# Valores de dispersión

Realicemos un breve resumen de lo que venimos estudiando.

Recopilamos datos para hacer una investigación, luego presentamos esos valores en tablas y gráficos. Estudiamos las medidas que resumen ese conjunto de datos, que son las medidas de posición.

Con las medidas de posición, tenemos más información de un conjunto de datos, pero no es suficiente. Puede suceder que, por ejemplo, la media sea más o menos representativa de los valores de una muestra.

El estudio se completa con las medidas de dispersión, que estudiaremos en esta lectura y nos dan más información sobre los valores de una muestra o población.

Con esto cerramos la estadística descriptiva y también descubrirás lo importantes que son las medidas de variabilidad o dispersión en muchos campos, en especial para hacer control de calidad.

# Valores de dispersión para datos no agrupados

El gerente del sucursal número 5 del banco HSUR está estudiando la necesidad de reforzar el personal de atención al cliente en el Área Cuentas Corrientes desde las 10 horas hasta las 13 horas. Para esto encarga un estudio estadístico a un grupo de expertos. El estudio se realiza durante una semana sobre un total de 50 clientes tomados al azar. Se registra el tiempo de demora en minutos (definido como el tiempo transcurrido desde que el cliente retira su turno en la máquina situada al ingreso del sector hasta que es llamado por la pantalla electrónica) de cada cliente muestreado.

En el siguiente cuadro, se muestran los resultados obtenidos.

A continuación, los tiempos que demoran los clientes de la sucursal del banco HSUR en ser atendidos.

### Tabla 1: Distribución de clientes de una sucursal del banco HSUR, según la demora en ser atendidos

| N.° de clase | Tiempo en minutos[Li -Ls) | N.° de clientes fi |
|---|---|---|
| 1 | 0-2 | 2 |
| 2 | 2-4 | 3 |
| 3 | 4-6 | 5 |
| 4 | 6-8 | 12 |
| 5 | 8-10 | 15 |
| 6 | 10-12 | 10 |
| 7 | 12-14 | 3 |
| | | **50** |

Fuente: elaboración propia.

Luego de este estudio, se decidió aumentar el número de empleados en el sector para mejorar los tiempos de atención. Se volvió a tomar una muestra aleatoria de 50 clientes durante una semana y se obtuvieron los siguientes resultados.

Se muestran los tiempos que demoran los clientes de la sucursal del banco HSUR en ser atendidos después de aumentar el personal de atención al público en el Área Cuentas Corrientes.

### Tabla 2: Distribución de clientes de una sucursal del banco HSUR, según la demora en ser atendidos

| N.° de clase | Tiempo en minutos[Li -Ls) | N.° de clientes fi |
|---|---|---|
| 1 | 0-2 | 1 |
| 2 | 2-4 | 12 |
| 3 | 4-6 | 14 |
| 4 | 6-8 | 15 |
| 5 | 8-10 | 5 |
| 6 | 10-12 | 2 |
| 7 | 12-14 | 1 |
| | | **50** |

Fuente: elaboración propia.

El gerente desea saber —con fundamento en las medidas más representativas para este caso— si han mejorado algunos indicadores. Además, quiere obtener más información sobre la Tabla 2, con base en la regla empírica.

## Valores o medidas de variación o dispersión

En la Figura 1, mostramos dos conjuntos de datos con la misma media, pero uno con mayor dispersión que el otro.

Observa que la media de las dos curvas toma el mismo valor, pero la curva A tiene menor dispersión (o variabilidad) que la curva B.Si medimos solo la media de estas dos distribuciones, estaremos pasando por alto una distinción importante que existe entre las dos curvas.

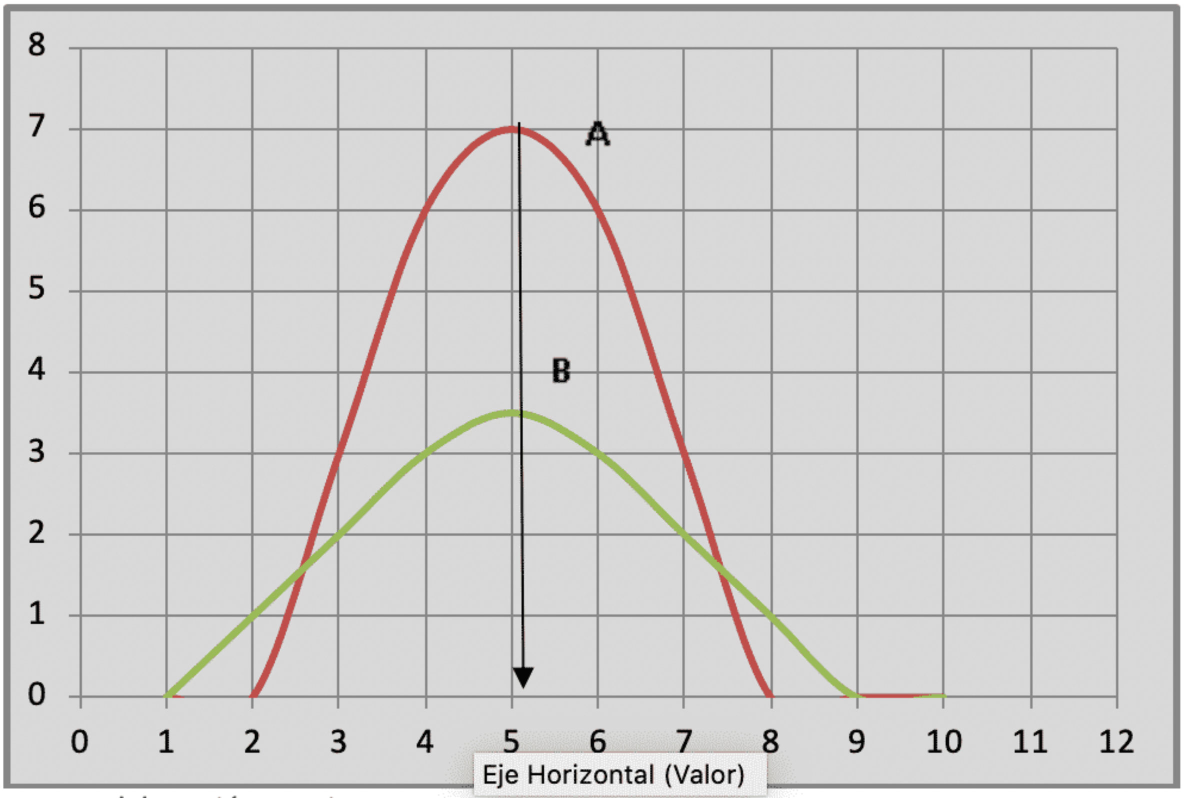
Al igual que sucede con cualquier conjunto de datos, la media, la mediana y la moda solo nos muestran parte de la información que debemos conocer acerca de las características de los datos de una muestra o población.

