estándar (figura 20.15).

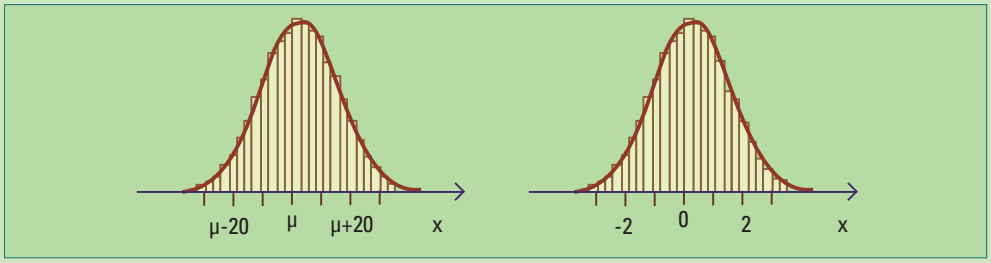


Figura 20.15. *Histogramas de un conjunto de datos en su escala original (x) y transformados en puntaje z.*

Recordemos (sección 17.1.1) que la curva Normal Estándar, también llamada N(0,1), está **dada por** $f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ **, y no depende de parámetros desconocidos.**

Las áreas bajo esta curva se pueden calcular.

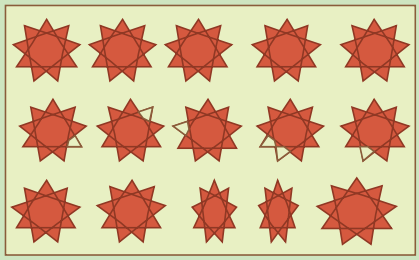
¡Conocer el área bajo la curva Normal Estándar sobre cualquier intervalo, permite conocerla para todos los intervalos bajo cualquier curva Normal!

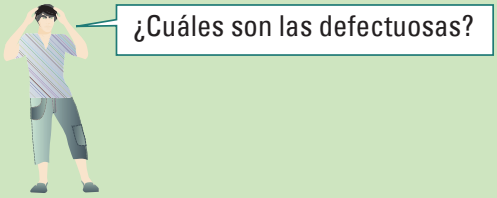
En particular, las áreas sobre los intervalos, (**μ-σ**; **μ+σ**); (**μ-2σ**; **μ+2σ**) y (**μ-3σ**; **μ+3σ**) bajo la curva N (**μ**, **σ**) son iguales a las áreas sobre los intervalos (-1,1) ; (-2,2) y (-3;3) bajo la curva N(0,1).

Los valores 68; 95 y 99,7 para los porcentajes de áreas por encima de los intervalos (-1,1); (-2,2) y (-3;3) bajo la curva Normal Estándar son aproximados. Valores más precisos son: 68,27; 95,45 y 99,73 respectivamente.

21. Control de calidad

Idealmente todas las unidades fabricadas serán idénticas y perfectas.





¿Para qué se realiza el control de calidad?

El control estadístico de calidad de productos y servicios tiene como objetivo **reducir su variabilidad** e idealmente eliminar sus defectos. Como un ideal, se busca que todas las unidades fabricadas sean idénticas y perfectas. En la práctica real se consigue reducir los desperdicios, minimizar los reprocesamientos y mejorar la opinión del cliente, tratando de hacer las cosas bien de una primera vez.

En el desarrollo de un producto puede servir en experimentos para comparar materiales, componentes o ingredientes. Durante el proceso de producción y distribución, los métodos estadísticos permiten identificar problemas que atentan contra la calidad buscada.

Hemos visto (secciones 17.1.1 y 20.2.3) que diferentes piezas, correspondientes a un mismo producto, pueden parecer iguales pero al medir sus características detalladamente se encuentran diferencias. Por más cuidado que se tenga en la calibración de las máquinas, se controlen los factores ambientales, se vigilen los materiales y se capaciten los operarios, **las piezas no serán idénticas.** Se trata de una **variabilidad** natural o **aleatoria.** Puede considerarse como un "ruido de fondo" inevitable.

