Colegio de
San Francisco de Paula

Tema 0. Análisis estadístico

Germán Tenorio
Biología NS-Diploma BI

International Baccalaureate

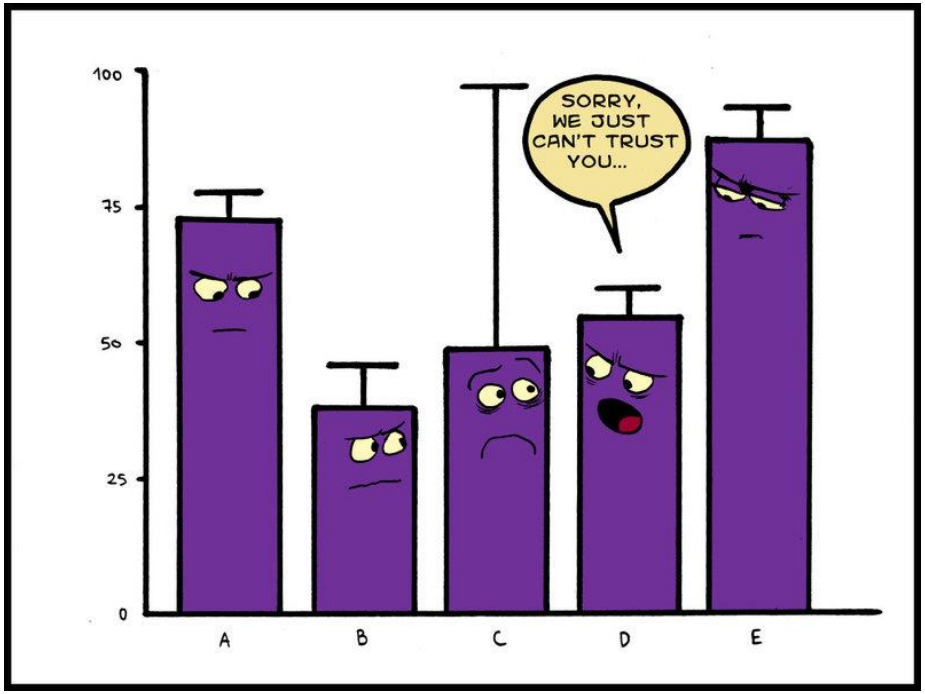


IMAGEN: delusionalmiasma.blogspot.com.es

Para empezar….un vídeo

Bioestadística

- La Bioestadística es la rama de las Matemáticas que permite tomar muestras de pequeñas porciones de hábitats, comunidades o poblaciones biológicas, y sacar conclusiones sobre la población mayor.

IMAGEN:unpasomas.fundacion.telefonica.com

IMAGEN: educabarrie.org

- La estadística mide las diferencias y relaciones entre dos o más conjuntos de **datos**.

¿Qué es un dato?

- Un dato es un antecedente necesario para llegar al conocimiento exacto de una cosa (Diccionario Espasa-Calpe,1988).

IMAGEN: unpasomas.fundacion.telefonica.com

- Los datos pueden tomarse a partir de una población (encuesta).
- Los datos también pueden tomarse a partir de un proceso (experimentación).
- Existen 2 tipos de datos: Cualitativos y cuantitativos.

Datos cuantitativos

- La información que nos proporcionan hace referencia a datos **contados** (discretos) **o medidos** (continuos) usando una escala numérica.
- Ejemplos:
  - Concentración
  - Temperatura
  - Longitud
  - Actividad enzimática…,etc.
- Por tanto, los datos cuantitativos se obtienen mediante la medición (cuantificación).
- Cuando se hacen mediciones pueden ocurrir errores, lo cual hace que los datos obtenidos sean imprecisos. Se distinguen **dos tipos de errores: aleatorios y sistemáticos**.

IMAGEN: danielaza.files.wordpress.com/2012/01/medici1.jpg

Errores sistemáticos versus aleatorios

- Los **errores aleatorios** son aquellos errores inevitables que se producen por eventos imposibles de controlar durante el proceso de medición.
- Son numerosos y pequeños, y su acumulación hace que las medidas fluctúen alrededor de una media.
- Las fuentes de los errores aleatorios son difíciles de identificar o sus efectos no pueden corregirse del todo, pero se pueden mantener en un nivel mínimo mediante una selección meticulosa del material y bajo un cuidado control de las variables.