Existen distribuciones que presentan la misma media, la misma moda y la misma mediana, pero aun así son diferentes.

Podemos tener también distribuciones de igual número de elementos o conformadas por los mismos elementos que, no obstante, son distintas.

Deberíamos, entonces, considerar valores que nos determinen cuán dispersas están dichas distribuciones. Estos valores se denominan *valores de dispersión.*

### Figura 1: Distribuciones con igual media y distinta variabilidad de datos




Fuente: elaboración propia.

En la figura quiere destacarse que, aunque las curvas A y B tengan la misma media, esta no es suficiente para informar las características de las distribuciones. La dispersión en ambas curvas es distinta.

Para mejorar nuestra comprensión del conjunto de datos de cada distribución, debemos medir también su *dispersión* o variabilidad.

## ¿Cuáles son las medidas de dispersión más comunes?

### AMPLITUD O RANGO

**Notación:** Rg.

*Es la diferencia entre el valor más grande y el más pequeño del conjunto de datos.*

$Rg = x_M - x_m$

El rango no toma en cuenta cómo se distribuyen los datos entre el valor más grande y el más pequeño.

**Ejemplo 1**: Los pesajes de 7 bolsas de harina, en kg, tomadas como muestra en un comercio mayorista, son:

4,550   5,102   4,860   5   4,930   5,150   5,200

Rg= $x_M$ - $x_m$ = 5,200 – 4,550 = 0,650

El rango es de 0,650 kg.

## DESVÍO MEDIO

Notación: DM.

Es el promedio de las desviaciones tomadas en valor absoluto.

Una desviación es la diferencia entre un dato y la media aritmética de la distribución.

En el **ejemplo 1:**

- primero calculamos la media de la muestra:

$$\overline{x} = \frac{\sum x_i}{n} = \frac{4{,}550 + 5{,}102 + 4{,}860 + 5 + 4{,}930 + 5{,}150 + 5{,}200}{7} = 4{,}97$$

- Calculamos cada desviación: diferencia de cada dato con la media, tomando los valores absolutos (no nos interesa el signo, no importa la dirección de la desviación, sino cuánto se desvía de la media):

$$|4{,}550 - 4{,}97| = 0{,}42$$
$$|5{,}102 - 4{,}97| = 0{,}132$$
$$|4{,}860 - 4{,}97| = 0{,}11$$
$$|5 - 4{,}97| = 0{,}03$$
$$|4{,}930 - 4{,}97| = 0{,}04$$
$$|5{,}150 - 4{,}97| = 0{,}18$$
$$|5{,}200 - 4{,}97| = 0{,}23$$

- Calculamos el promedio de las desviaciones:

$$DM = \frac{\sum|x_i - \bar{x}|}{n} = \frac{0{,}42 + 0{,}132 + 0{,}11 + 0{,}03 + 0{,}04 + 0{,}18 + 0{,}23}{7} = 0{,}16$$

- Si los datos están en una tabla de frecuencias, se multiplica cada desviación por la frecuencia respectiva, ya que el dato está desviado de la media tantas veces como lo indique la frecuencia. Después se hace el promedio de todas las desviaciones y la fórmula queda:

$$\text{Para una muestra: } DM = \frac{\sum|x_i - \bar{x}| \times f_i}{n}$$

$$\text{Para una población: } DM = \frac{\sum|x_i - \mu| \times f_i}{N}$$

**VARIANZA**

$$\textbf{Notación:}\begin{cases} \sigma^2 \text{para la población} \\ s^2 \text{ para la muestra} \end{cases}$$

*Es el promedio de los cuadrados de las desviaciones.*

- Esta medida es muy parecida en su composición al desvío medio, pero, con el fin de evitar las diferencias con resultados negativos, elevamos las desviaciones al cuadrado en lugar de aplicar valor absoluto. ¡Pero cuidado!, se ha convenido en que, cuando se trata de una muestra, al divisor se le quita una unidad como factor de corrección.
- Por lo tanto, prueba que para el **ejemplo 1:**

$$s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1} = \frac{0{,}2937}{6} = 0{,}0489$$

- Si los datos están en una tabla de frecuencias, se multiplica cada cuadrado de cada desviación por la frecuencia respectiva, ya que el dato está desviado de la media tantas veces como lo indique la frecuencia. Después se hace el promedio de todas las desviaciones y la fórmula queda:

$$\text{Para una muestra: } s^2 = \frac{\sum(x_i - \bar{x})^2 \times f_i}{n-1}$$

$$\text{Para una población: } \sigma = \frac{\sum(x_i - \mu)^2 \times f_i}{N}$$

**DESVIACIÓN ESTÁNDAR**

$$\textbf{Notación:} \begin{cases} \sigma \text{ para la población} \\ s \text{ para la muestra} \end{cases}$$

