# FUNDAMENTOS DE MACHINE LEARNING

## Estadística Descriptiva II

Medidas de Dispersión

DuocUC®

ESCUELA DE
INFORMÁTICA Y
TELECOMUNICACIONES

# ¿Por qué la Estadística Descriptiva es la Base del Machine Learning?

Imagina que quieres entrenar un modelo de Machine Learning para predecir el precio de casas.

Primero debemos …

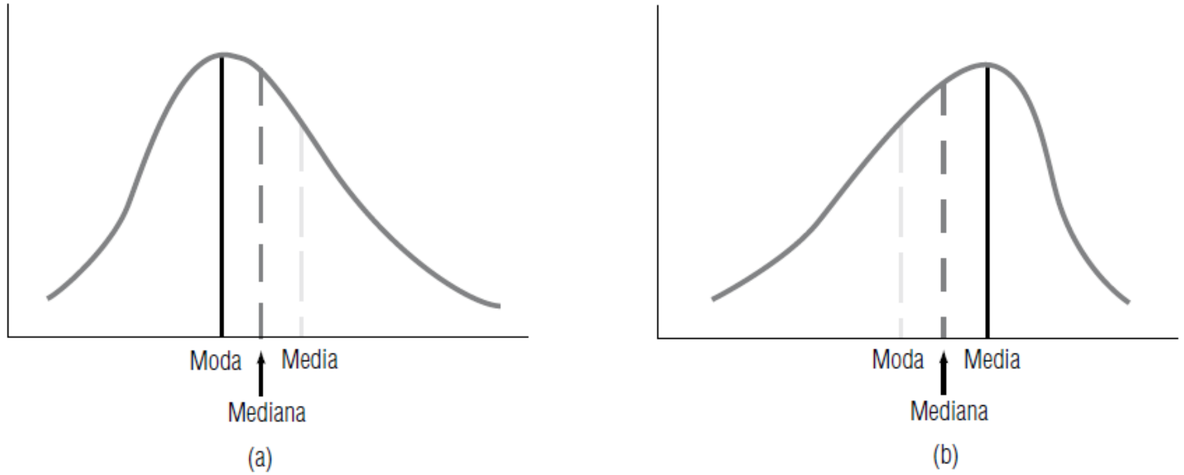
**entenderlos, limpiarlos y organizarlos**.

# ESTADÍSTICOS BÁSICOS

## TENDENCIA CENTRAL

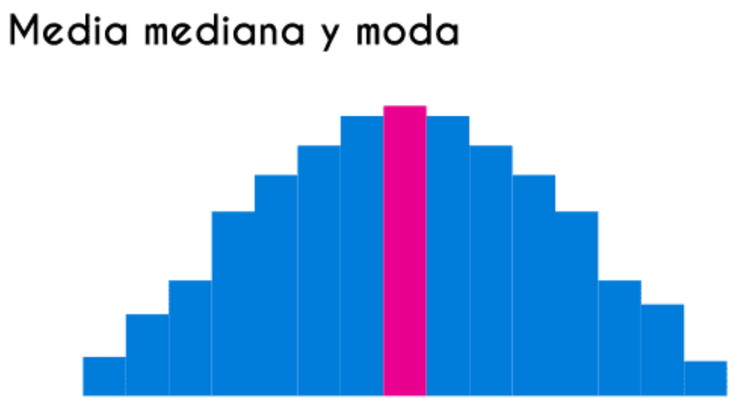
**COMPARACIÓN:** En distribuciones **simétricas** de datos, la moda, la mediana y la media tienen el mismo valor. Cuando la población de datos está sesgada negativa o positivamente, la mediana suele ser la mejor medida de posición, debido a que siempre está entre la moda y la media. Pero, elegir la medida adecuada, dependerá de cada caso que se esté analizando.

Media mediana y moda




Distribuciones con sesgo (a) positivo y (b) negativo que muestran las posiciones de la media, la mediana y la moda.

# ESTADÍSTICOS BÁSICOS | TENDENCIA CENTRAL




**La moda sería la estatura que más se repite en el grupo.**

**La mediana sería la estatura de la persona que está justo en el medio de la fila si los ordenamos de menor a mayor.**

**La media sería el promedio de todas las estaturas sumadas y divididas entre el número de amigos.**

**Si todos tienen estaturas similares, estos tres valores estarán cerca.**

# ESTADÍSTICOS BÁSICOS | TENDENCIA CENTRAL

Media mediana y moda

**La moda sería la estatura que más se repite en el grupo.**

**La mediana sería la estatura de la persona que está justo en el medio de la fila si los ordenamos de menor a mayor.**

**La media sería el promedio de todas las estaturas sumadas y divididas entre el número de amigos.**

**Si todos tienen estaturas similares, estos tres valores estarán cerca.**

**Pero si hay un amigo extremadamente alto o muy bajo,**

# ESTADÍSTICOS BÁSICOS | TENDENCIA CENTRAL


**La moda sería la estatura que más se repite en el grupo.**

**La mediana sería la estatura de la persona que está justo en el medio de la fila si los ordenamos de menor a mayor.**

**La media sería el promedio de todas las estaturas sumadas y divididas entre el número de amigos.**

**Si todos tienen estaturas similares, estos tres valores estarán cerca.**

**Pero si hay un amigo extremadamente alto o muy bajo, eso puede afectar el promedio (media), desplazándolo hacia un lado, mientras que la mediana y la moda podrían quedarse en el mismo lugar.**

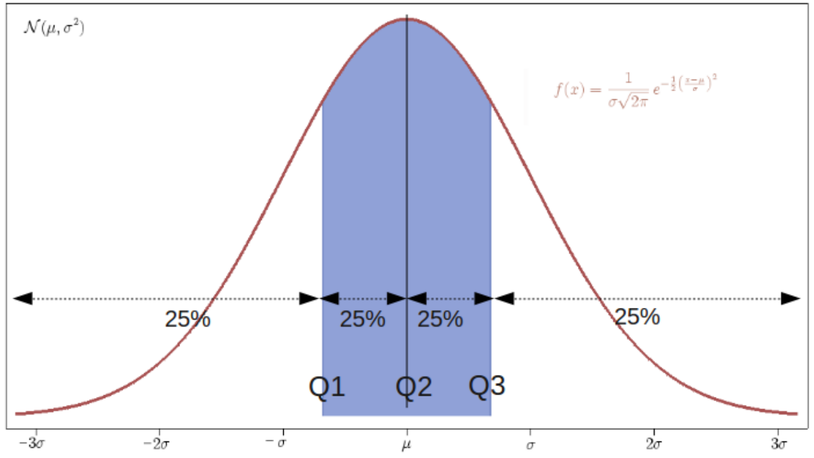
# ESTADÍSTICOS BÁSICOS

## POSICIÓN NO CENTRAL

**CUANTILES:** Los cuantiles son aquellos valores de la variable que, ordenados de menor a mayor, dividen la distribución en partes. De tal manera, que cada una de ellas contiene el mismo número de frecuencias.

Los cuantiles más conocidos son :

**CUARTILES**: son valores de la variable que dividen a la distribución en cuatro partes, cada una de las cuales engloba el 25% de las mismas.

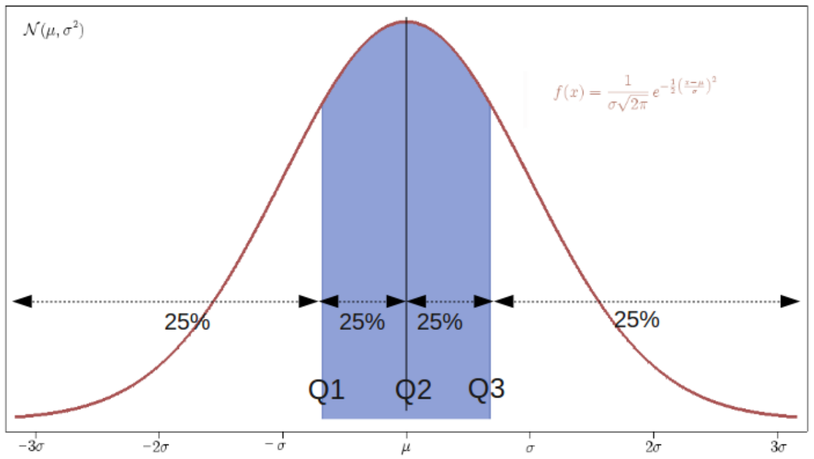



**DECILES**: son los valores de la variable que dividen a la distribución en diez partes iguales, cada una de ellas engloba el 10% de los datos.

**CENTILES O PERCENTILES**: son los valores que dividen a la distribución en cien partes iguales, cada una de las cuales engloba el 1% de las observaciones.

# ESTADÍSTICOS BÁSICOS

# POSICIÓN NO CENTRAL




Imagina que 100 personas participan en una maratón. Queremos saber en qué posición están en comparación con los demás.

Si dividimos a los corredores en **4 grupos de 25 personas**, estamos usando **cuartiles**.

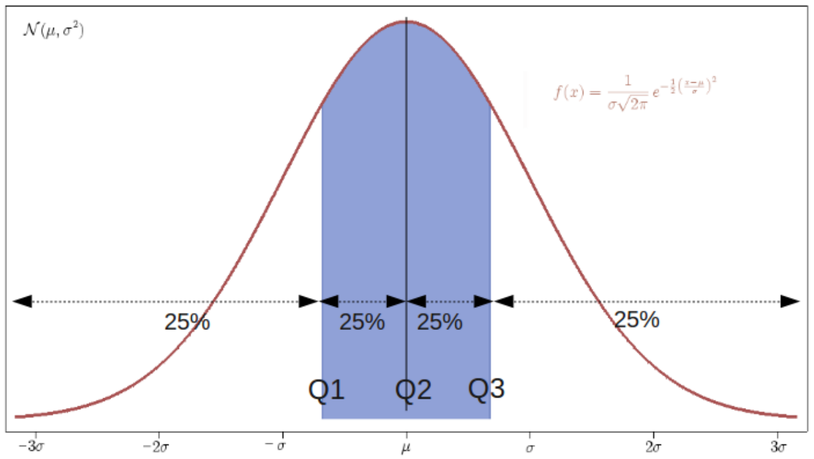
El primer 25% de corredores más rápidos están en **Q1**.

El 50% (la mitad) está en Q2 (que también es la mediana).

El 75% de los corredores ya habrán llegado a Q3.

Y el último 25% está en la última parte de la carrera.

# ESTADÍSTICOS BÁSICOS | POSICIÓN NO CENTRAL




Imagina que 100 personas participan en una maratón. Queremos saber en qué posición están en comparación con los demás.

Si dividimos a los corredores en **4 grupos de 25 personas**, estamos usando **cuartiles**.

Si dividimos en **10 grupos de 10 corredores**, estamos usando **deciles**.

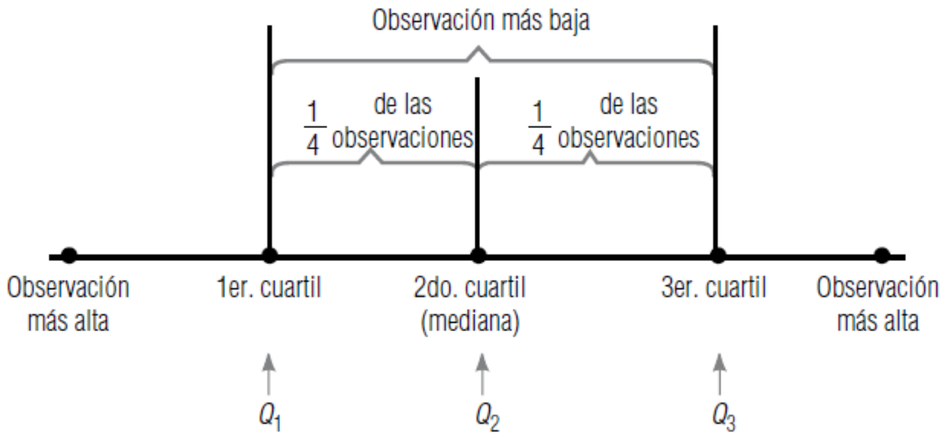
Si dividimos en **100 grupos de 1 corredor**, estamos usando **percentiles**.