Cuando el ruido de fondo de un proceso de producción es relativamente pequeño se lo considera aceptable. Cuando un proceso sólo está afectado por esa variabilidad aleatoria decimos que se trata de un "sistema estable" y "bajo control estadístico" o simplemente "en control".

Walter A. Shewhart identificó las variaciones que se presentan cuando el proceso productivo opera normalmente. Son el resultado de muchas causas generalmente pequeñas e inevitables, que ocurren todo el tiempo. Las llamó variaciones "debidas a **causas comunes"**, en contraposición con un segundo tipo de variaciones, las debidas a **"causas especiales o asignables"** y que ocurren de vez en cuando.

Walter Andrew Shewhart. (1891-1967). Físico, matemático y estadístico norteamericano, también conocido como el padre del control estadístico de la calidad.

No es posible –ni tiene sentido- perder tiempo en averiguar la causa de una variación debida a causas comunes cuando el proceso ya satisface las especificaciones. Sí es útil dedicarse a las debidas a causas especiales; usualmente provienen de tres fuentes: de los equipos, del operador o de las materias primas utilizadas.

Por ejemplo, se habla de causas especiales cuando la calidad del producto es afectada por haberse utilizado materias primas defectuosas o por el accionar inadecuado de los operarios que, por cansancio ó distracción, en muchas oportunidades continúan la producción sin advertir un desajuste en su máquina.

No son causas comunes. Se trata de **causas asignables:** máquinas mal ajustadas, materias primas defectuosas, errores del operador, software incorrecto, etc. Las piezas producidas bajo estas condiciones anómalas no tendrán su variabilidad habitual. Esta variabilidad extra suele ser grande, en comparación con el ruido de fondo, y representa un nivel inaceptable del rendimiento del proceso.

Cuando un proceso de producción opera en presencia de **causas asignables** decimos que **está fuera de control.**

Al controlar un proceso interesa restringir la variación únicamente a la debida a las causas comunes; **las causas asignables deben ser detectadas y eliminadas.** Cuando **la única variación presente** es debida a **causas comunes,** y no a una causa asignable decimos que **el proceso opera en estado bajo control** o que **está en control.** Un proceso en control es **estable en el tiempo** respecto a las variaciones y no muestra indicaciones de causas extrañas.

Un proceso en control (o bajo control) es un proceso estable.

No significa necesariamente que se satisfagan las especificaciones del producto.

La variabilidad debida a causas comunes puede exceder los límites de tolerancia del producto. Puede ocurrir que la proporción de piezas defectuosas sea mayor a lo tolerable en términos económicos. En ese caso se deberán introducir modificaciones al proceso.

Cuando ya se ha establecido un proceso que cumple con las especificaciones de tolerancia del producto, el problema principal consiste en determinar cuándo el proceso está fuera de control.

Los gráficos de control, también llamados cartas de control de Shewhart, permiten reconocer situaciones en las cuales las causas asignables pueden estar afectando negativamente la calidad de un producto. Se trata de una secuencia de puntos obtenidos de muestras de piezas tomadas a través del tiempo. Son los valores de algún estadístico, tal como la media de la longitud o la proporción de piezas defectuosas.

□ 21.1. Gráficos de control

Un gráfico de control muestra en el eje vertical los valores de los datos y en el horizontal el orden en que fueron recogidos a través del tiempo. Es esencial, para este tipo de

gráfico que las muestras hayan sido tomadas en forma sucesiva en el tiempo, pues es esa evolución -de alguna o algunas características de un producto- lo que interesa controlar.

Tiene una línea horizontal a la altura de un **valor central,** ésta puede provenir de las especificaciones del producto o de valores históricos y dos líneas -una en el **límite inferior de control** (LIC) y otra en el **límite superior de control** (LSC)- para indicar cuan lejos por encima o por debajo del valor objetivo se espera que se obtengan los valores de las muestras.