*Es la raíz cuadrada de la varianza.*

- Por lo tanto, su fórmula para el caso del **ejemplo 1,** que es una muestra, queda:

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

- Prueba que para el caso del **ejemplo 1:**

$$s = \sqrt{0{,}0489} = 0{,}22$$

- La mayor ventaja es que permite trabajar con magnitudes lineales y no al cuadrado, como ocurre con la varianza. Tendría sentido para la varianza hablar de centímetros cuadrados, ¡pero no de hombres o autos al cuadrado!
- Si los datos están en una tabla de frecuencias, se multiplica cada cuadrado de cada desviación por la frecuencia respectiva, ya que el dato está desviado de la media tantas veces como lo indique la frecuencia. Después se hace el promedio de todas las desviaciones y la fórmula queda:

$$\text{Para una muestra: } s = \sqrt{\frac{\sum(x_i - \bar{x})^2 \times f_i}{n-1}}$$

$$\text{Para una población: } \sigma = \sqrt{\frac{\sum(x_i - \mu)^2 \times f_i}{N}}$$

## COEFICIENTE DE VARIACIÓN

**Notación**: CV.

*Es el coeficiente que se obtiene dividiendo la desviación estándar y la media de una misma distribución.*

- Es una medida relativa de dispersión, pues indica cuántas veces está contenida la desviación estándar dentro de la media.
- Es más significativa que la desviación estándar, ya que, si la multiplicamos por 100, podremos ver el porcentaje de variabilidad que tiene la muestra o la población para una variable determinada.
- Su fórmula es:

$$\begin{cases} \text{Para una muestra: } CV = \dfrac{s}{\bar{x}} \\ \text{Para una población: } CV = \dfrac{\sigma}{\mu} \end{cases}$$

- Para el caso del **ejemplo 1**, comprueba:

$$CV = \frac{s}{\bar{x}} = \frac{0{,}22}{4{,}97} = 0{,}04$$

La dispersión es pequeña, representa un 4%.

## Algunas consideraciones de importancia

- El desvío medio se utiliza muy poco: lo superan en exactitud y adaptabilidad a las investigaciones la varianza y la desviación estándar.
- Observa que la desviación estándar es levemente superior al desvío medio para una misma distribución. Probablemente, sea por eso que mide con más perfección la variabilidad.
- Sobre el coeficiente de variación:
    - El coeficiente de variación es muy útil. Interpreta mejor la dispersión de los datos, es una medida de dispersión relativa.
    - La media no puede ser 0, pues no existe la división por cero. Tampoco puede ser negativa, pues nos daría un CV negativo y no es lo que buscamos. En estos casos existen otros coeficientes para poder medir la variabilidad que verás más adelante en otro curso de estadística.
    - A mayor coeficiente de variación, mayor será la variabilidad de los datos; inversamente, a menor CV, menor variabilidad.
    - La desviación estándar sola dice poco, pero, si se la pondera en la media, es más objetiva. De allí la aplicabilidad del coeficiente de variación para comparar distintas distribuciones y ver el grado de variabilidad de cada una.
    - El coeficiente de variación es muy utilizado en los procesos industriales, en los que la variabilidad de un producto en algún proceso de su fabricación puede quedar fuera de los límites especificados de antemano y, así, ocasionar pérdidas a las empresas.

- Algunos autores concluyen con que, si el coeficiente de variación es menor a 0,20, la distribución tiene datos homogéneos; en caso contrario, los datos son más heterogéneos. Pero en esto influye el criterio del investigador y el tipo de datos que se maneje.

> Aunque el coeficiente de variación es un poco más complejo que el cociente del ejemplo, el concepto es el mismo: se usa para comparar la cantidad de variación en grupos de datos que tienen medias diferentes. Advertencia: no compare la dispersión en los conjuntos de datos usando las desviaciones estándar, a menos que las medias sean parecidas.(Levin y Rubin, 2012, p.108).

*Es de importancia que recurras al texto básico, capítulo 3, para ampliar sobre las medidas de dispersión, especialmente cuando los datos están en una tabla de frecuencias. Es muy enriquecedor hacerlo, hay comentarios y ejemplos que te resultarán de mucha utilidad.*

Con lo estudiado hasta aquí, serás capaz de reflexionar y realizar esta actividad. Para resolverla te puede ayudar mirar las fórmulas de los valores de dispersión.

La desviación estándar puede ser negativa.

- ○ Es verdadero, porque hay datos que son menores que la media.
- ○ Es falso, porque no puede haber una dispersión negativa.

SUBMIT

**Caso banco HSUR**

¿Podremos, con lo estudiado hasta aquí, ir resolviendo lo que desea saber el gerente? Por ahora lo importante es ir comprendiendo de qué se tratan las medidas de dispersión.

Se supone que el gerente querrá saber si la media de atención a los clientes en el Área Cuentas Corrientes ha bajado. Además, es importante agregar alguna medida de variabilidad para comparar las dos distribuciones, antes y después de la inclusión de personal en el sector.

Pero la realidad es que los datos que se presentan en los cuadros son agrupados en intervalos. ¿No te parece que tendremos que investigar qué pasa con los datos agrupados? No es mucha la diferencia, solo cambian algunos cálculos, pero la esencia es la misma.

Veamos ahora cómo se calculan las medidas de dispersión para datos agrupados y resolvamos el caso del banco HSUR.

# Valores de dispersión para datos agrupados

## ¿Cómo se calculan las medidas de dispersión para datos agrupados?

Las medidas de dispersión para datos agrupados son las mismas que acabamos de estudiar para una serie de datos pequeña o para una tabla de frecuencias.

Observa: ¿Qué es lo único que cambia en el desvío medio, en la varianza y en la desviación estándar?