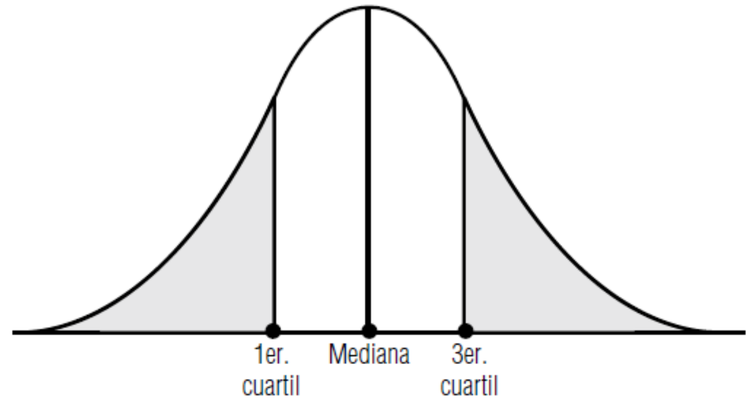
# ESTADÍSTICOS BÁSICOS | POSICIÓN NO CENTRAL

## RANGO INTERCUARTIL:

El rango intercuartílico mide aproximadamente qué tan lejos de la mediana debemos ir en cualquiera de las dos direcciones antes de recorrer una mitad de los valores del conjunto de datos. Para calcular este rango, se calcula la diferencia entre los valores del primer y tercer cuartil

$$\text{Rango intercuartil} = Q_3 - Q_1$$



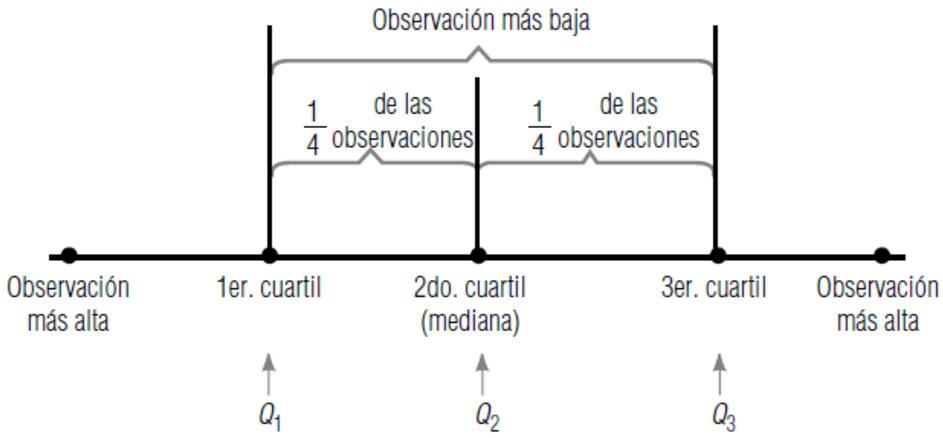



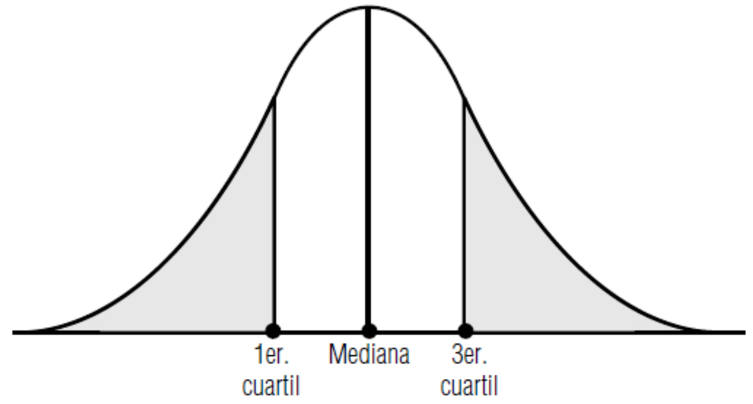
# ESTADÍSTICOS BÁSICOS | POSICIÓN NO CENTRAL

## RANGO INTERCUARTIL:

El rango intercuartílico mide aproximadamente qué tan lejos de la mediana debemos ir en cualquiera de las dos direcciones antes de recorrer una mitad de los valores del conjunto de datos. Para calcular este rango, se calcula la diferencia entre los valores del primer y tercer cuartil

$$\text{Rango intercuartil} = Q_3 - Q_1$$




Nos dice qué tan dispersos están los datos en la parte central de una distribución

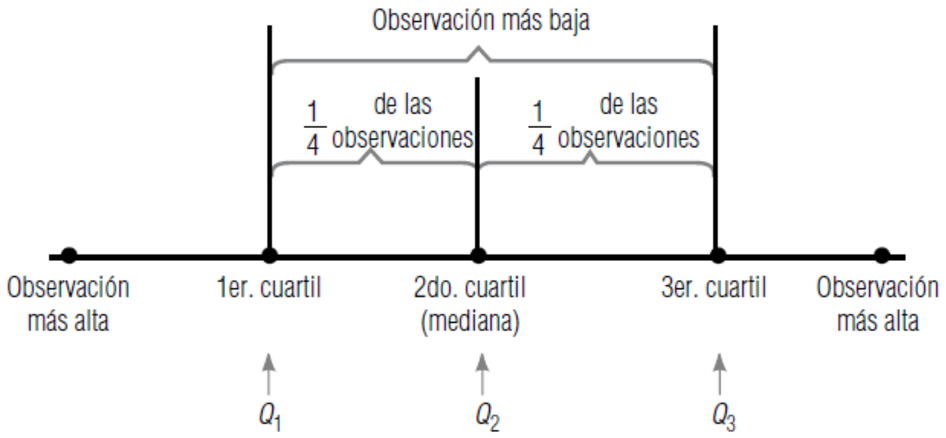
# ESTADÍSTICOS BÁSICOS | POSICIÓN NO CENTRAL

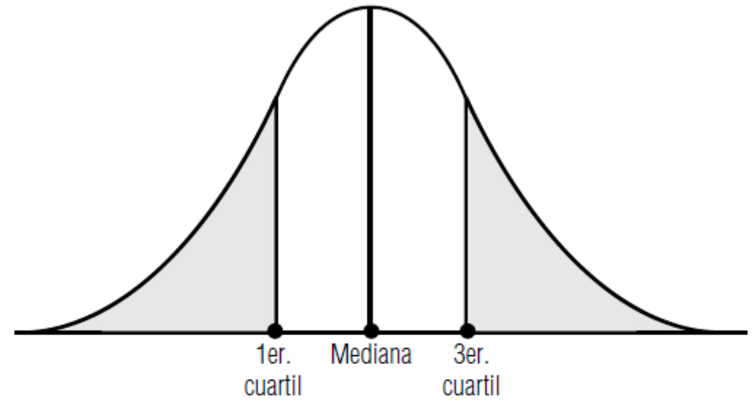
## RANGO INTERCUARTIL:

El rango intercuartílico mide aproximadamente qué tan lejos de la mediana debemos ir en cualquiera de las dos direcciones antes de recorrer una mitad de los valores del conjunto de datos. Para calcular este rango, se calcula la diferencia entre los valores del primer y tercer cuartil

$$\text{Rango intercuartil} = Q_3 - Q_1$$

**representa el 50% central de los datos**




Nos dice qué tan dispersos están los datos en la parte central de una distribución

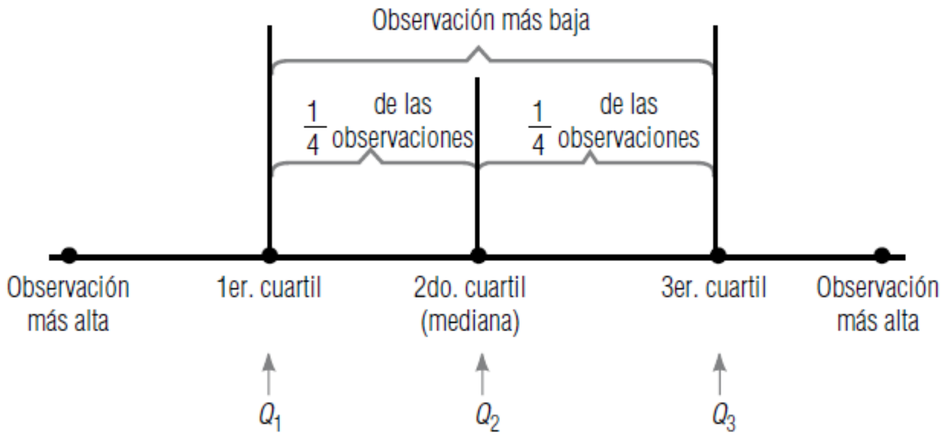
# ESTADÍSTICOS BÁSICOS | POSICIÓN NO CENTRAL

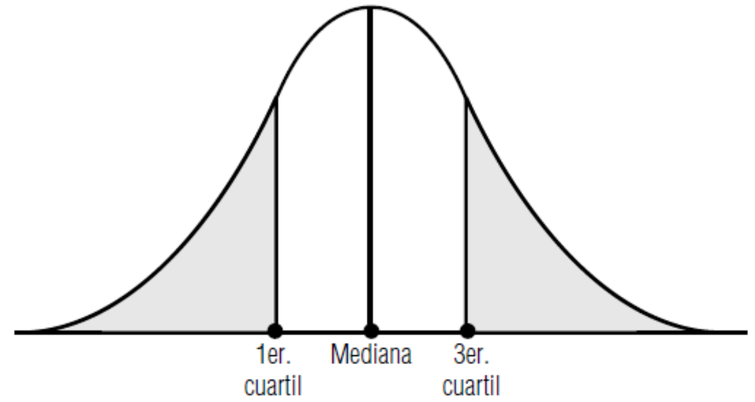
## RANGO INTERCUARTIL:

El rango intercuartílico mide aproximadamente qué tan lejos de la mediana debemos ir en cualquiera de las dos direcciones antes de recorrer una mitad de los valores del conjunto de datos. Para calcular este rango, se calcula la diferencia entre los valores del primer y tercer cuartil

$$\text{Rango intercuartil} = Q_3 - Q_1$$

**representa el 50% central de los datos**




**elimina los valores más extremos y se enfoca en la parte más representativa del conjunto de datos.**

Nos dice qué tan dispersos están los datos en la parte central de una distribución

# ESTADÍSTICOS BÁSICOS | POSICIÓN NO CENTRAL

## RANGO INTERCUARTIL:

**Imagina que estás en un vagón del metro en hora punta. Hay mucha gente y queremos medir qué tan apretado está el espacio entre las personas.**

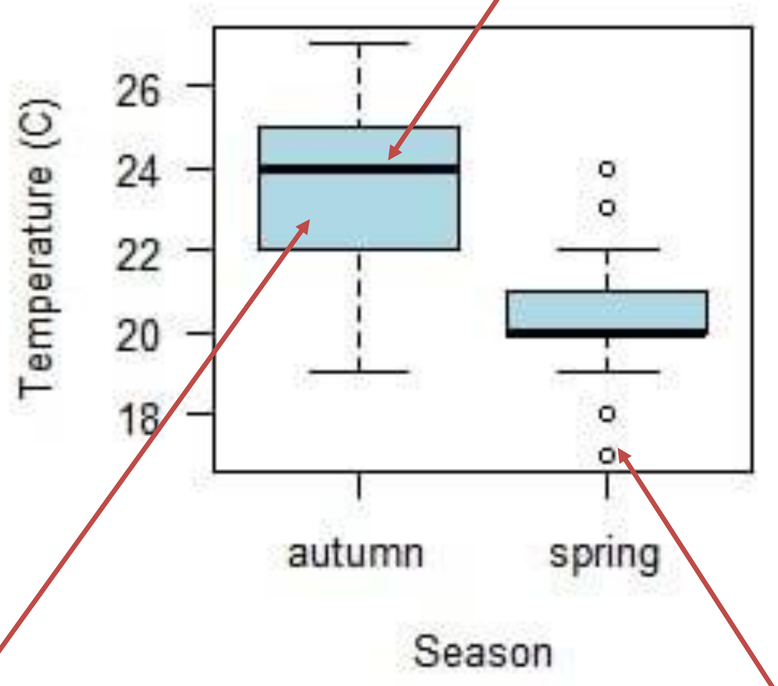
Si medimos desde la persona más baja hasta la más alta, tendremos una idea del **rango total** de alturas.