Se grafican pesos, volúmenes, cantidades o más frecuentemente pesos promedio, el promedio de volúmenes o cantidades promedio (proporciones). Si los puntos caen dentro de los límites inferior y superior se considera que el proceso se encuentra en el estado de control estadístico.

Ejemplo: Las bolsas de galletitas siempre parecen tener menos unidades que las que debieran. Supongamos que un fabricante está llenando bolsas con galletitas de agua, y el valor objetivo es de 50 piezas por bolsa, con LIC = 45 piezas y LSC = 55 piezas. Supongamos, además, que 8 bolsas de galletitas son seleccionadas mediante un muestreo sistemático. Una de cada 200 bolsas, de una línea de producción, son inspeccionadas obteniéndose los siguientes resultados: 49 galletitas, 47 galletitas, 51 galletitas, 49 galletitas, 46 galletitas, 53 galletitas, 48 galletitas, y 55 galletitas.

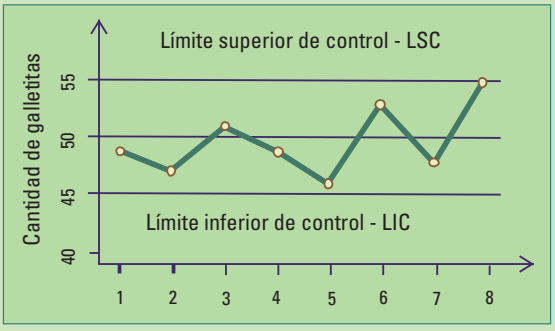


Fig. 21.1. *Gráfico de control para la cantidad de galletitas.*

El proceso que muestra la figura 21.1. parece estar operando en control, al menos por el momento.

> Un gráfico de control puede indicar una condición fuera de control cuando uno o más puntos caen más allá de los límites o presenta algún patrón de comportamiento no aleatorio.

21.1.1. ¿Cómo se establecen el valor central y los límites en un gráfico de control?

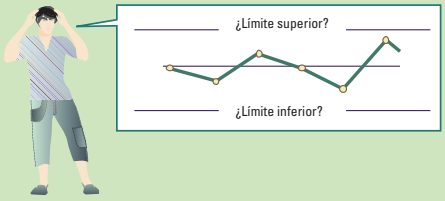
La idea detrás de los gráficos de control es obtener límites, fijar cotas para toda la variabilidad aleatoria, excluyendo la variabilidad asignable no deseada. De esta forma, las causas asignables tenderán a dar valores fuera de los límites de control mientras que la variabilidad aleatoria tenderá a generar puntos que se encuentran dentro de esos límites.

En general el valor central o **valor objetivo** es establecido por las especificaciones del producto (50 en el ejemplo de las galletitas). Otras veces es el promedio de muchos valores

(datos históricos) de la característica de interés, obtenidos con el proceso en estado de control. Por ejemplo, en una producción de almohadones el valor central podría ser su peso promedio cuando la variabilidad presente se debió únicamente a causas comunes.

Las especificaciones de los **límites de control** dependen de los límites de tolerancia del producto y de qué proporción de artículos está dispuesto a perder el fabricante.

Supongamos que el fabricante puede tener hasta un **5% de artículos** fuera de las especificaciones. Para el caso del llenado de bolsas de galletitas, esto significa que el 95% de las bolsas deben contener entre 45 y 55 galletitas. Si la distribución de la cantidad de galletitas por bolsa puede aproximarse por la Normal (ya vimos que suele ocurrir), cerca del 95% de los valores se encontrará dentro de 2 desvíos de la media (sección 20.2.3). El intervalo de cantidades aceptables [45;55] tiene longitud 10 y debe corresponder a 2 desvíos para que se cumpla que el 95% de las bolsas contengan entre 45 y 55 galletitas (sección 20.2.3). Luego la cantidad de galletitas por bolsa producida puede tener un desvío menor o igual a 10/2=5.