DESVÍO MEDIO

$$\text{Para una muestra: } DM = \frac{\sum |x_{mi} - \bar{x}| \times f_i}{n}$$

$$\text{Para una población: } DM = \frac{\sum |x_{mi} - \mu| \times f_i}{N}$$

1 of 3

VARIANZA

$$\text{Para una muestra: } s^2 = \frac{\sum (x_{mi} - \bar{x})^2 \times f_i}{n-1}$$

$$\text{Para una población: } \sigma = \frac{\sum (x_{mi} - \mu)^2 \times f_i}{N}$$

2 of 3

DESVIACIÓN ESTÁNDAR

$$\text{Para una muestra: } s = \sqrt{\frac{\sum (x_{mi} - \bar{x})^2 \times f_i}{n - 1}}$$

$$\text{Para una población: } \sigma = \sqrt{\frac{\sum (x_{mi} - \mu)^2 \times f_i}{N}}$$

3 of 3

Seguramente, ya lo descubriste: se sustituye en la fórmula el valor de la variable $x_i$ por la marca de clase $x_{mi}$. No olvides que es el representante de la clase. Entonces, para calcular cada desviación de un dato con la media, en un intervalo, interviene ese representante, que es la marca de clase.

Por supuesto que, si ese intervalo tiene, por ejemplo, frecuencia 8, habrá 8 desviaciones que tienen que promediarse con el resto de los datos de la tabla. Es por eso que cada desviación se multiplica por la frecuencia.

El resto de las medidas casi no cambia, solo el rango se considera con una variante:

AMPLITUD O RANGO

*Ra= Ls*(de la última clase) − *Li* (de la primera clase)

AMPLITUD O RANGO

*Rg*= *Ls*(de la última clase) – *Li* (de la primera clase)

1 of 2

COEFICIENTE DE VARIACIÓN

$$\begin{cases} \text{Para una muestra:} \quad CV = \dfrac{s}{\bar{x}} \\ \text{Para una población:} \quad CV = \dfrac{\sigma}{\mu} \end{cases}$$

2 of 2

### Caso banco HSUR

Ahora sí, con estas nuevas fórmulas, podemos avanzar sobre el caso del banco HSUR.

El gerente desea saber —con fundamento en las medidas más representativas para este caso— si han mejorado algunos indicadores. Para poder dar respuesta al gerente, lo más indicado es calcular la media en ambas distribuciones y la desviación estándar para poder, así, calcular el coeficiente de variación. Estas medidas nos ayudarán a dar una respuesta con sustento estadístico al gerente.

Comencemos entonces a resolver el caso, calculando la media en ambos muestreos.

Se muestran los tiempos que demoran los clientes de la sucursal del banco HSUR en ser atendidos. Estos datos fueron tomados antes de realizar el experimento de aumentar el personal. Se agregan las columnas auxiliares para calcular la media.

### Tabla 3: Distribución de clientes de una sucursal del banco HSUR, según la demora en ser atendidos, antes del experimento

| N.° de clase | Tiempo en minutos [Li - Ls) | $X_{mi}$ | N.° de clientes *fi* | Auxiliar $X_{mi}$ x fi |
|---|---|---|---|---|
| 1 | 0-2 | 1 | 2 | 2 |
| 2 | 2-4 | 3 | 3 | 9 |
| 3 | 4-6 | 5 | 5 | 25 |
| 4 | 6-8 | 7 | 12 | 84 |
| 5 | 8-10 | 9 | 15 | 135 |
| 6 | 10-12 | 11 | 10 | 110 |
| 7 | 12-14 | 13 | 3 | 39 |
| | | | **50** | **404** |

Fuente: elaboración propia.

*$\bar{x}_1$ media correspondiente a la Tabla 1, experimento anterior.*

$$\bar{x}_1 = \frac{\sum x_{mi} \times f_i}{50} = \frac{404}{50} = 8{,}08$$

El promedio de demoras en la Tabla 1 o en la Tabla 3 es de 8,08 minutos. Esta demora es anterior a la investigación realizada por el gerente.

Calculamos ahora la media para la Tabla 2.

Se muestran los tiempos que demoran los clientes de la sucursal del banco HSUR en ser atendidos después de aumentar el personal de atención al público en el Área Cuentas Corrientes.

### Tabla 4: Distribución de clientes de una sucursal del banco HSUR, según la demora en ser atendidos. Esta tabla se realizó después del experimento

| N.° de clase | Tiempo en minutos [Li - Ls) | $X_{mi}$ | N.° de clientes *fi* | Auxiliar $X_{mi}$ x fi |
|---|---|---|---|---|
| 1 | 0-2 | 1 | 1 | 1 |
| 2 | 2-4 | 3 | 12 | 36 |
| 3 | 4-6 | 5 | 14 | 70 |
| 4 | 6-8 | 7 | 15 | 105 |
| 5 | 8-10 | 9 | 5 | 45 |
| 6 | 10-12 | 11 | 2 | 22 |
| 7 | 12-14 | 13 | 1 | 13 |
| | | | **50** | **292** |

Fuente: elaboración propia.

*$\bar{x}_2$ media correspondiente a la Tabla 2, experimento anterior.*

$$\bar{x}_2 = \frac{\sum x_{mi} \times f_i}{50} = \frac{292}{50} = 5{,}84$$

El promedio de demoras en la Tabla 2 o de la Tabla 4 es de 8,08 minutos. Esta demora es posterior a la investigación realizada por el gerente.

*Podemos adelantar que, con el refuerzo de empleados, se logró disminuir la media de demora de atención a los clientes.*

Con el fin de comparar las distribuciones mediante el coeficiente de variación, calculamos la desviación estándar en ambas distribuciones.