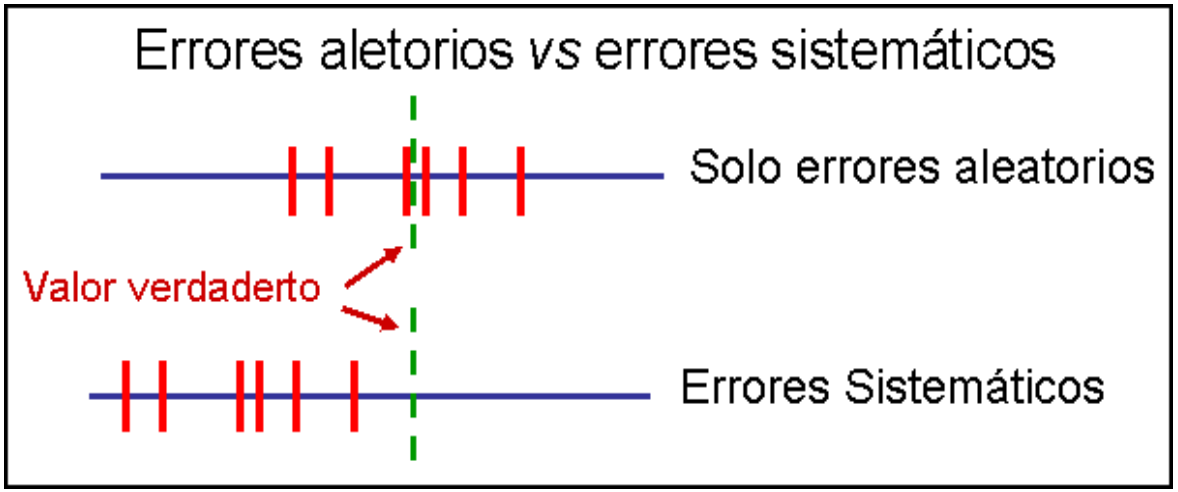


IMAGEN: virtual.unal.edu.co

Errores sistemáticos versus aleatorios

- **Ejemplos de errores aleatorios** son:

  - El tiempo de respuesta a la hora de pulsar un cronómetro, ya que algunas veces se pulsará antes y otras después, por lo que se medirá más o menos.

  - Las fluctuaciones aleatorias a temperatura ambiente.

  - El contenido en agua de distintas patatas procedentes de la misma bolsa.

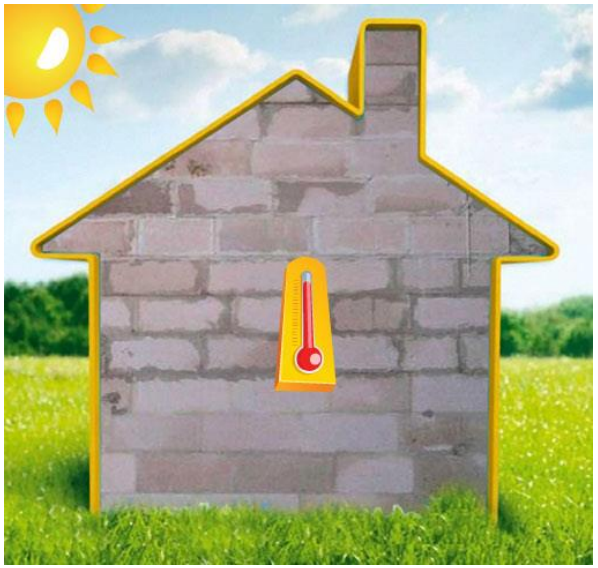


IMAGEN: googledrive.com

IMAGEN: viverosdelpozo.com

IMAGEN: mscjoseapacheco.files.wordpress.com

Errores sistemáticos versus aleatorios

- Los **errores sistemáticos** son aquellos errores que se producen de igual modo y en la misma dirección en todas las mediciones que se realizan de una magnitud.
- Puede estar originado en un defecto del instrumento o deberse a una particularidad del operador o del proceso de medición, etc.
- Este error no tiende a cero al aumentar el tamaño de la muestra, ya que está implícito en el diseño del estudio, pero puede prevenirse su aparición a través de un buen diseño del estudio y de la elección de instrumental.