Si ahora el fabricante es más exigente, admitiendo sólo un 0,3% de productos defectuosos, entonces el 99.7% de las bolsas deben contener entre 45 y 55 galletitas, el intervalo [45; 55] de longitud 10 debe corresponder a 3 desvíos y el desvío debe ser menor o igual a 10/ 3=3,33. El proceso de fabricación ahora requiere un desvío de 3,33 galletitas por bolsa.

Una vez logrados los requerimientos de calidad se debe monitorear el proceso para garantizar que se mantenga estable.

21.1.1.1. Límites de control 3 - sigma

Es habitual colocar los límites de control a una distancia de más y menos 3 desvíos, a partir de la línea central, de la variable graficada. Estos límites se conocen como **límites de control 3 – sigma.**

- Límite inferior de control (LIC) = $\boldsymbol{\mu - 3\sigma}$
- Límite superior de control (LSC) = $\boldsymbol{\mu + 3\sigma}$

Si la variable graficada tiene distribución Normal con media igual al valor central del gráfico el **99,7% de los valores** (casi todos) **caerán dentro de los límites de control 3 – sigm**a (ver sección 20.2.3). Con estos límites tendremos un 0,3 % de falsas alarmas.

Ejemplo: Una fábrica produce garrafas de gas comprimido de uso doméstico con 20 kg de capacidad nominal. La capacidad volumétrica es una de las características importantes de las mismas. Cada hora se selecciona una garrafa de la línea de producción y se mide su capacidad volumétrica interna en dm^3. En una jornada de 16 horas (2 turnos) se obtuvieron los siguientes valores: 45,91; 46,34; 47,52; 46,52; 47,15; 47,15; 47,99; 46,81; 45,70; 47,25; 45,85; 48,14; 47,56; 48,01; 46,55; 47,27.

Se sabe que la capacidad volumétrica interna de una garrafa, cuando el proceso de producción opera en condiciones de control, es una variable con distribución Normal con media 47 dm^3 y desvío estándar de 0,666 dm^3.

> No confundir la capacidad nominal de 20 kg con la capacidad volumétrica que se mide en dm^3 y su valor está alrededor de 47.

La carta de control de 3 - σ (3 desvíos) tendrá los siguientes límites:

- Límite inferior de control (LIC) $= 47\text{-}3 \times 0{,}666$
  $= 45{,}002$
  $= 45$
- Límite superior de control (LSC) $= 47+3 \times 0{,}666$
  $= 48{,}998$
  $= 49$

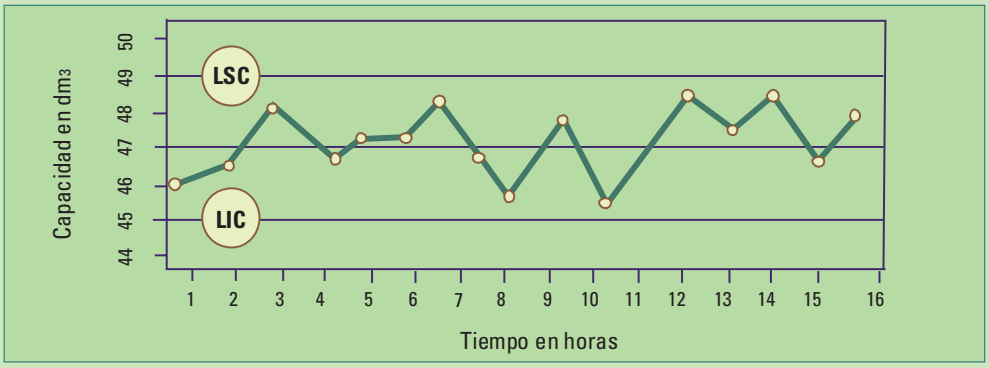


Figura 21.2. *Gráfico de control para la capacidad volumétrica de garrafas de uso doméstico.*

La figura 21.2 muestra los siguientes valores: 45,91; 46,34; 47,52; 46,52; 47,15; 47,15; 47,99; 46,81; 45,70; 47,25; 45,85; 48,14; 47,56; 48,01; 46,55; 47,27 de las capacidades volumétricas de 16 garrafas, seleccionadas una cada hora en una jornada laboral de 2 turnos. Los puntos se encuentran dentro de los límites de control; el proceso se encuentra bajo control estadístico.