Observa las columnas de auxiliares de la tabla 5, estos pasos en la tabla son para ir armando la fórmula de la desviación estándar. Esto puede realizarse mediante calculadora científica o un *software* estadístico.

El objetivo de explicarla paso a paso es para que comprendas que la desviación estándar es, en definitiva, un promedio de los cuadrados de las desviaciones.

Calculamos la desviación estándar y el coeficiente de variación de la **Tabla 1.**

### Tabla 5: Distribución de clientes de una sucursal del banco HSUR, según la demora en ser atendidos, antes del experimento

Se muestran los tiempos que demoran los clientes de la sucursal del banco HSUR en ser atendidos. Estos datos fueron tomados antes de realizar el experimento de aumentar el personal. Se agregan las columnas auxiliares para calcular la desviación estándar.

| N.° de clase | Tiempo en minutos [Li -Ls) | $X_{mi}$ | N.° de clientes $fi$ | Auxiliar $(x_{mi}-\bar{x})^2$ | Auxiliar $(x_{mi}-\bar{x})^2 \times f_i$ |
|---|---|---|---|---|---|
| 1 | 0-2 | 1 | 2 | 50,1264 | 100,2528 |
| 2 | 2-4 | 3 | 3 | 25,8064 | 77,4192 |
| 3 | 4-6 | 5 | 5 | 9,4864 | 47,432 |
| 4 | 6-8 | 7 | 12 | 1,1664 | 13,9968 |
| 5 | 8-10 | 9 | 15 | 0,8464 | 12,696 |
| 6 | 10-12 | 11 | 10 | 8,5264 | 85,264 |
| 7 | 12-14 | 13 | 3 | 24,2064 | 72,6192 |
| | | | **50** | | **409,68** |

Fuente: elaboración propia:

Calculamos primero la varianza:

$$s^2 = \frac{\sum(x_{mi}-\bar{x})^2 \times f_i}{n-1} = \frac{409{,}68}{49} = 8{,}3608$$

Calculamos la desviación estándar:

$$s = \sqrt{\frac{\sum(x_{mi} - \bar{x})^2 \times f_i}{n-1}} = 2{,}89$$

Calculamos el coeficiente de variación:

$$CV = \frac{s}{\bar{x}} = 0{,}3579$$

o la distribución tiene una variabilidad del 35,79 %.

Calculamos la desviación estándar y el coeficiente de variación de la Tabla 2.

### Tabla 6: Distribución de clientes de una sucursal del banco HSUR, según la demora en ser atendidos. Esta tabla se realizó después del experimento

Se muestran los tiempos que demoran los clientes de la sucursal del banco HSUR en ser atendidos después de aumentar el personal de atención al público en el Área Cuentas Corrientes. Se agregan columnas auxiliares para calcular la desviación estándar.

| N.° de clase | Tiempo en minutos [Li -Ls) | $X_{mi}$ | N.° de clientes *fi* | Auxiliar $(x_{mi}-\bar{x})^2$ | Auxiliar $(x_{mi}-\bar{x})^2 \times f_i$ |
|---|---|---|---|---|---|
| 1 | 0-2 | 1 | 1 | 23,4256 | 23,4256 |
| 2 | 2-4 | 3 | 12 | 8,0656 | 96,7872 |
| 3 | 4-6 | 5 | 14 | 0,7056 | 9,8784 |
| 4 | 6-8 | 7 | 15 | 1,3456 | 20,184 |
| 5 | 8-10 | 9 | 5 | 9,9856 | 49,928 |
| 6 | 10-12 | 11 | 2 | 26,6256 | 53,2512 |
| 7 | 12-14 | 13 | 1 | 51,2656 | 51,2656 |
| | | | **50** | | **304,72** |

Fuente: elaboración propia.

Calculamos primero la varianza:

$$s^2 = \frac{\sum(x_{mi} - \bar{x})^2 \times f_i}{n-1} = \frac{304,72}{49} = 6,2188$$

Calculamos la desviación estándar:

$$s = \sqrt{\frac{\sum(x_{mi} - \bar{x})^2 \times f_i}{n-1}} = 2,49$$

Calculamos el coeficiente de variación:

$$CV_2 = \frac{s}{\bar{x}} = 0,4270$$

o la distribución tiene una variabilidad del 42,70%.

### Conclusión

Si bien la media de demoras de atención al cliente ha disminuido, la dispersión de la muestra tomada después de la investigación arroja una mayor variabilidad que la primera, aproximadamente un 7% más dispersa. Lo que sucede es que en ambas distribuciones hay una variabilidad bastante alta.

El hecho de que, en el último muestreo, la variabilidad haya aumentado significa que hay algunos datos extremos con muy poca frecuencia. Si bien la media bajó, todavía quedan algunos pocos clientes que tardan en ser atendidos.

En dos muestras de tornillos tomadas en forma aleatoria en una fábrica de autopartes, una de la máquina A y otra de la máquina B, se midió la longitud los tornillos producidos. Los resultados después de realizar las mediciones fueron: en la máquina A, la media de la longitud de los tornillos fue de 15mm, y la desviación estándar, de 0,15 mm. Si la máquina B produce tornillos con una media de 14,20 mm y una desviación estándar de 0,22 mm, ¿en cuál de las dos máquinas se producen tornillos con mayor variabilidad de longitud?

- En la máquina B.
- En la máquina A.
- En ambas máquinas.
- En ninguna de las máquinas.

SUBMIT

## Consejos para recordar las fórmulas de la estadística descriptiva

Te damos algunos consejos para no tener que recordar todas las fórmulas estudiadas en las lecturas del Módulo 1. Solo te mostraremos las más importantes, pero sobre el resto podrás —por analogía— crear tus propios *tips*.

Conviene que aprendas la fórmula para datos agrupados y luego la cambies en función de si se trata de una serie simple, de una tabla de frecuencias sin agrupar o de una tabla de frecuencias agrupada.