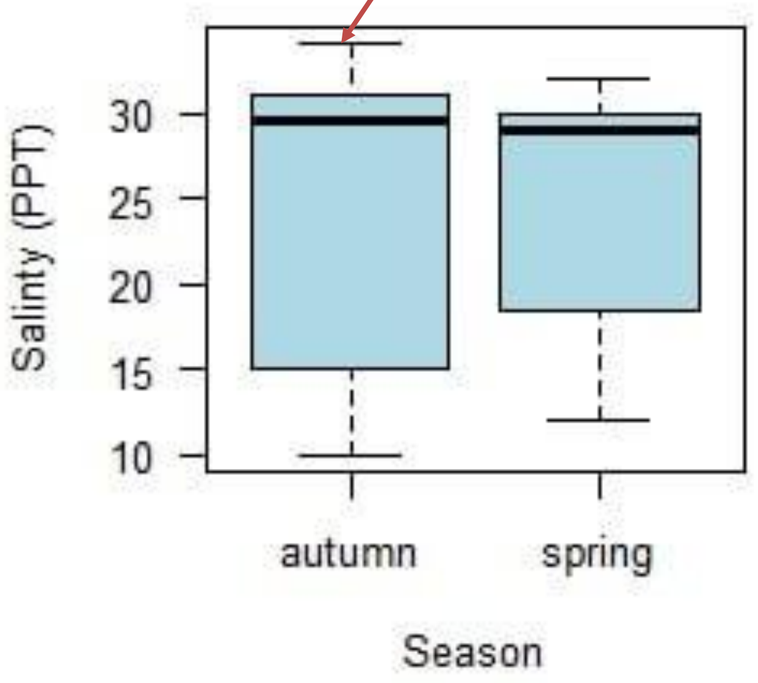
Pero en realidad, queremos saber qué tan apretado está el grupo del medio, sin contar a las personas extremadamente bajas o altas.

Para eso, eliminamos al **25% de las personas más bajas** (Q1) y al **25% de las personas más altas** (Q3), quedándonos con el **50% central**.

Mediana

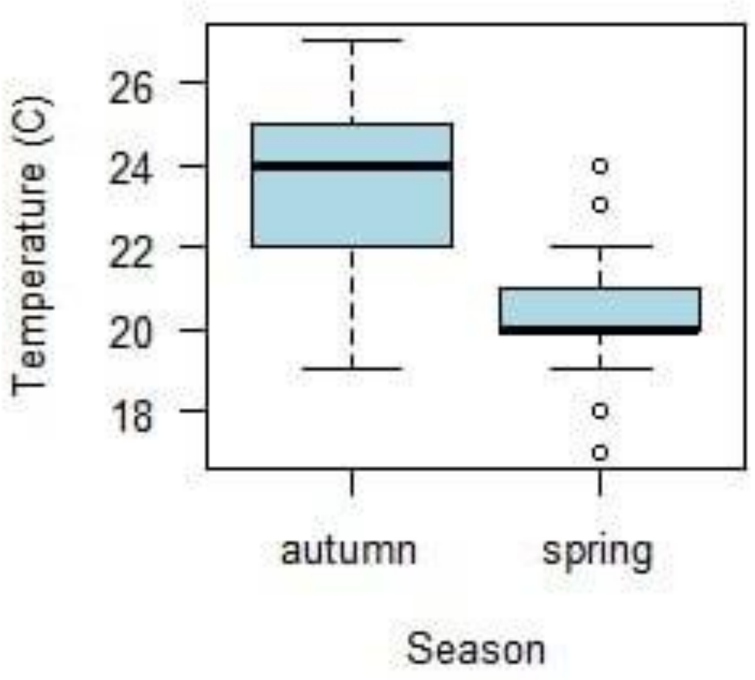
Rango Extendido

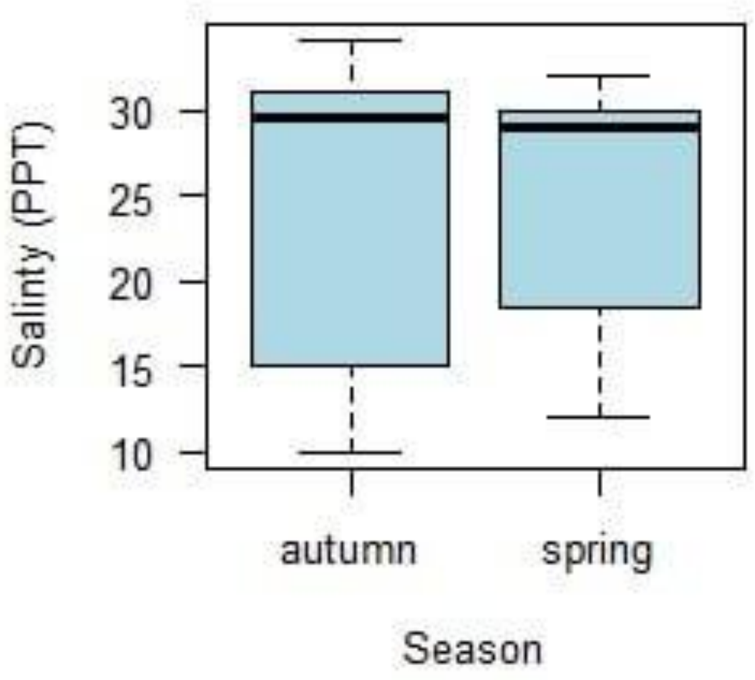


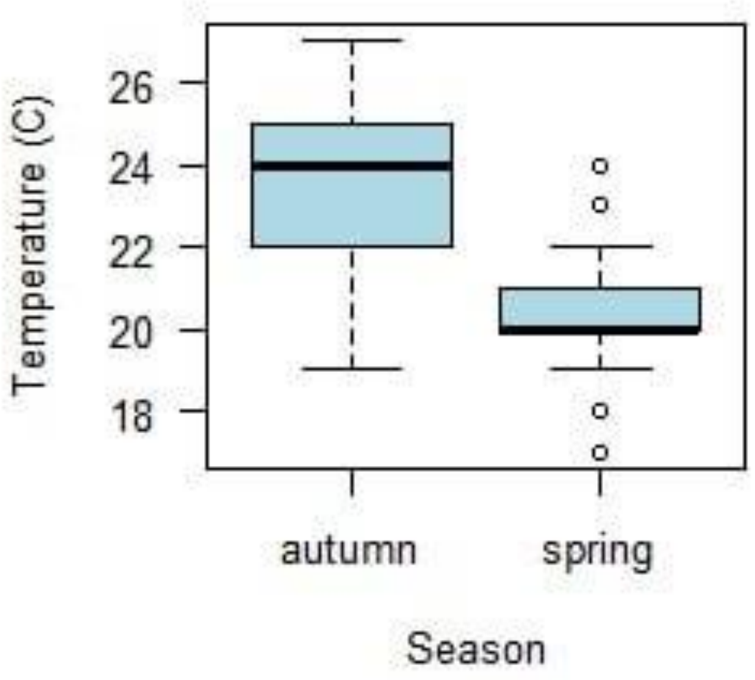

Representa el rango donde se encuentra la mayoría de los datos

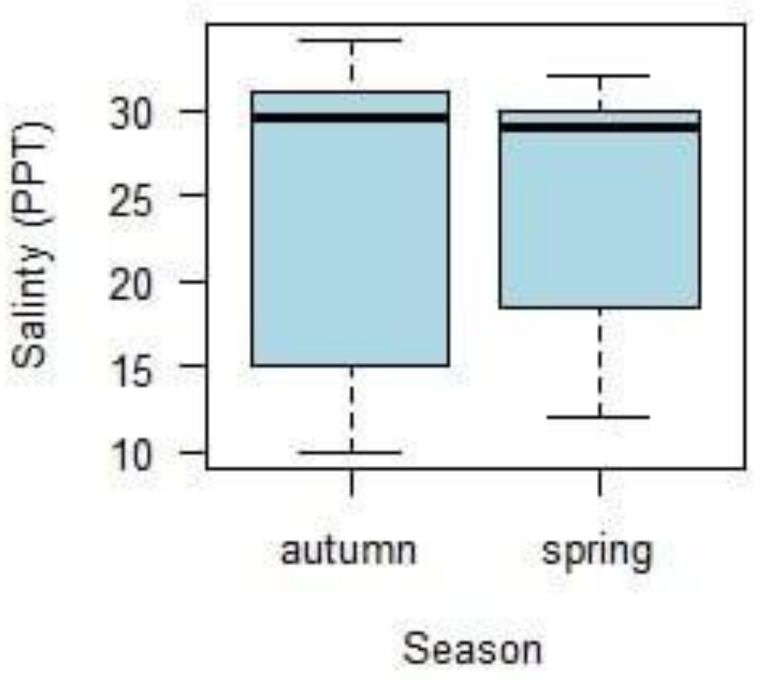
Dato Atípicos o Extremos




No siempre el **mínimo y máximo** valor están representados por los extremos de los bigotes.

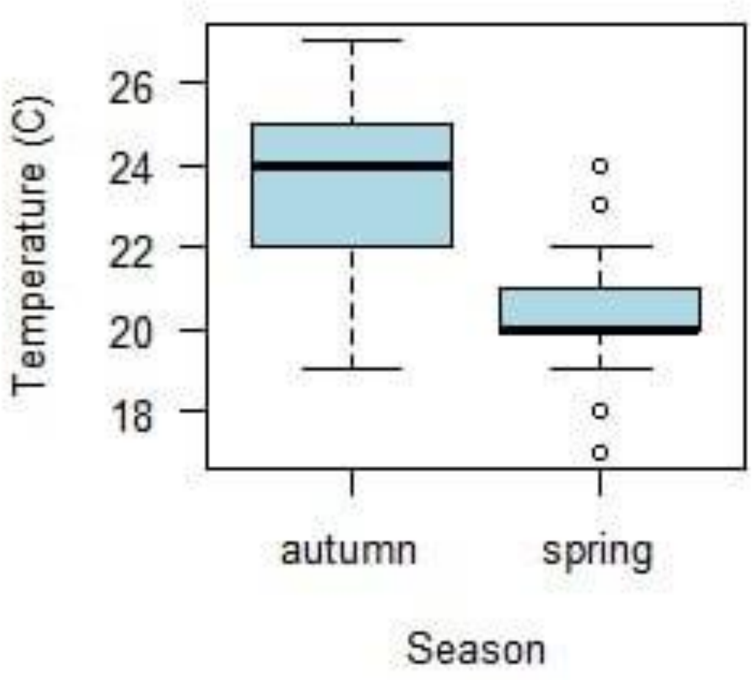


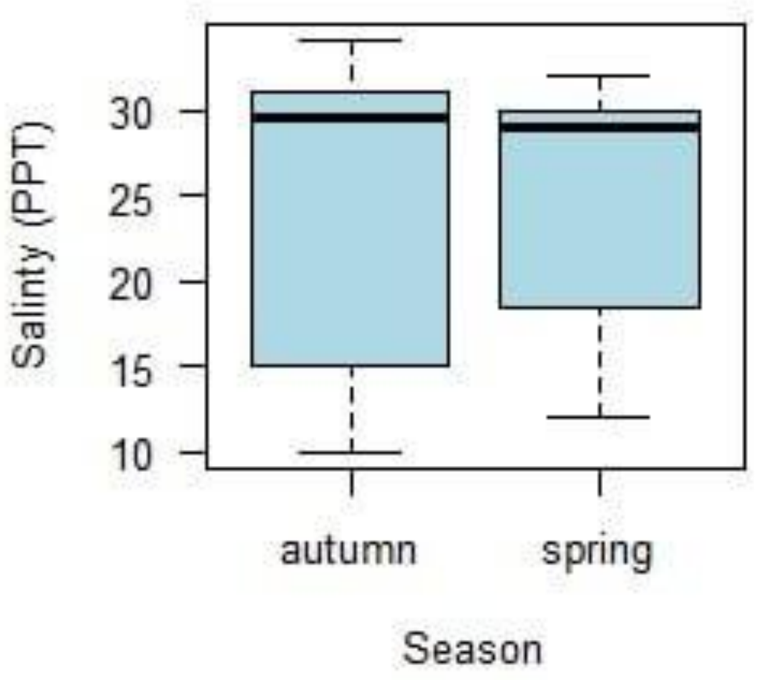




No siempre el **mínimo y máximo** valor están representados por los extremos de los bigotes.

Los bigotes **no marcan el valor más extremo posible**, sino que siguen una regla matemática (generalmente 1.5 veces el rango intercuartil).

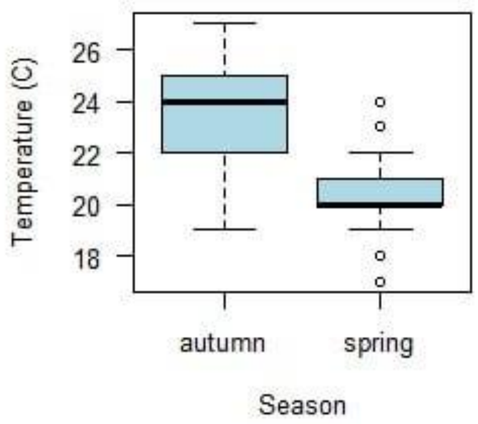
DuocUC®

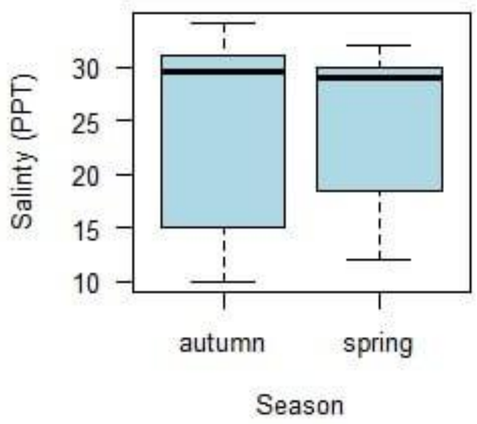



No siempre el **mínimo y máximo** valor están representados por los extremos de los bigotes.