21.1.1.2. Estimación de los parámetros del proceso

Cuando un gráfico de control muestra un proceso bajo control estadístico es posible utilizar los puntos del mismo para estimar la media (**μ**), el desvío (**σ**) y la fracción que no cumple con los requerimientos. Esto permite tomar decisiones respecto a realizar o no modificaciones en el ciclo de producción y actualizar los límites de control si fuera necesario.

□ 21.2. Gráficos de control $\bar{x}$ (equis barra)

La mayoría de los procesos productivos son monitoreados seleccionando muestras y calculando promedios. En los gráficos de control, en vez de valores individuales se grafican **promedios,** calculados a partir de los datos de subgrupos o **muestras** de artículos.

Los gráficos de control que utilizan promedios se denominan "Gráficos de control equis barra".

21.2.1. ¿Cómo se eligen las muestras de artículos?

Los subgrupos de artículos se eligen, en lo posible, de manera que contengan la variabilidad natural del proceso y excluyan la variabilidad debida a causas asignables. Describiremos dos tipos de criterios para elegirlos.

21.2.1.1. Unidades cercanas en el tiempo

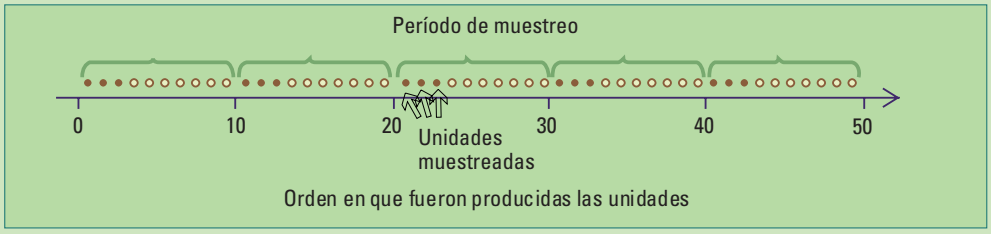


Figura 21.3. *Esquema de muestreo de 3 unidades cercanas en el tiempo cada 10 producidas (período de muestreo). Los puntos sólidos representan las unidades muestreadas.*

Cada subgrupo está formado por unidades producidas al mismo tiempo o casi al mismo tiempo. De esta manera se espera que un cambio en el proceso ocurra entre subgrupos y no dentro de cada subgrupo. Se obtiene así como una foto del proceso en cada instante en que se tomó la muestra. Este tipo de muestreo suele utilizarse para detectar corrimientos en la media del proceso.

La figura 21.3 muestra un esquema de **muestreo sistemático** de unidades cercanas en el tiempo. Se eligen 3 artículos sucesivos al comienzo de cada período de 10 unidades. Los puntos muestreados aparecen como puntos negros sólidos. Es mejor elegir al azar, dentro de cada período, el momento a partir del cual se seleccionan los 3 artículos sucesivos.