### Media aritmética

Recuerda esta fórmula:

#### Figura 2: Consejos para recordar la fórmula de lamedia

$$\mu = \frac{\sum xm_i \times f_i}{N}$$

$$\mu = \frac{\sum xm_i \times f_i}{N}$$

**μ**

media poblacional. Si se trata de una muestra, es $\bar{x}$.

$$\mu = \frac{\sum xm_i \times f_i}{N}$$

**fi**

frecuencia absoluta. Si se trata de datos distintos, $f_i$ vale 1 y no hace falta que figure la fórmula.

$$\mu = \frac{\sum xm_i \times f_i}{N}$$

N

cantidad de elementos de una población. Si se trata de una muestra es n.

$$\mu = \frac{\sum xm_i \times f_i}{N}$$

**xmi**

marca de clase para datos agrupados. Si se trata de datos sin agrupar, es $x_i$.

### Desviación estándar

Recuerda esta fórmula:

### Figura 3:Consejos para recordar la fórmula de la desviación estándar

$$\sigma = \sqrt{\frac{\sum (xm_i - \mu)^2 \times f_i}{N}}$$

$$\sigma = \sqrt{\frac{\sum (xm_i - \mu)^2 \times f_i}{N}}$$

**σ**

desviación estándar poblacional. Si se trata de una muestra, es ***s***.

$$\sigma = \sqrt{\frac{\sum (xm_i - \mu)^2 \times f_i}{N}}$$

N

cantidad de elementos de una población. Si se trata de una muestra, es n-1.

$$\sigma = \sqrt{\frac{\sum (xm_i - \mu)^2 \times f_i}{N}}$$

**fi**

frecuencia absoluta. Si se trata de datos distintos, $f_i$ vale 1 y no hace que figure en la fórmula.

$$\sigma = \sqrt{\frac{\sum (xm_i - \mu)^2 \times f_i}{N}}$$

**μ**

media aritmética para una población. Si se trata de una muestra es $\bar{x}$.

$$\sigma = \sqrt{\frac{\sum (xm_i - \mu)^2 \times f_i}{N}}$$

**xmi**

marca de clase para datos agrupados. Si se trata de datos sin agrupar, es $x_i$.

# Regla empírica

## ¿A qué se llama *distribución normal*?

La *regla empírica* es una aplicación práctica de la *distribución normal*, que relaciona a la media y a la desviación estándar de una población.

¿Recuerdas en la Lectura 3 cuando hablamos de curvas en lugar de polígonos para representar la simetría y los tipos de asimetrías en una distribución? Estas simetrías y asimetrías se daban al comparar la media, la mediana y la moda.

Cuando la curva es simétrica (media = mediana = moda) y, además, tiene forma acampanada, decimos que es una *curva normal* o acampanada (¡cuidado!, una distribución puede ser simétrica y no tener forma de campana).

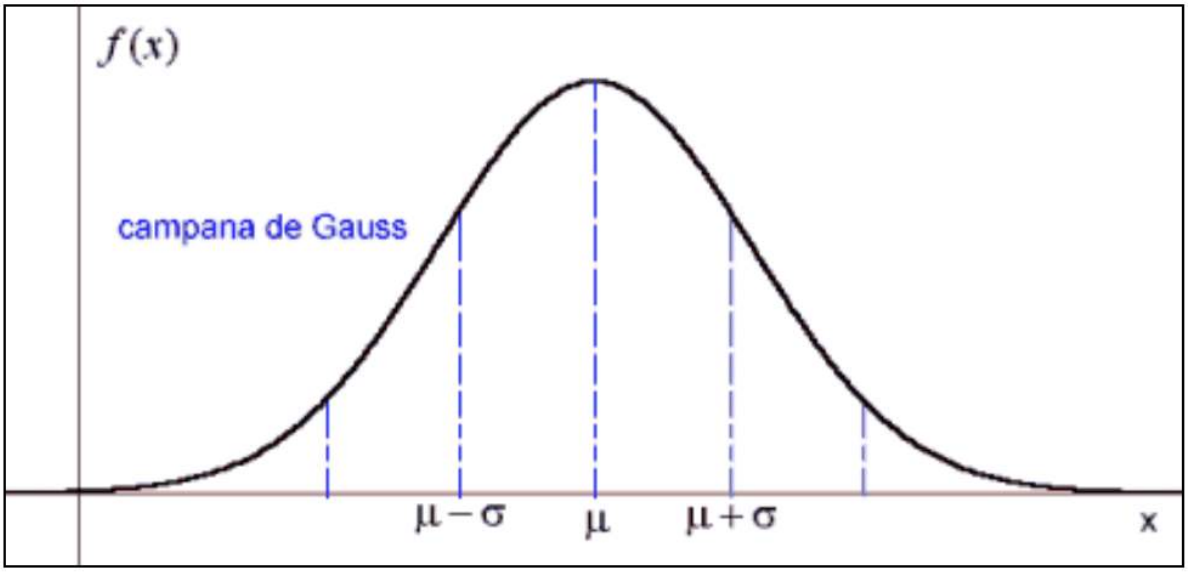
También se la llama *curva de Gauss* y tiene múltiples aplicaciones.

Los gráficos A y B de la Figura 1 son curvas normales.

A continuación, mostramos un gráfico de curva normal.

### Figura 4: Distribución normal o campana de Gauss

En la gráfica se observa la forma acampanada y simétrica que tiene la distribución normal. La media, la mediana y la moda coinciden. El eje vertical mide la frecuencia relativa de cada punto del eje horizontal. El área bajo la curva representa a todos los valores de la población.