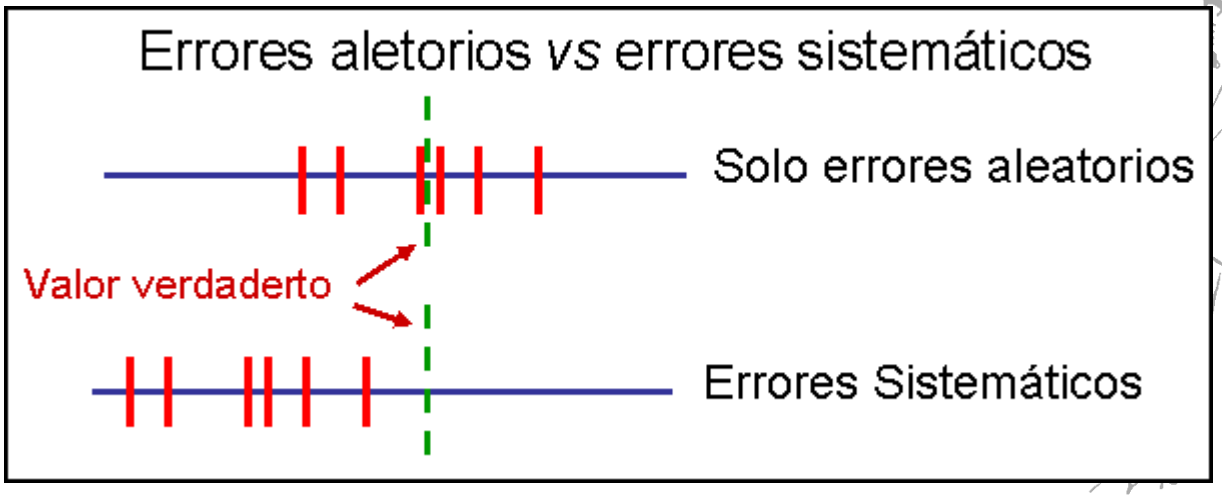


IMAGEN: virtual.unal.edu.co

Errores sistemáticos versus aleatorios

- **Ejemplos de errores sistemáticos** son:

  - Cuando una balanza está mal calibrada y no mide el cero apropiadamente.

  - Cuando un cronómetro no funciona bien y siempre se obtiene el mismo intervalo de tiempo mayor (o menor).

IMAGEN: femto.es

IMAGEN: mscjoseapacheco.files.wordpress.com

Precisión versus exactitud

- **Precisión** se refiere a la dispersión del conjunto de valores obtenidos de mediciones repetidas de una magnitud. Cuanto menor es la dispersión mayor la precisión. Una medida común de la variabilidad es la **desviación estándar** de las mediciones y la precisión se puede estimar como una función de ella. Es el grado de ausencia de error aleatorio.
- **Exactitud** se refiere a cómo de cerca del valor real se encuentra el valor medido.