21.2.1.2. Unidades representativas de un período completo

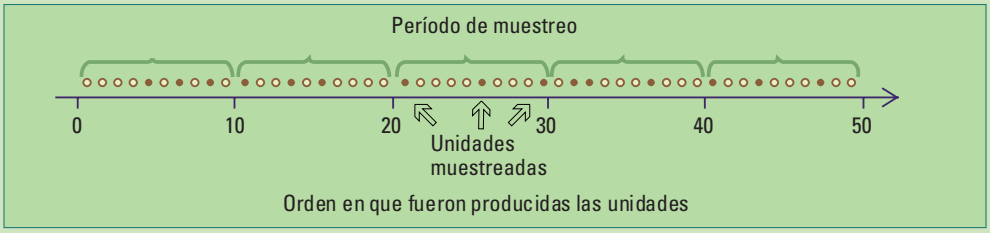


Figura 21.4. *Esquema de muestreo de 3 unidades representativas de todo el período de muestreo. Los puntos sólidos representan las unidades muestreadas.*

Cada muestra está formada por unidades representativas de todas las unidades producidas desde que se tomó la última muestra. Se trata de una muestra aleatoria de toda la salida del proceso sobre el intervalo de muestreo.

Este procedimiento suele utilizarse para tomar decisiones respecto de la aceptación o no de todas las unidades producidas en ese período.

La figura 21.4 muestra un ejemplo en el cual 3 unidades fueron elegidas al azar dentro de cada período de 10 unidades.

□ 21.2.2. ¿Cómo se calculan los límites de control tres sigma en un gráfico $\bar{X}$?

En un gráfico de control $\bar{X}$ (equis barra) **se grafican promedios** y para los límites de control se utilizan los **errores estándar.**

Un **error estándar** es el **desvío estándar** de las **medias muestrales.**

Esto significa que los límites de control 3 - sigma se establecerán como el valor objetivo más menos 3 veces el error estándar.

El error estándar se calcula como el desvío estándar (σ) dividido la raíz cuadrada de n, siendo n el tamaño de la muestra: Error estándar = $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

El error estándar siempre es menor que el desvío estándar. Esto es porque los promedios utilizan más información que un único dato y por lo tanto varían menos entre una muestra y la siguiente. Vimos un ejemplo de esa situación en el Capítulo 10. Allí también el error aleatorio se reducía al aumentar el tamaño de la muestra de acuerdo con $\frac{1}{\sqrt{n}}$.

Los límites de control tres sigma para un gráfico de control $\bar{x}$ son:

- Límite inferior de control (LIC) = $\mu - 3\,\sigma_{\bar{x}}$
- Límite superior de control (LSC) = $\mu + 3\,\sigma_{\bar{x}}$

O sea:

- Límite inferior de control (LIC) = $\mu - 3\sigma/\sqrt{n}$
- Límite superior de control (LSC) = $\mu + 3\sigma/\sqrt{n}$

Ejemplo: Continuemos con datos de capacidad volumétrica (sección 21.1.1.1.) para un proceso con media 47 dm^3 y desvío estándar de 0,666 dm^3. Utilizamos esta vez una carta de control $\bar{X}$, con n=5. Es decir, se promedia la capacidad volumétrica de 5 garrafas cada hora, durante 16 horas. Los datos son los siguientes:

Capacidad volumétrica de 5 garrafas						
	Garrafa 1	**Garrafa 2**	**Garrafa 3**	**Garrafa 4**	**Garrafa 5**	**Promedio**
Muestra 1	45,36	46,53	47,36	47,27	46,78	46,66
Muestra 2	47,59	46,10	47,10	47,01	47,52	47,06
Muestra 3	47,44	47,91	46,07	47,11	47,97	47,30
Muestra 4	47,84	46,19	47,01	47,43	46,39	46,97
Muestra 5	46,79	48,21	47,37	46,61	46,39	47,07
Muestra 6	48,11	47,45	46,65	48,01	48,02	47,65
Muestra 7	46,66	47,06	47,95	46,51	46,53	46,94
Muestra 8	46,92	47,87	47,05	47,96	47,18	47,40
Muestra 9	46,59	47,45	45,81	46,55	47,22	46,72
Muestra 10	47,28	46,53	48,17	45,93	47,01	46,98