Fuente: Plaza Martínez, s. f., http://bit.ly/2YKUPbr

En realidad, es una función de probabilidad, como veremos en el Módulo 3. Es importante, por ahora, que sepas las características generales, pues a la distribución normal la utilizaremos en toda la materia y es básica para comprender la estadística inferencial.

## ¿Qué dice la regla empírica?

> Cuando se cree que los datos tienen aproximadamente esta distribución, se puede emplear la **regla empírica** para determinar el porcentaje de los valores de los datos que deben encontrarse dentro de un determinado número de desviaciones estándar de la media. (Anderson, Sweeney y Williams, 2008, p. 101).

El autor aquí se refiere a la distribución normal, y la regla —basada en la experiencia— dice lo siguiente:

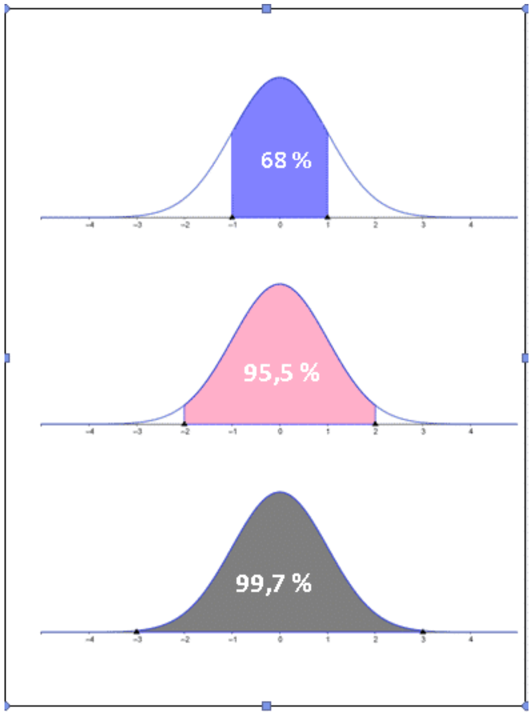
*El área bajo la curva normal limitada por*

- *"μ ± σ c% de los valores de la población;*
- *"μ ± 2σ c% de los valores de la población;*
- *"μ ± 3σ c% de los valores de la población.*

Esta regla también se aplica a las observaciones de una muestra.

### Figura 5: Regla empírica

Se muestran los porcentajes que ocupan las áreas bajo la curva para la media ±1, ±2 y ±3 desviaciones estándar.




Fuente: elaboración propia.

## Uso de la regla empírica

Veamos el siguiente ejemplo:

Los envases con detergente líquido se llenan de forma automática en una línea de producción. Los pesos de llenado suelen tener una distribución en forma de campana. Si el peso medio de llenado es de 500 ml y la desviación estándar es de 0,03 ml, la regla empírica es aplicada para sacar las siguientes conclusiones:

- El 68 % de la línea de producción tiene un pesaje aproximado que está entre 499,97 ml y 500,03 ml.
- El 95,5 % de la línea de producción tiene un pesaje aproximado que está entre 499,94 ml y 500,06 ml.
- El 99,7 % de la línea de producción tiene un pesaje aproximado que está entre 499,91 ml y 500,09 ml.

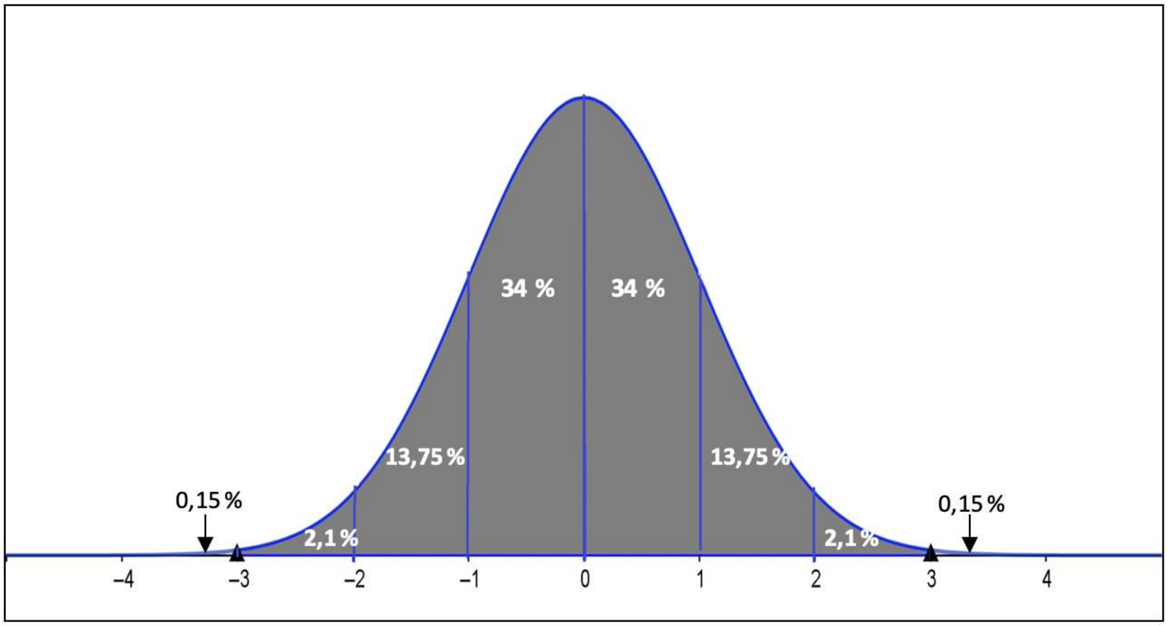
Si, por ejemplo, la empresa tiene rechazos internos y externos cuando el pesaje del líquido de llenado es menor a 499 ml, entonces con esta regla empírica la empresa puede concluir con que no tendrá ningún rechazo por ese motivo.