Si hay valores muy alejados de la mayoría, se dibujan como puntos fuera de los bigotes para indicar que son excepcionales o raros.



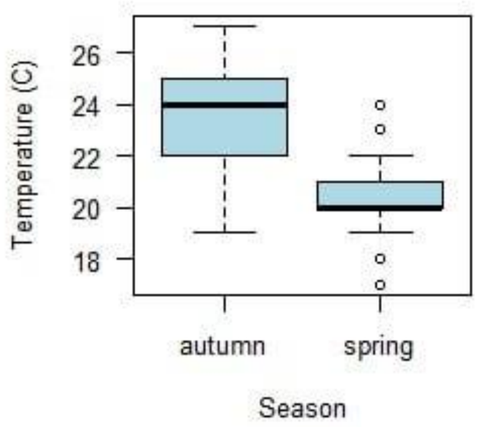

**Imagina que tienes estos datos de temperatura en otoño:**
**[20, 21, 22, 23, 24, 24, 25, 25, 26, 27]**

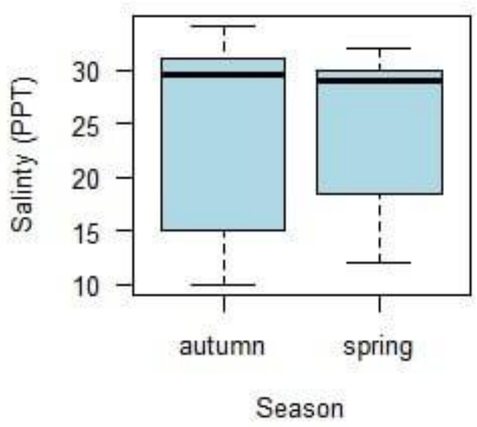
**Q1 = 22.5 (porque el 25% de los datos están por debajo)**

**Q3 = 25.5 (porque el 75% de los datos están por debajo)**

**IQR** = Q3 - Q1 = **25.5 - 22.5 = 3**

**Ahora calculamos los límites de los bigotes:**
- **Límite inferior: 22.5 - (1.5 × 3) = 18**
- **Límite superior: 25.5 + (1.5 × 3) = 30**




**Imagina que tienes estos datos de temperatura en primavera:**
[18, 19, 20, 20, 21, 22, 22, 23, 23, 24, 30]

**Q1 = 20 (el 25% de los datos están por debajo)**

**Q3 = 23 (el 75% de los datos están por debajo)**

## IQR = Q3 - Q1 = 23 - 20 = 3

**Ahora calculamos los límites de los bigotes:**
- **Límite inferior = Q1 - 1.5 × IQR = 20 - 1.5(3) = 15.5**
- **Límite superior = Q3 + 1.5 × IQR = 23 + 1.5(3) = 27.5**

# FUNDAMENTOS DE MACHINE LEARNING

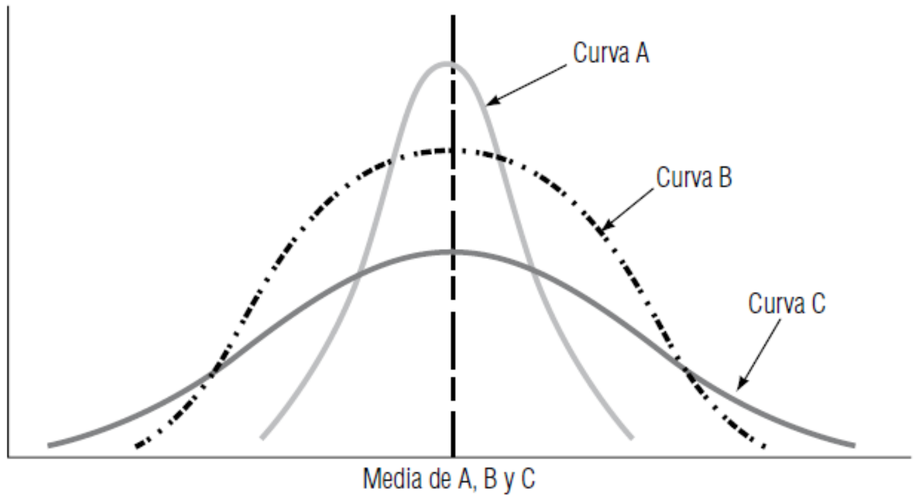
## Estadística Descriptiva II

Medidas de Dispersión

DuocUC®

ESCUELA DE
INFORMÁTICA Y
TELECOMUNICACIONES

# DISPERSIÓN




## Dispersión

En esta figura, la media de las tres curvas es la misma, pero la curva A tiene menor separación (o *variabilidad*) que la curva B, y esta a su vez tiene menor variabilidad que la C.

Si medimos sólo la media de estas tres distribuciones, estaremos pasando por alto una diferencia importante que existe entre las tres curvas.

Al igual que sucede con cualquier conjunto de datos, la media, la mediana y la moda sólo nos revelan una parte de la información que debemos conocer acerca de las características de los datos.

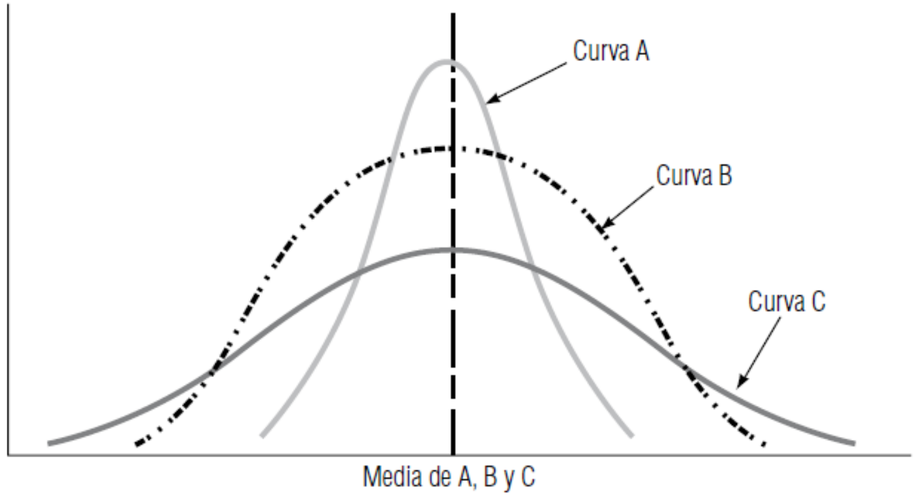
Para aumentar nuestro entendimiento del patrón de los datos, **debemos medir también su *dispersión*, separación o variabilidad**.

# DISPERSIÓN

## Dispersión

Tiempos de entrega en dos empresas.

| Empresa A (minutos de retraso) | Empresa B (minutos de retraso) |
|---|---|
| 8 | 1 |
| 9 | 15 |
| 10 | 30 |
| 11 | 45 |
| 12 | 60 |

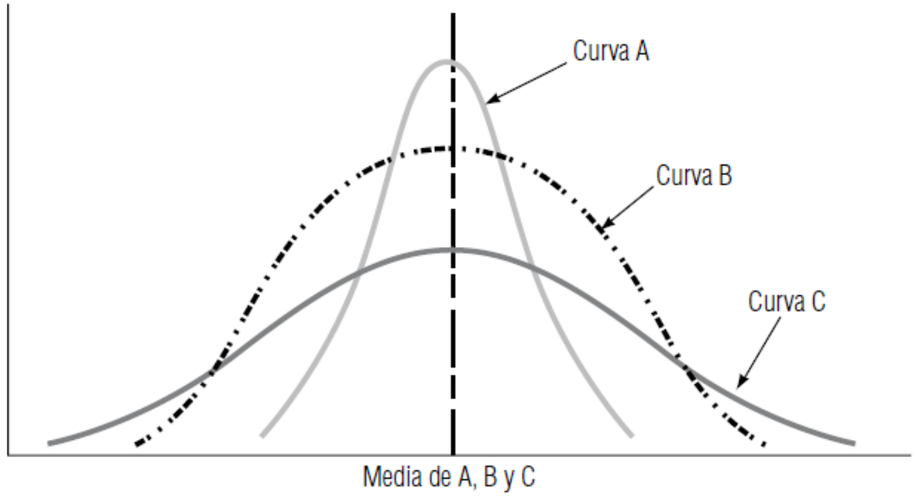


# DISPERSIÓN

## Dispersión

Tiempos de entrega en dos empresas.

| Empresa A (minutos de retraso) | Empresa B (minutos de retraso) |
|---|---|
| 8 | 1 |
| 9 | 15 |
| 10 | 30 |
| 11 | 45 |
| 12 | 60 |




**Empresa A tiene baja dispersión** (todos los valores están entre 8 y 12 minutos).

**Empresa B tiene alta dispersión (los valores van desde 1 hasta 60 minutos).**

# DISPERSIÓN

## Dispersión

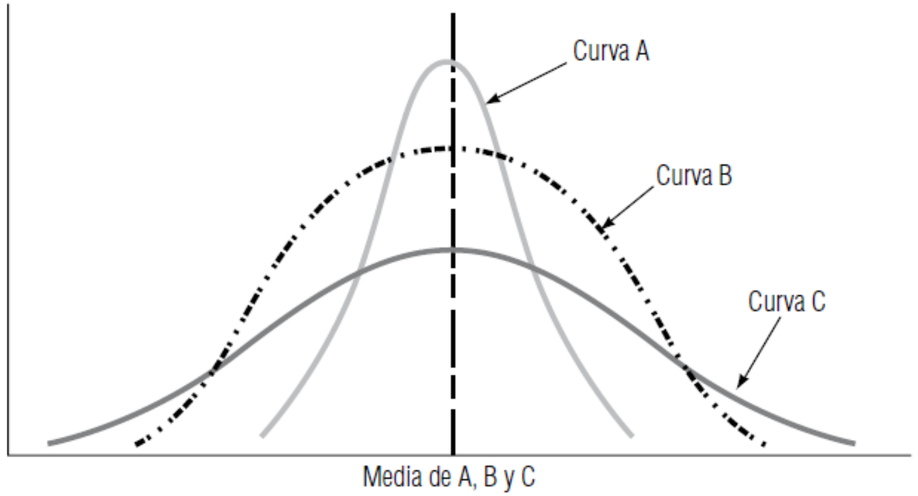
Tiempos de entrega en dos empresas.

| Empresa A (minutos de retraso) | Empresa B (minutos de retraso) |
|---|---|
| 8 | 1 |
| 9 | 15 |
| 10 | 30 |
| 11 | 45 |
| 12 | 60 |

**Empresa A tiene baja dispersión** (todos los valores están entre 8 y 12 minutos).

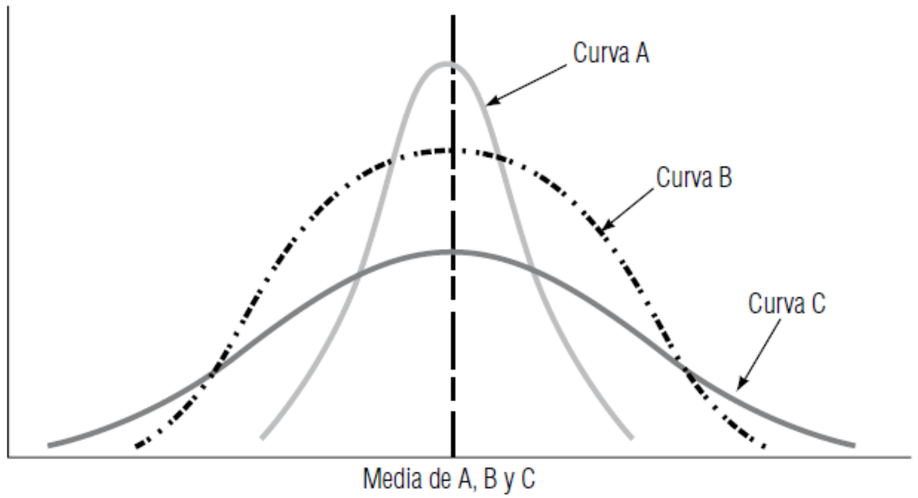
**Empresa B tiene alta dispersión (los valores van desde 1 hasta 60 minutos).**

Esto significa que **Empresa A es más consistente**, mientras que **Empresa B es impredecible**: a veces entrega muy rápido, pero otras veces se retrasa demasiado.