Capacidad volumétrica de 5 garrafas						
	Garrafa 1	**Garrafa 2**	**Garrafa 3**	**Garrafa 4**	**Garrafa 5**	**Promedio**
Muestra 11	47,18	47,12	47,70	47,09	47,27	47,27
Muestra 12	46,58	47,02	45,82	47,35	46,31	46,62
Muestra 13	46,89	47,39	46,33	47,50	48,18	47,26
Muestra 14	46,17	46,89	46,63	45,00	47,46	46,43
Muestra 15	47,82	47,24	46,86	46,01	47,04	46,99
Muestra 16	46,29	47,59	47,40	45,81	47,62	46,94

Tenemos n=5, μ=47 dm^3 y σ=0,66 dm^3, por lo tanto:

- Límite inferior de control (LIC) $= 47 - 3 \times \frac{0,66}{\sqrt{5}}$
  $= 47 - 0,89$
  $= 46,11$

- Límite superior de control (LSC) $= 47 + 3 \times \frac{0,66}{\sqrt{5}}$
  $= 47 - 0,89$
  $= 47,89$

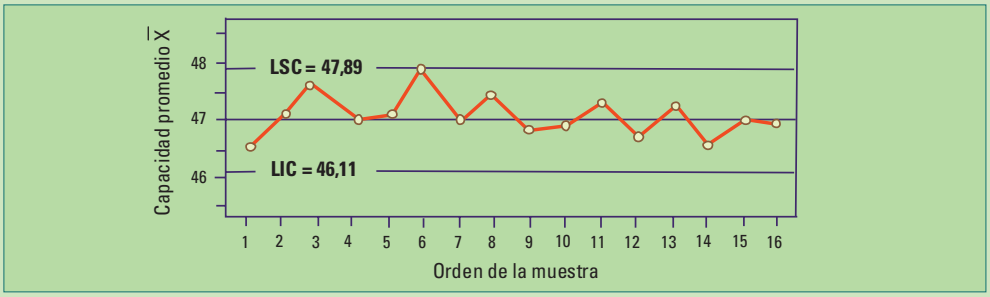


Figura 21.5. *Gráfico de control $\bar{X}$ para la capacidad volumétrica de garrafas, n=5.*

La figura 21.5 no muestra evidencias de que el proceso se haya salido de control, todos sus valores están dentro de las bandas y no aparece ningún patrón no aleatorio.

Nuevo ejemplo: El sector de control de calidad de una fábrica que produce dardos registró los diámetros de 10 submuestras sucesivas de tamaño 4, resultado en los siguientes promedios (en milímetros):

Muestra	1	2	3	4	5	6	7	8	9	10
Promedio	3,01	2,97	3,12	2,99	3,03	3,02	3,1	3,14	3,09	3,2

Registros históricos indican que cuando el proceso opera en control, los diámetros sucesivos tienen distribución gaussiana con media μ=3 y desvío σ=0.1 por lo tanto para n=4 los límites de control 3-sigma son:

$$LIC = \frac{3 - 3\,(0,1)}{\sqrt{4}} \qquad LSC = \frac{3 + 3\,(0,1)}{\sqrt{4}}$$

$$LIC = 2,85 \qquad LSC = 3,15$$

Como la media muestral número 10 se encuentra por encima del límite superior concluimos que hay razones para sospechar que la media de los diámetros de los dardos difiere de 3. Más aún el gráfico de control de la figura 21.6 parece sugerir que a partir de la muestra 6 aumentó la media del diámetro de los dardos.

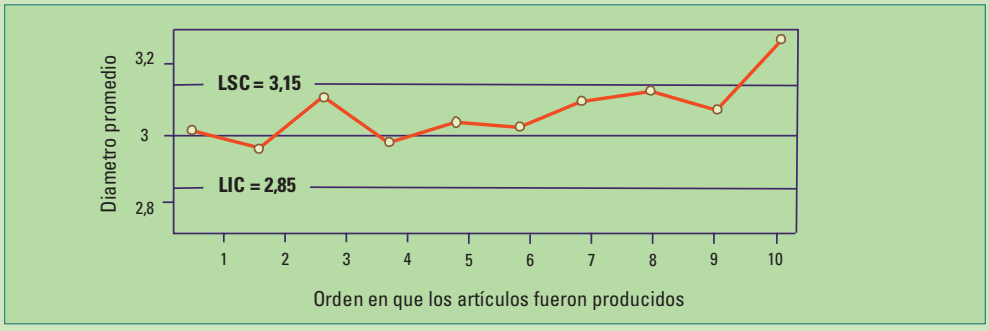


Figura 21.6. *Gráfico de control para diámetros de dardos, de 10 submuestras sucesivas de tamaño 4.*