Para calcular cualquier magnitud que se quiera, en ejemplos concretos, es útil calcular las áreas de cada sector para poder responder a las posibles preguntas que se puedan realizar en algún problema. Estos porcentajes son siempre iguales, cualquiera sea la situación que haya que resolver, siempre que la distribución sea normal.

Verifica los siguientes porcentajes, dada la simetría de la distribución normal.

Se muestran las distintas áreas bajo la curva de la normal expresada en porcentajes, de acuerdo con la regla empírica.

**Figura 6: Porcentajes por sectores de la regla empírica**




Fuente: elaboración propia

## La regla empírica para el caso del banco HSUR

Para poder aplicar la regla empírica para el caso del banco HSUR, primero tendríamos que analizar si es aplicable. Esto significa que en la Tabla 2 (la realizada después del experimento) tendría que darse una distribución aproximadamente normal o acampanada.

Pareciera a la vista que no lo es, pero comparemos la media, la mediana y la moda. Si son aproximadamente iguales, podría aplicarse.

Ya habíamos calculado la media:

$$\bar{x}_2 = \frac{\sum x_{mi} \times f_i}{50} = \frac{292}{50} = 5{,}84$$

Calculemos ahora la mediana:

1) Determinamos la primera clase cuya frecuencia acumulada es mayor a n/2.

**Descripción de la tabla:** En esta tabla se muestra el caso del banco HSUR después de haber realizado la investigación, con el objeto de calcular las medias de posición.

### Tabla 7: Tabla de frecuencias para el caso del banco HSUR

| N.° de clase | Tiempo en minutos [*Li* -*Ls*) | *Marca de clase xmi* | N° de clientes *fi* | *Frecuencia acumulada Fi* |
|---|---|---|---|---|
| 1 | 0 - 2 | 1 | 1 | 1 |
| 2 | 2 - 4 | 3 | 12 | 13 |
| **3** | **4 - 6** | **5** | **14** | **27** |
| 4 | 6 - 8 | 7 | 15 | 42 |
| 5 | 8 - 10 | 9 | 5 | 47 |
| 6 | 10 - 12 | 11 | 2 | 49 |
| 7 | 12 - 14 | 13 | 1 | 50 |
| | | | **50** | |

Clase que contiene a la mediana

Clase que contiene a la moda

Fuente: elaboración propia

La clase mediana es el número 3.

2) Aplicar la fórmula:

$$\bar{x} = L_i + \frac{\Delta x\left(\frac{n}{2} - F_{(m-1)}\right)}{f_m} = 4 + \frac{2(25-13)}{27} = 4 + \frac{24}{27} = 4{,}89$$

Cálculo de la moda:

1. Determinamos la clase modal, que es la clase 4, por tener la mayor frecuencia absoluta.
2. Aplicamos la fórmula:

$$\hat{x} = Li + \Delta x \times \frac{d_1}{(d_1 + d_2)} = 6 + 2\,\frac{1}{1 + 10} = 6 + 2 \times \frac{1}{11} = 6{,}18$$

### Conclusión

Claramente, en el caso de la sucursal 5 del banco HSUR, no puede aplicarse la regla empírica. La distribución de clientes que esperan ser atendidos después de haber sido reforzado el sector no sigue una distribución normal. La moda es mayor que la media, y esta, a su vez, es mayor que la mediana, por lo que tiene una asimetría negativa. Tendríamos que calcularlo mediante otra regla, que no es objeto de los contenidos que pensamos para este curso.

Entonces, no podemos darle respuestas al gerente, quien solicitó información basada en la regla empírica, pero sí le hemos dado respuesta sobre el estudio de variabilidad que había solicitado, tal cual lo mostramos en su momento.

# Referencias

**Anderson, D. R., Sweeney, D. J. y Williams, T. A.** (2008). *Estadística Descriptiva: Medidas numéricas*. En Autores, *Estadística para Administración y Economía* (pp. 89-109). D.F., MX: Cengage Learning.

**Berenson, M., Levine, D. y Krehbiel, T.** (2006). Medidas numéricas descriptivas. En Autores, *Estadística descriptiva* (pp. 72-121). Naucalpan de Juárez, MX. Pearson.

**Levin, R. y Rubin, D.** (2004). *Estadística para Administración y Economía* (7.a ed.). Naucalpan de Juárez, MX: Pearson.

**Plaza Martínez, J.** (s. f.). Representación gráfica de esta función de densidad [Imagen]. Recuperado de https://www.matematicasonline.es/BachilleratoCCSS/segundo/archivos/distribucion_normal/DISTRIBUCION%20NORMAL.htm

# Revisión del módulo

## Hasta acá aprendimos

### La estadística

Presentamos la terminología y conceptos básicos de la Estadística que se manejarán a lo largo de toda la materia. Además, se han relacionado con ejemplos y casos de la vida profesional y cotidiana.

### Presentación de datos

En esta lectura has aprendido a recopilar, organizar y presentar la información obtenida de la realidad. El énfasis se puso en la presentación y construcción de gráficos como las tablas de distribución de frecuencias; definición de tipos de frecuencias; y, construcción de gráficos según el tipo de datos que manipule. Siempre en aplicaciones concretas.

### Medidas de posición

En esta lectura has aprendido a analizar los datos presentados, mediante el cálculo de las medidas de posición: media, mediana y moda. Interpretándolas posteriormente y relacionándolas entre sí. Estos cálculos e interpretación posterior se aplicaron siempre a problemas concretos ya que se calculan distinto según el tipo de datos que posea el problema.

### Valores de dispersión

En esta lectura has aprendido a analizar los datos presentados, complementando las medidas de posición con las de dispersión, percibiendo la importancia de la variabilidad de los procesos y los costos que puedan ocasionar a cualquier emprendimiento el no tener control de estos valores. Has aprendido a calcular en especial la desviación estándar, la varianza y el coeficiente de variación en problemas concretos.