# DISPERSIÓN

¿Por qué es importante entenderla?




## PRIMERO

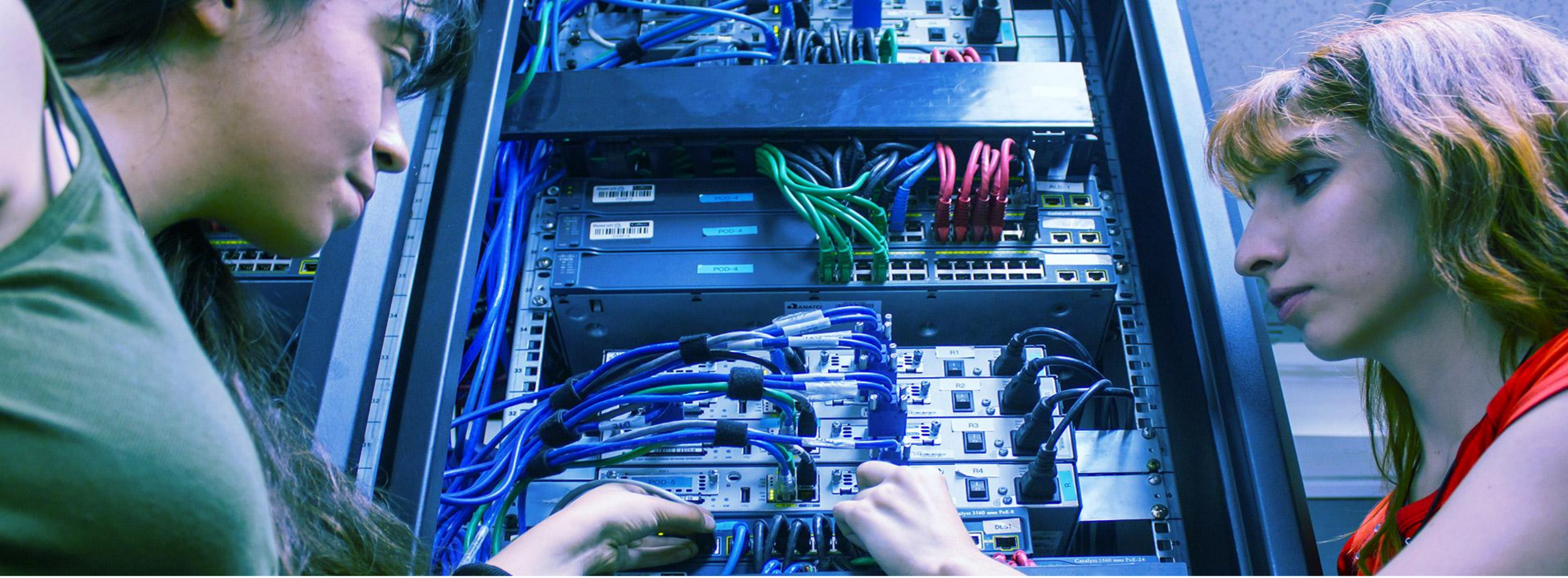
Nos proporciona información adicional que **nos permite juzgar la confiabilidad de nuestra medida de tendencia central**. Si los datos se encuentran muy dispersos, como los que representa la curva C de la figura**, la posición central es menos representativa de los datos,** como un todo, que cuando éstos se agrupan más cerca alrededor de la media, como en la curva A de la misma figura.

## SEGUNDO

Ya que existen problemas característicos para datos muy dispersos, **debemos ser capaces de reconocer esa dispersión** amplia para poder abordar esos problemas.

## TERCERO

Si no se desea tener una amplia dispersión de valores con respecto del centro de distribución, o esto presenta riesgos inaceptables, **necesitamos poder reconocerla y evitar elegir distribuciones que tengan las dispersiones más grandes**.



# DISPERSIÓN

MEDIDAS DE DISPERSIÓN PROMEDIO : Las descripciones más completas de la dispersión son aquellas que manejan la desviación promedio respecto a alguna medida de tendencia central. Dos de estas medidas son importantes: la varianza y la desviación estándar.

# DISPERSIÓN

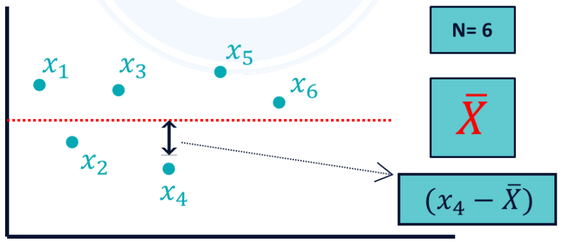
VARIANZA DE LA POBLACIÓN

## VARIANZA

$$\sigma^2 = \frac{\sum_1^N (x_i - \bar{X})^2}{N}$$

- **X** → Variable
- **$x_i$** → Observación número i de la variable X.
- **N** → Número de observaciones.
- **$\bar{X}$** → Es la media de la variable X.

**Es una medida de dispersión que representa la variabilidad de una serie de datos respecto a su media.**




## VARIANZA

Para calcular la varianza de una población, la suma de los cuadrados de las distancias entre la media y cada elemento de la población se divide entre el número total de observaciones en la población.

**Varianza de población**

$$\sigma^2 = \frac{\Sigma(x-\mu)^2}{N} = \frac{\Sigma x^2}{N} - \mu^2$$

donde:

- $\sigma^2$ = varianza de la población
- $x$ = elemento u observación
- $\mu$ = media de la población
- $N$ = número total de elementos de la población
- $\Sigma$ = suma de todos los valores $(x-\mu)^2$, o todos los valores $x^2$

# DISPERSIÓN

## VARIANZA DE LA POBLACIÓN

Supongamos que cinco estudiantes sacan estas calificaciones en un examen:

✅ **Grupo A (baja varianza):** 70, 72, 71, 69, 70
✅ **Grupo B (alta varianza):** 50, 90, 40, 100, 60

Aunque el promedio puede ser similar, el **Grupo B tiene más varianza**, porque las notas están más separadas entre sí.

# DISPERSIÓN

## VARIANZA

VARIANZA DE LA POBLACIÓN

$$\sigma^2 = \frac{\Sigma (x - \mu)^2}{N}$$

- Al elevar al cuadrado cada distancia, logramos que todos los números sean positivos y, al mismo tiempo, asignamos más peso a las desviaciones más grandes (desviación es la distancia entre la media y un valor).

- Sin embargo, esta operación (elevar al cuadrado) provoca que las unidades son ***el cuadrado de las unidades de los datos***; por ejemplo “metros al cuadrado”.

- Estas unidades no son intuitivamente claras o fáciles de interpretar. Por esto debemos hacer un cambio significativo en la varianza para calcular una medida útil de la desviación que no nos dé problemas con las unidades de medida y, en consecuencia, sea menos confusa.

# DISPERSIÓN

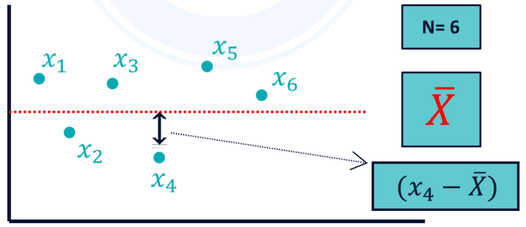
DESVIACIÓN ESTÁNDAR DE LA POBLACIÓN

## DESVIACIÓN ESTÁNDAR

$$\sigma = \sqrt{\frac{\sum_1^N (x_i - \bar{X})^2}{N}}$$

- **X** → Variable
- $\mathbf{x_i}$ → Observación número i de la variable X.
- **N** → Número de observaciones.
- $\bar{\mathbf{x}}$ → Es la media de la variable X.

**También conocida como desviación típica $\sigma$ es una medida que ofrece información sobre la dispersión media de una variable.**




## DESVIACIÓN ESTÁNDAR

La desviación estándar, es simplemente la raíz cuadrada de la varianza de la población. Como la varianza es el promedio de los cuadrados de las distancias de las observaciones a la media, **la desviación estándar es la raíz cuadrada del promedio de los cuadrados de las distancias entre las observaciones y la media**.

**Desviación estándar de la población**

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\Sigma(x-\mu)^2}{N}} = \sqrt{\frac{\Sigma x^2}{N} - \mu^2}$$

donde,

- $x$ = observación
- $\mu$ = media de la población
- $N$ = número total de elementos de la población
- $\Sigma$ = suma de todos los valores $(x-\mu)^2$, o todos los valores $x^2$
- $\sigma$ = desviación estándar de la población
- $\sigma^2$ = varianza de la población

# DISPERSIÓN

DESVIACIÓN ESTÁNDAR DE LA POBLACIÓN

$$\sigma = \sqrt{\frac{\sum_i^N (X_i - \bar{X})^2}{N}}$$

## DESVIACIÓN ESTÁNDAR

- Mientras que la varianza se expresa con el cuadrado de las unidades utilizadas para medir los datos, la desviación estándar está en las mismas unidades que las que se usaron para medir los datos.

- La raíz cuadrada de un número positivo puede ser positiva o negativa. Sin embargo, cuando obtenemos la raíz cuadrada de la varianza para calcular la desviación estándar, los especialistas en estadísticas sólo consideran la raíz cuadrada positiva.

# DISPERSIÓN

DISPERSIÓN RELATIVA

| COEFICIENTE DE VARIACIÓN | $CV$ |
|---|---|
| El *coeficiente de variación* indica que tan grande es la desviación estándar en relación al promedio. | |
| Como se Calcula | Ejemplo |
| $CV = \frac{S}{\bar{x}}\ 100\ \%$ | $CV = \frac{9{,}13}{443{,}39}\ 100\% = 2{,}06\ \%$ |
| Interpretación | |
| La distribución promedio de los precios de un apartamento de una habitación presenta menor variación o es menos heterogénea. | |

## EL COEFICIENTE DE VARIACIÓN

La desviación estándar es una ***medida absoluta*** de la dispersión que expresa la variación en las mismas unidades que los datos originales. Muchas veces se requiere una ***medida relativa*** que proporcione una estimación de la magnitud de la desviación respecto a la magnitud de la media.

El coeficiente de variación es una de las medidas relativas de dispersión. Relaciona la desviación estándar y la media , expresando la desviación estándar como porcentaje de la media. La unidad de medida, entonces, es “porcentaje”, en lugar de las unidades de los datos originales.

**Coeficiente de variación**

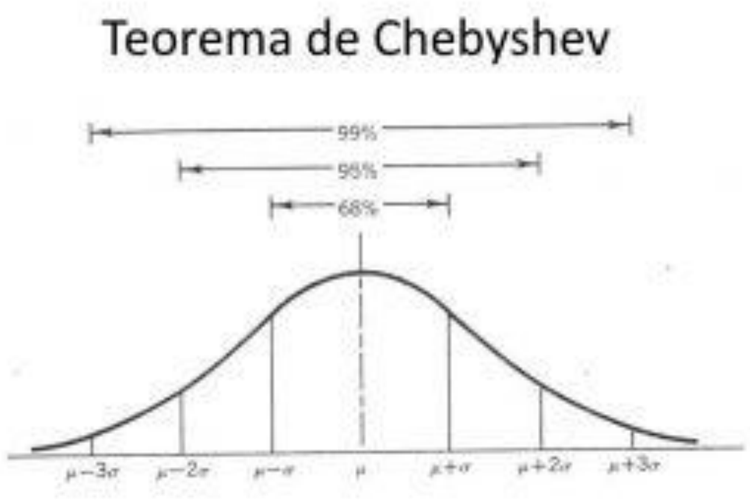
Desviación estándar de la población

$$\text{Coeficiente de variación de la población} = \frac{\sigma}{\mu}\ (100)$$

Media de la población

# DISPERSIÓN

DESVIACIÓN ESTÁNDAR DE LA POBLACIÓN

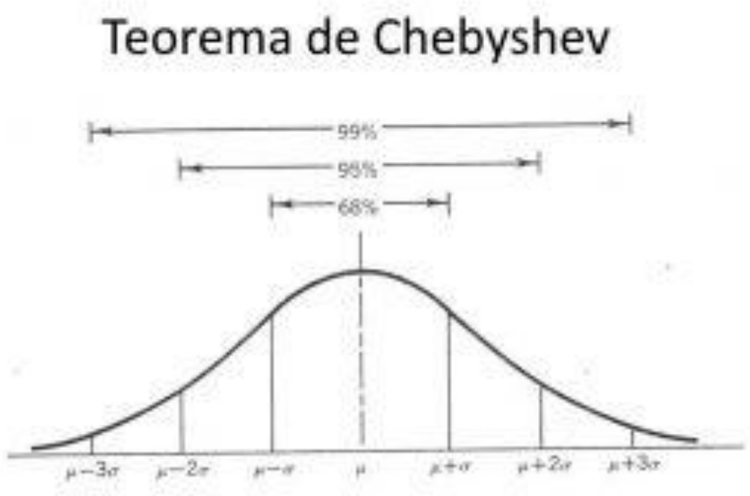



## USOS DE LA DESVIACIÓN ESTÁNDAR

- La desviación estándar nos permite determinar, con un buen grado de precisión, dónde están localizados los valores de una distribución de frecuencias con relación a la media.
- Podemos hacer esto de acuerdo con un teorema establecido por el matemático ruso Chebyshev.
- El teorema establece que independiente de la forma de la distribución, al menos 75% de los valores caen dentro de +/- 2 desviaciones estándar a partir de la media de la distribución, y al menos 89% de los valores caen dentro de los +/- 3 desviaciones estándar a partir de la media.