□ 21.3. Análisis de patrones no aleatorios en cartas de control

Una carta de control puede indicar una condición fuera de control cuando uno ó más puntos caen fuera de los límites de control o cuando los puntos muestran un comportamiento no aleatorio. Por ejemplo, la figura 21.6 muestra un punto fuera de los límites de control pero además desde el punto cuatro hasta el ocho se observa una marcada tendencia creciente con apariencia no aleatoria.

El manual de la empresa Western Electric (Western Electric Handbook (1956)) establece que un proceso está fuera de control cuando se cumple alguna de las siguientes pautas:

1. Hay un punto fuera de los límites 3-**σ**.
2. Dos de tres puntos consecutivos se encuentran del mismo lado fuera de un límite 2-**σ**.
3. Cuatro de 5 puntos consecutivos están se encuentran del mismo lado fuera de un límite 1-**σ**.
4. Ocho puntos consecutivos del mismo lado de la línea central.

Este criterio aumenta la sensibilidad para detectar un proceso fuera de control pero también aumenta la probabilidad de falsa alarma.

La figura 21.7 muestra el gráfico $\bar{X}$ (equis barra) para el ejemplo de los dardos con los límites 1-sigma, 2-sigma y 3-sigma utilizados en el procedimiento Western Electric. A veces estos límites son llamados límites de advertencia. Estos límites dividen el gráfico de control en tres zonas (A, B y C) a cada lado de la línea central. Nótese que los **cuatro últimos puntos** caen en la zona B ó más allá de ella. Luego tenemos en este caso **una doble evidencia de que el proceso no está en control ya que se cumplen las reglas 1 y 3.**

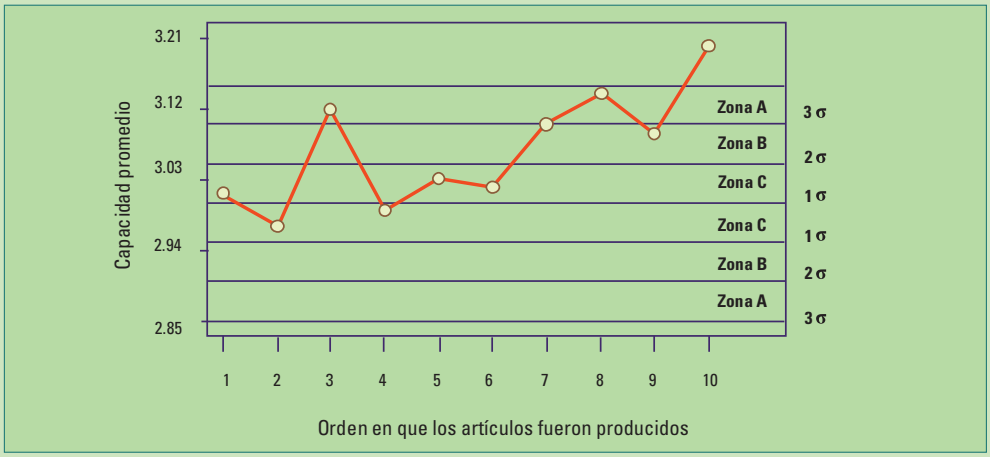


Figura 21.7. *Gráfico de control $\bar{X}$, para diámetros de dardos, con los límites 1-sigma, 2-sigma y 3-sigma.*