# DISPERSIÓN

DESVIACIÓN ESTÁNDAR DE LA POBLACIÓN




## USOS DE LA DESVIACIÓN ESTÁNDAR

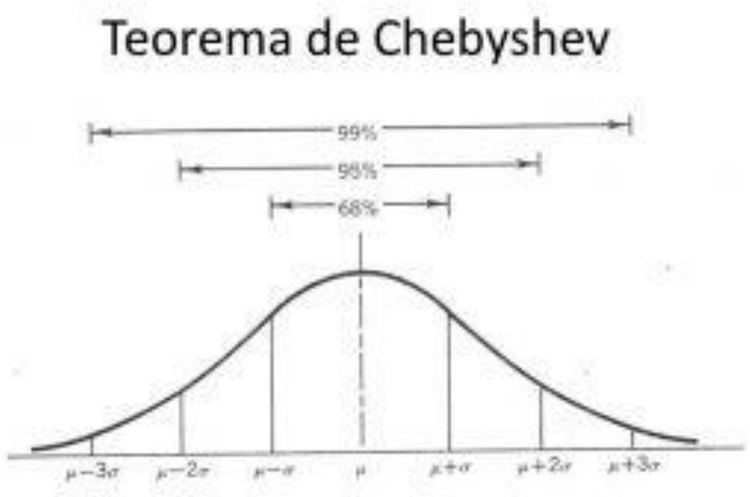
Supongamos que una empresa de envíos tiene un **promedio** de entrega de **5 días** y una **desviación estándar de 2 días**.

📌 **Usando el Teorema de Chebyshev**:

- **75% de los paquetes** se entregarán entre **3 y 7 días** (5 ± 2).
- **89% de los paquetes** se entregarán entre **1 y 9 días** (5 ± 3).

# DISPERSIÓN

DESVIACIÓN ESTÁNDAR DE LA POBLACIÓN




## USOS DE LA DESVIACIÓN ESTÁNDAR

Supongamos que una empresa de envíos tiene un **promedio** de entrega de **5 días** y una **desviación estándar de 2 días**.

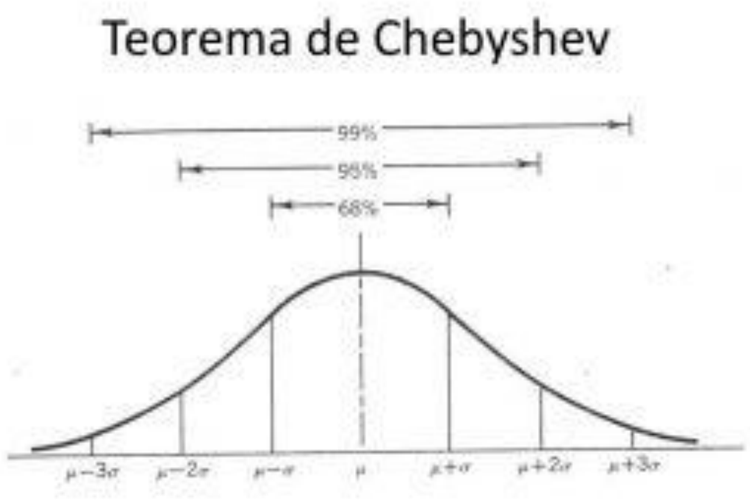
📌 **Usando el Teorema de Chebyshev**:

- **75% de los paquetes** se entregarán entre **3 y 7 días** (5 ± 2).
- **89% de los paquetes** se entregarán entre **1 y 9 días** (5 ± 3).

Esto ayuda a la empresa a saber cuánto tiempo tardarán en entregar la mayoría de los pedidos.

# DISPERSIÓN

DESVIACIÓN ESTÁNDAR DE LA POBLACIÓN

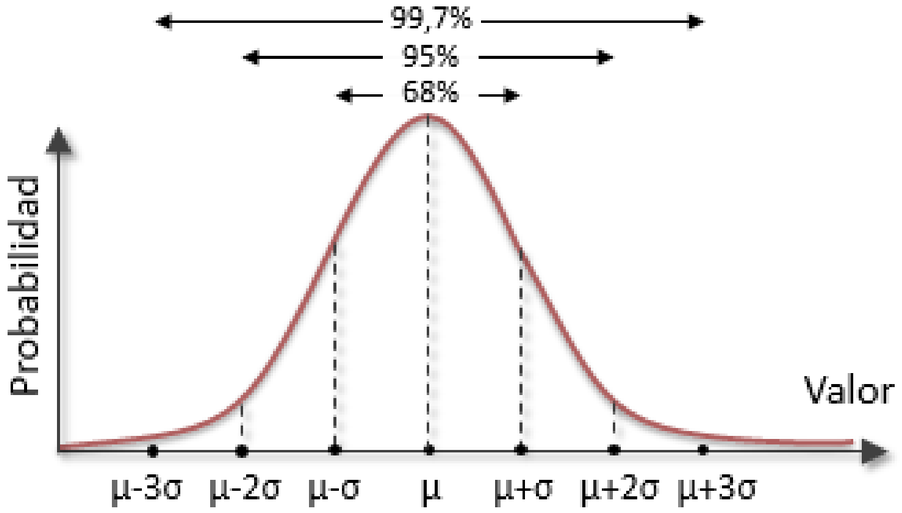



## USOS DE LA DESVIACIÓN ESTÁNDAR

- Podemos medir aún con más precisión el porcentaje de observaciones que caen dentro de un rango específico de una curva simétrica con forma de campana.
- Aproximadamente 68% de los valores de la población cae dentro de +/- 1 desviación estándar a partir de la media.
- Aproximadamente 95% de los valores de la población cae dentro de +/-2 desviación estándar a partir de la media.
- Aproximadamente 99% de los valores de la población cae dentro de +/-3 desviación estándar a partir de la media.

# DISPERSIÓN

DESVIACIÓN ESTÁNDAR DE LA POBLACIÓN

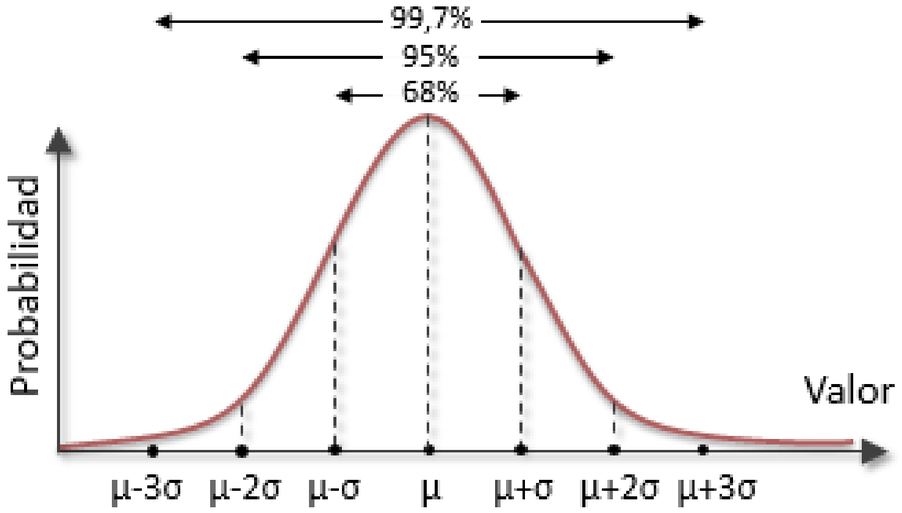



## USOS DE LA DESVIACIÓN ESTÁNDAR

- Más adelante veremos una aplicación práctica de esto con la llamada Distribución Gaussiana, la cual es una popular ***distribución de probabilidad continua*** para cualquier variable aleatoria.
- Este tipo de distribución se caracteriza por dos parámetros : la media y la desviación estándar.
- La mayoría de los conjuntos de datos en Machine Learning siguen este tipo de distribución.

# DISPERSIÓN

DESVIACIÓN ESTÁNDAR DE LA POBLACIÓN




**Notas en un examen** 📚

Si en un examen la **media es 70 puntos** y la **desviación estándar es 10**:

- **68% de los estudiantes** tendrán notas entre **60 y 80** (70 ± 10).
- **95% de los estudiantes** tendrán notas entre **50 y 90** (70 ± 2×10).
- **99.7% de los estudiantes** tendrán notas entre **40 y 100** (70 ± 3×10).

Esto ayuda a los profesores a entender **qué tan comunes son ciertos puntajes** y si hay muchos alumnos con calificaciones extremas.

# RESUMEN

- Concepto de Dispersión
- Varianza
- Desviación Estándar

# FUNDAMENTOS DE MACHINE LEARNING

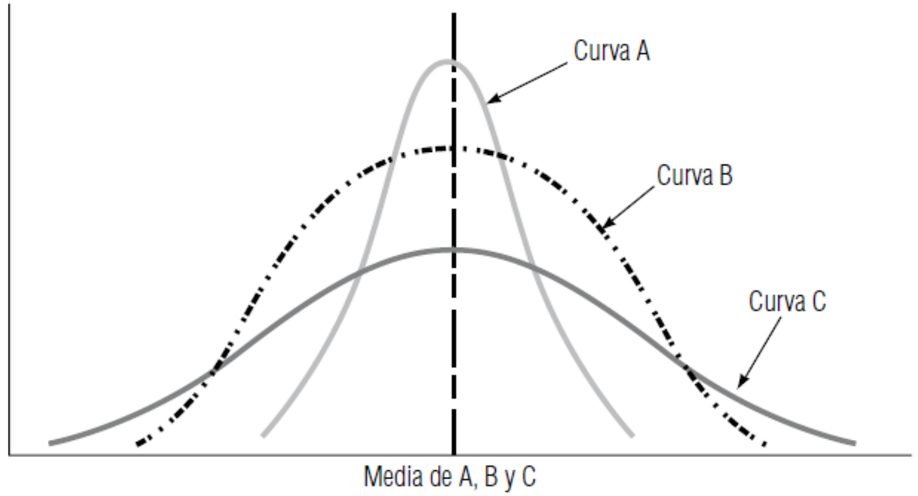
## Estadística Descriptiva II

Análisis de Varianza

DuocUC®

ESCUELA DE
INFORMÁTICA Y
TELECOMUNICACIONES

# ANOVA




## ANALISIS DE VARIANZA

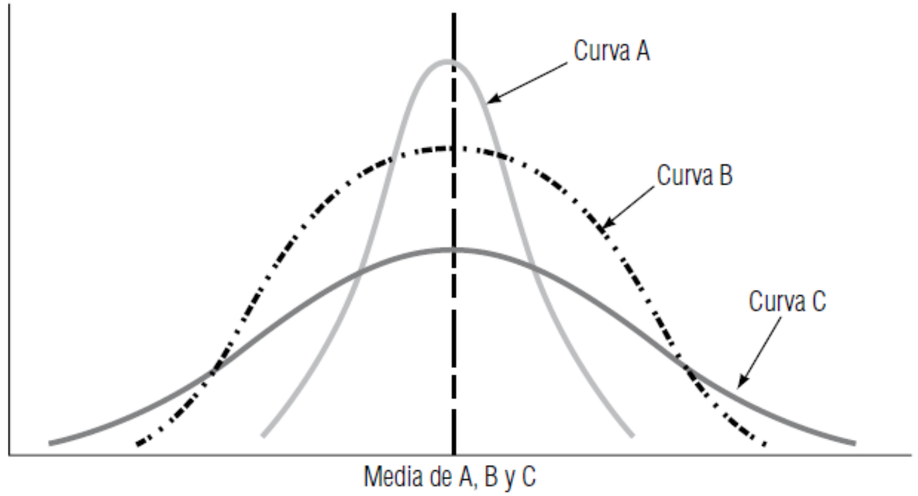
Uno de los grandes desafíos en Machine Learning es poder seleccionar los datos adecuados (conocidos como "**características**") que permitan entrenar adecuadamente un modelo, con el fin **de predecir** correctamente los resultados.

Para ello, sólo se requiere identificar los datos (características) que influyen en gran medida en la variable a predecir. Pero ¿qué ocurre si dicha variable es continua y nuestros datos (características predictoras) son categóricos?

Piensa, por ejemplo, en como predecir el precio de la gasolina (variable continua) a partir de precios por tipo de gasolina (variable categórica).

El análisis de varianza, en inglés Analysis of Variance o ANOVA, nos puede ayudar en esta selección.

# ANOVA




## PLANTEAMIENTO DE HIPÓTESIS

ANOVA requiere que, en primer lugar, se haga un planteamiento formal de las hipótesis nula y alternativa que se desean probar:

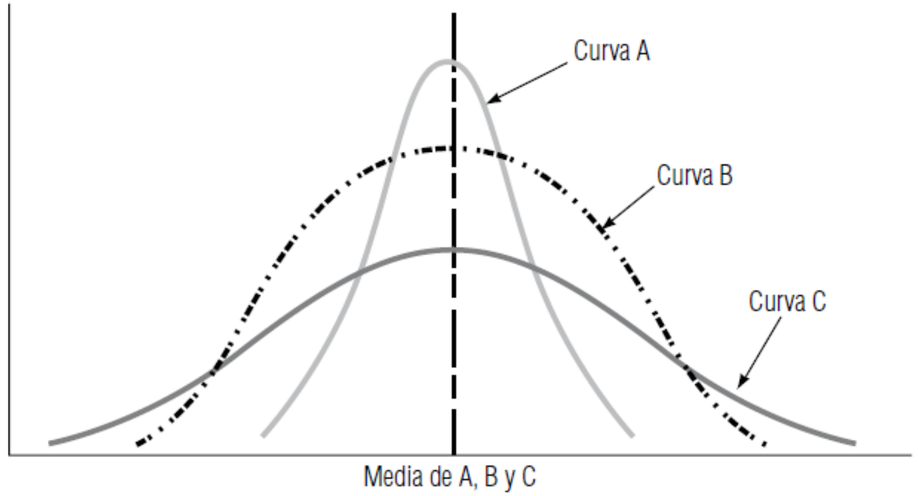
$H_0$: $\mu_1 = \mu_2 = \mu_3$ ← Hipótesis nula

$H_1$: $\mu_1$, $\mu_2$ y $\mu_3$ *no* son todas iguales ← Hipótesis alternativa

En este caso la Hipótesis nula (Ho) nos indica que, si las medias de todos los grupos de datos son iguales, dichos datos o características no están correlacionados con la predicción y sirven para el modelo de Machine Learning.

Por otra parte, si encontramos entre las medias de los grupos diferencias demasiado grandes para atribuirlas a un error aleatorio de muestreo (Hipótesis Alternativa o H1), podemos inferir que no aportan al modelo de Machine Learning.

# ANOVA




## TRES PASOS

El Análisis de Varianza está basado en una comparación de dos estimaciones de varianza de la población total de datos: la varianza entre las medias muestrales y la varianza dentro de las muestras.

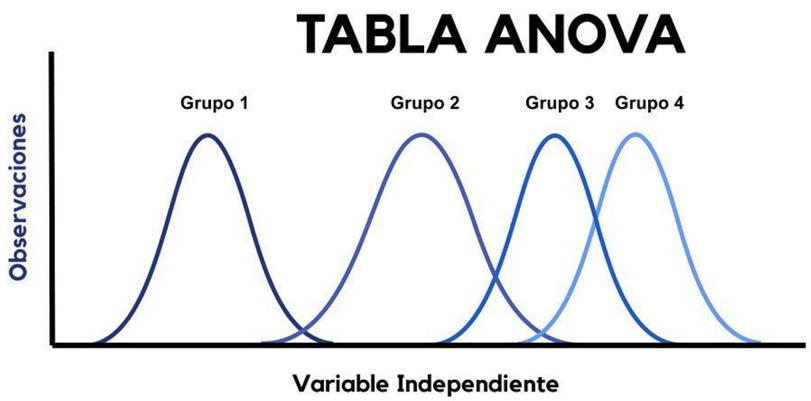
Así, los pasos del análisis de la varianza son:

1. Determinar una estimación de la varianza de la población a partir de **la varianza entre las medias** de las muestras.

2. Determinar una segunda estimación de la varianza de la población a partir de **la varianza dentro de las muestras**.

3. Comparar estas estimaciones. Si su valor es aproximadamente igual, se acepta la hipótesis nula.

# ANOVA




## CÁLCULO ENTRE LAS MEDIAS

**El paso 1** en el Análisis de la Varianza, indica que debemos obtener una estimación de la varianza de la población a partir de la varianza entre las medias de las muestras.

**Estimación de la varianza entre columnas**

Primera estimación de la varianza de la población

$$\hat{\sigma}_b^2 = \frac{\Sigma n_j (\bar{x}_j - \bar{\bar{x}})^2}{k - 1}$$

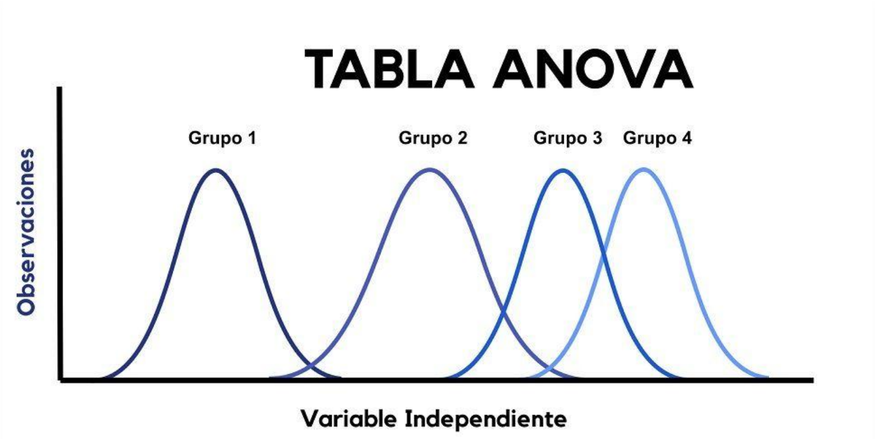
donde,

- $\hat{\sigma}_b^2$ = nuestra primera estimación de la varianza de la población, basada en la varianza entre las medias de las muestras (la *varianza entre columnas*)
- $n_j$ = tamaño de la $j$-ésima muestra
- $\bar{x}_j$ = media muestral de la $j$-ésima muestra
- $\bar{\bar{x}}$ = gran media
- $k$ = número de muestras

# ANOVA

## CÁLCULO DENTRO DE LAS MUESTRAS

**El paso 2** requiere una segunda estimación de la varianza de la población, basada en la varianza dentro de las muestras.




**Estimación de la varianza dentro de columnas**

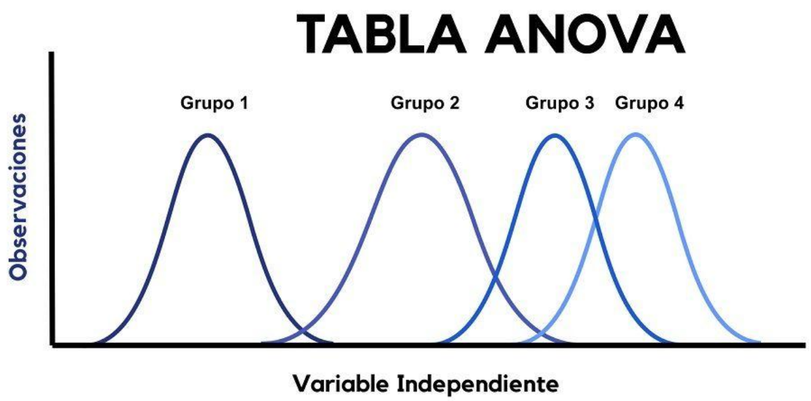
Segunda estimación de la varianza de la población

$$\hat{\sigma}_w^2 = \Sigma\left(\frac{n_j - 1}{n_T - k}\right)s_j^2$$

donde,

- $\hat{\sigma}_w^2$ = nuestra segunda estimación de la varianza de la población, basada en las varianzas dentro de las muestras (la *varianza dentro de columnas*)
- $n_j$ = tamaño de la $j$-ésima muestra
- $s_j^2$ = varianza muestral de la $j$-ésima muestra
- $k$ = número de muestras
- $n_T = \Sigma n_j$ = tamaño de la muestra total

# ANOVA




## PRUEBA DE HIPÓTESIS F

**En el paso 3** de ANOVA se comparan estas dos estimaciones de la varianza de la población mediante el cálculo de su cociente:

**Estadístico *F***

$$F = \frac{\text{varianza entre columnas}}{\text{varianza dentro de columnas}} = \frac{\hat{\sigma}_b^2}{\hat{\sigma}_w^2}$$

Como resultado, **el denominador y el numerador deben ser aproximadamente iguales si la hipótesis nula es verdadera**. Cuanto más cercano a 1 esté el cociente F, más nos inclinamos a aceptar la hipótesis nula. Al contrario, conforme el cociente F crece, nos inclinaremos más a rechazar la hipótesis nula y a aceptar la alternativa

# ANOVA

| Fuente de variación | Suma de cuadrados | Grados de libertad | Medias cuadráticas | Prueba F |
|---|---|---|---|---|
| Modelos | $\sum(\hat{y}_i - \bar{y})^2$ | 1 | $\frac{\Sigma(\hat{y}_i - \bar{y})^2}{1}$ | $\frac{\frac{\Sigma(\hat{y}_i - \bar{y})^2}{1}}{\frac{\Sigma(y_i - \hat{y}_i)^2}{n-2}}$ |
| Residuales | $\sum(y_i - \hat{y}_i)^2$ | n-2 | $\frac{\Sigma(y_i - \hat{y}_i)^2}{n-2}$ | |
| Total | $\sum(y_i - \bar{y})^2$ | n-1 | | |

## GRADOS DE LIBERTAD

Dependiendo del tamaño de las muestras, las distribuciones son diferentes. Esto se denomina ***grados de libertad***, que podría definirse como el número de valores que podemos escoger libremente.

### Grados de libertad del numerador

$$\text{Número de grados de libertad en el } \textit{numerador} \text{ del cociente } F = (\text{número de muestras} - 1)$$

### Grados de libertad del denominador

$$\text{Número de grados de libertad en el } \textit{denominador} \text{ del cociente } F = \Sigma(n_j - 1) = n_T - k$$

donde,

- $n_j$ = tamaño de la $j$-ésima muestra
- $k$ = número de muestras
- $n_T = \Sigma n_j$ = tamaño de la muestra total

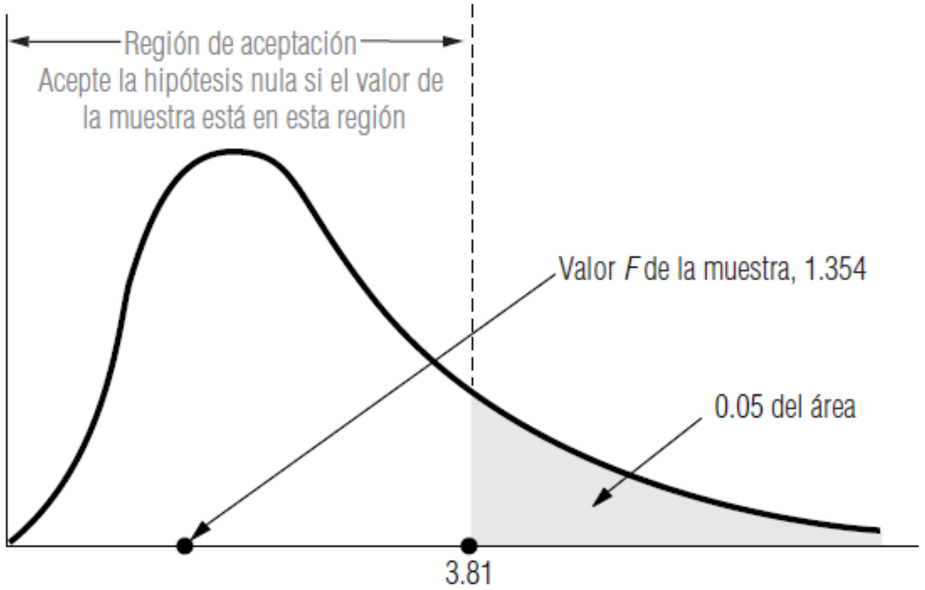
# ANOVA

Pruebas de Hipótesis

## PRUEBA DE HIPÓTESIS

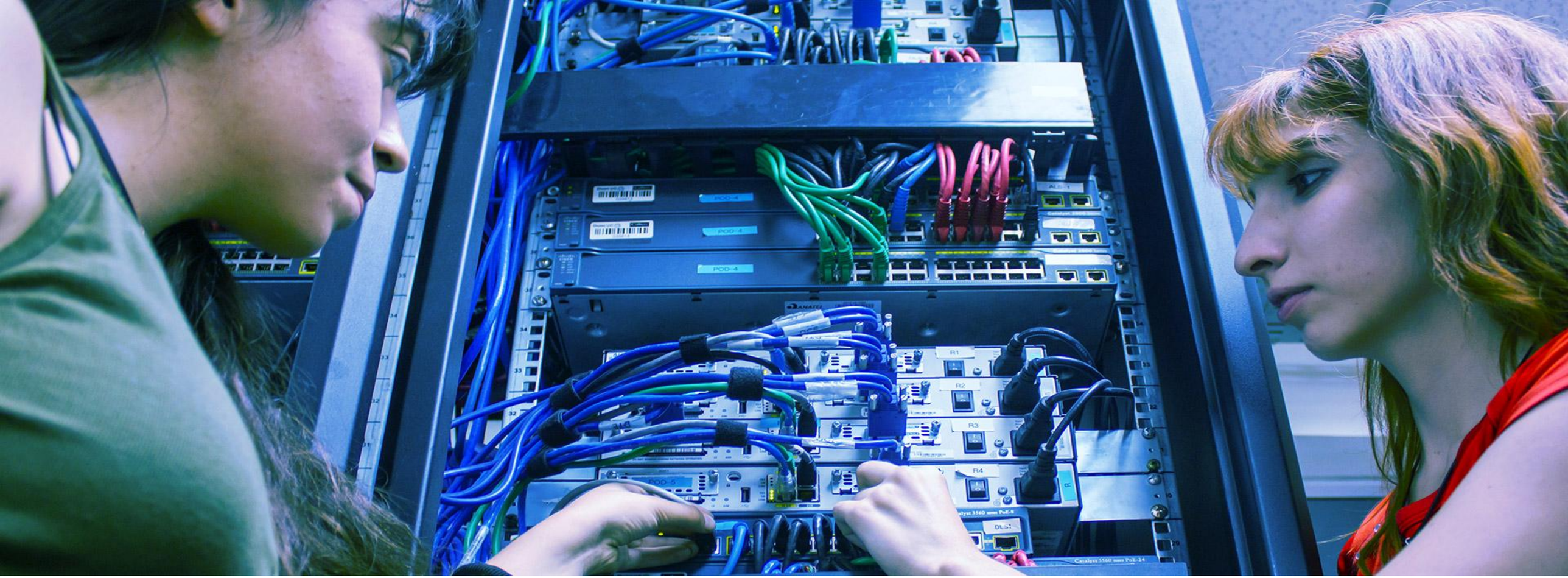
Para llevar a cabo pruebas de hipótesis F debemos utilizar ***una tabla F***, en la cual las columnas representan el número de grados de libertad del numerador y las filas el número de grados de libertad del denominador.

Si nuestro valor calculado de F excede este valor de la tabla, rechazamos la hipótesis nula. Si no es mayor, la aceptamos (ver ejemplo)




**Ejemplo:**

**Valor de F = 1.354**
**Grados de libertad Tabla F = 3.81**

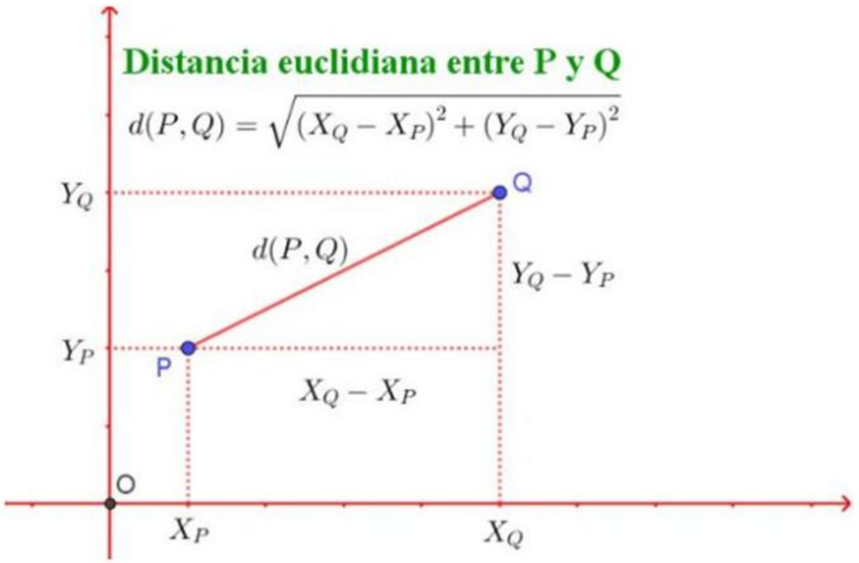


# DISTANCIAS

CÁLCULO DE DISTANCIAS : Muchos algoritmos de Machine Learning requieren calcular distancias entre los puntos observados. Hay diferentes métricas de distancias disponibles. Veremos algunas de las más utilizadas.

# DISTANCIA

Distancia Euclidiana




## DISTANCIA EUCLIDIANA

Es una medida de la distancia  en línea recta entre dos puntos en el espacio euclidiano.

El espacio euclidiano bidimensional es un plano. Los puntos de un plano euclidiano cumplen los axiomas de la geometría de Euclides, por ejemplo:

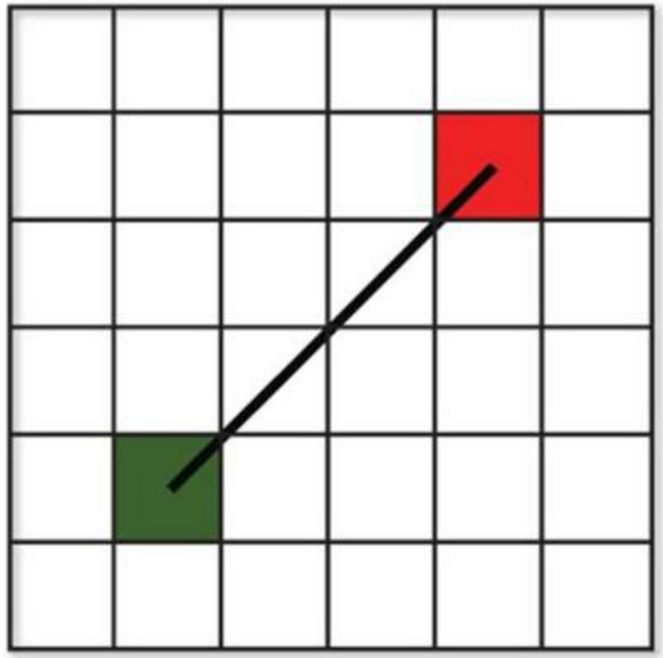
- Por dos puntos pasa una sola recta.
- Tres puntos sobre el plano forman un triángulo cuyos ángulos internos siempre suman 180º.
- En un triángulo rectángulo el cuadrado de la hipotenusa es igual a la suma de los cuadrados de sus catetos.

# DISTANCIA

Distancia Euclidiana

## DISTANCIA EUCLIDIANA

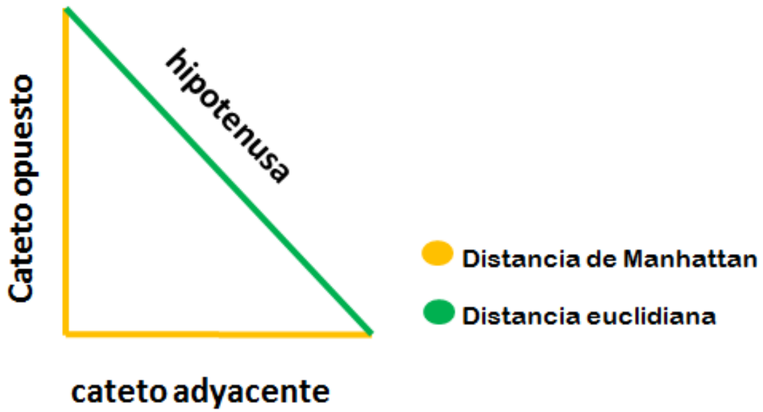
Si suponemos que tenemos dos puntos, el rojo (4,4) y el verde (1,1). Al calcular la distancia usando la métrica de distancia euclidiana se obtiene 4.24



$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

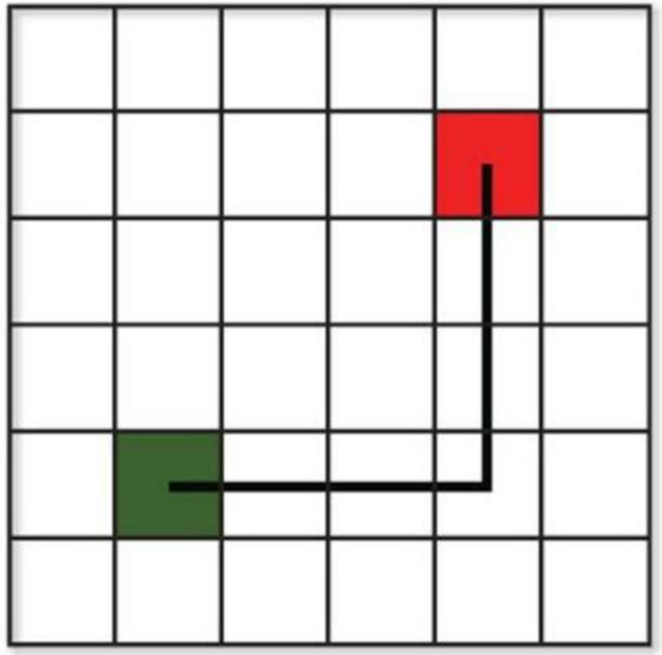
# DISTANCIA

Distancia de Manhattan




## DISTANCIA DE MANHATTAN

Esta distancia también se conoce como distancia de taxi o distancia de “cuadras” de ciudad, esto se debe a la forma en que se calcula. La distancia entre dos puntos es la suma de las diferencias absolutas de sus coordenadas cartesianas.



$$d = \sum_{i=1}^{n} |x_i - y_i|$$

Los dos puntos de la imagen, el rojo(4,4) y el verde(1,1), según el cálculo de distancia de Manhattan están a una distancia de valor 6.

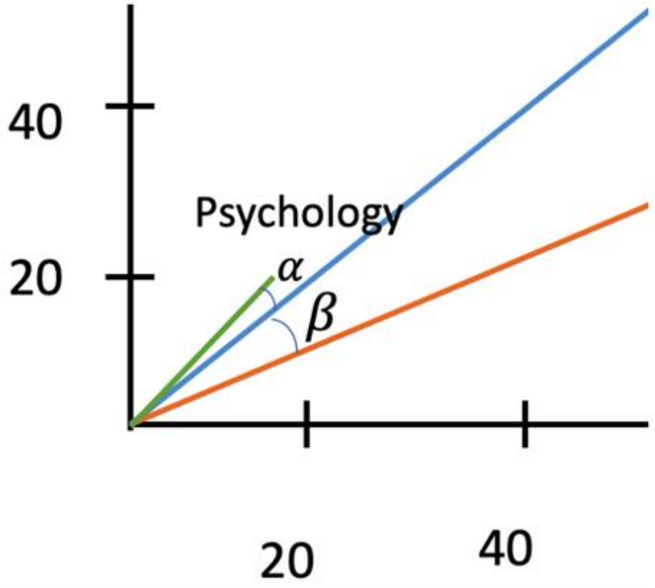
$$d = |4-1| + |4-1| = 6$$

# DISTANCIA

Distancia del Coseno

## DISTANCIA DEL COSENO

Esta métrica de distancia se utiliza principalmente para calcular la similitud entre dos vectores. Se mide por el coseno del ángulo entre dos vectores y determina si dos vectores apuntan en la misma dirección.




$$\cos\theta = \frac{\vec{a} \cdot \vec{b}}{\|\vec{a}\| \cdot \|\vec{b}\|}$$

El valor de similitud es el coseno del ángulo entre dos líneas rectas (vectores).

Cuanto más pequeño sea el ángulo, más similares son los dos vectores representados, y cuanto más grande es el ángulo, menos similares son los dos vectores representados.

# RESUMEN

- ANOVA
- Conceptos de distancias