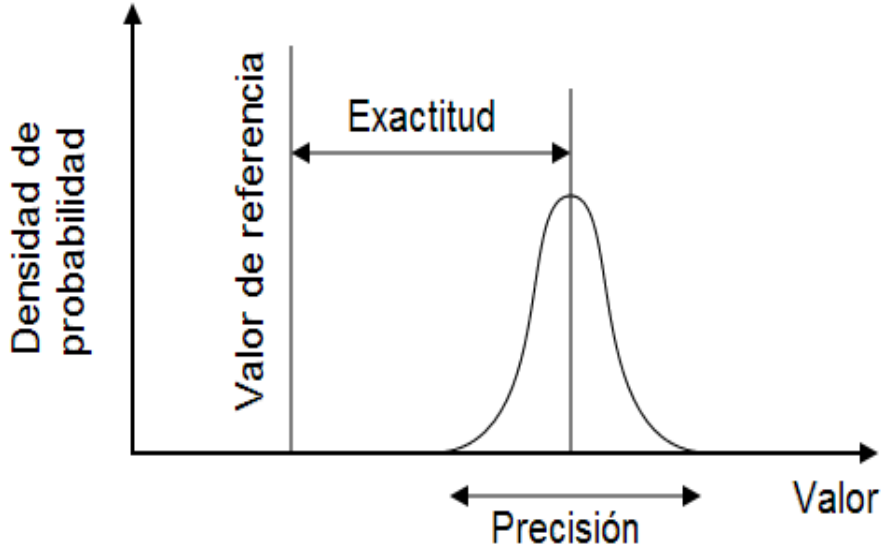


IMAGEN: es.wikipedia.org

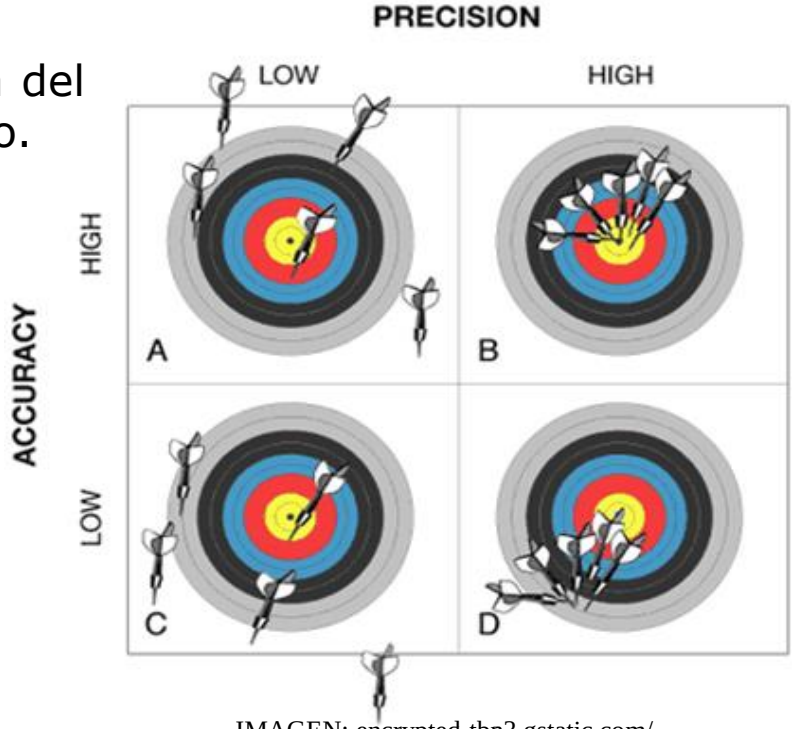


IMAGEN: encrypted-tbn3.gstatic.com/

Cuantificación = medición

- La toma de datos en la **muestra** de estudio debe ser medida cuidadosamente, utilizando el equipamiento del laboratorio (sensor de pH, balanza, pipeta, termómetro, etc.).

- El límite de medición del equipamiento usado añade cierta incertidumbre a los datos obtenidos.
- Todos los equipos tienen una cierta incertidumbre, por lo que en mediciones cuantitativas, **se debe indicar el nivel de incertidumbre del instrumento utilizado**, que dependerá de su grado de precisión.

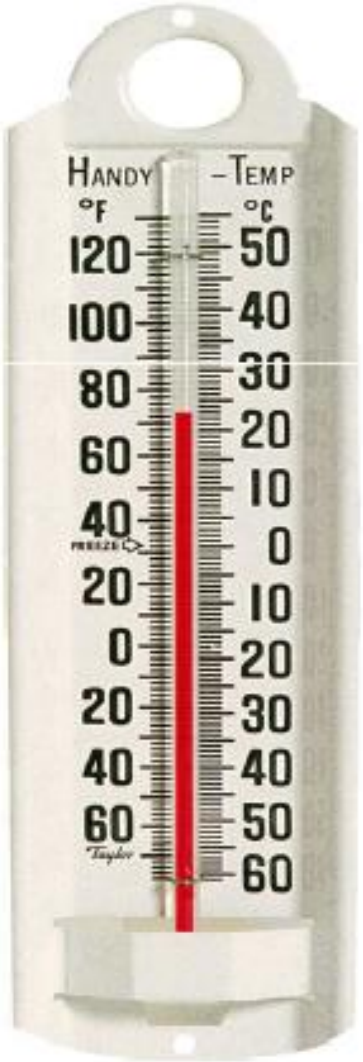
¿Cómo determinar la incertidumbre?

- Para los **grados de precisión**, la regla más sencilla es que el grado de precisión es más/menos (±) la división menor en el instrumento (división de lectura mínima). Esto se aplica a los **instrumentos con pantallas digitales y a las reglas**.

- El **límite de error** de un instrumento es con frecuencia la mitad del valor de la división menor, es decir, de su grado de precisión.

  Por ejemplo, si el valor de una bureta es 20,1 cm^3 se convierte en 20,10 cm^3 (±0,05 cm^3), ya que su lectura mínima es de 0,1 cm^3. Hay que tener en cuenta que el valor del volumen se menciona ahora con un decimal más para que haya coherencia con la incertidumbre.

¿Cómo determinar la incertidumbre?



- En la práctica, resulta imposible conocer el valor exacto de una magnitud medida, por lo que en lugar de dar un único valor numérico, se expresa como un intervalo (denominado **incertidumbre**) alrededor del valor resultante de la medida, que se entiende como el valor convencionalmente verdadero.
- Normalmente, **la incertidumbre viene indicada en el instrumento de medida** por el fabricante, pero **si no se especifica**, como regla general se usa +/- la mitad de la división menor de medida.
- Por ejemplo, en un termómetro de mercurio se suele leer hasta la mitad de la división de lectura mínima. Esto implica que si su lectura mínima es de 1 ºC y marca un valor de 23 ºC, se convierte en 23,0 ºC (±0,5 ºC).
- Hay que asegurarse de que el número de cifras significativas en la tabla de datos o gráfico refleje la la incertidumbre del instrumento utilizado.

Medida de la tendencia central

- **Moda**: Es el valor que aparece más frecuentemente. (Datos: 10, 11, 12, 12, 9, 8) Moda=12. Se usa con datos no numéricos.
- **Mediana**: Es el valor en el que la mitad de los datos son mayores que él, y la otra mitad son menores que él. Mediana=10.
- **Media**: El más usado para determinar la tendencia central de los datos, y es el promedio aritmético de los datos. Media= 10+11+12+12+9+8= 62/6 = 10.3

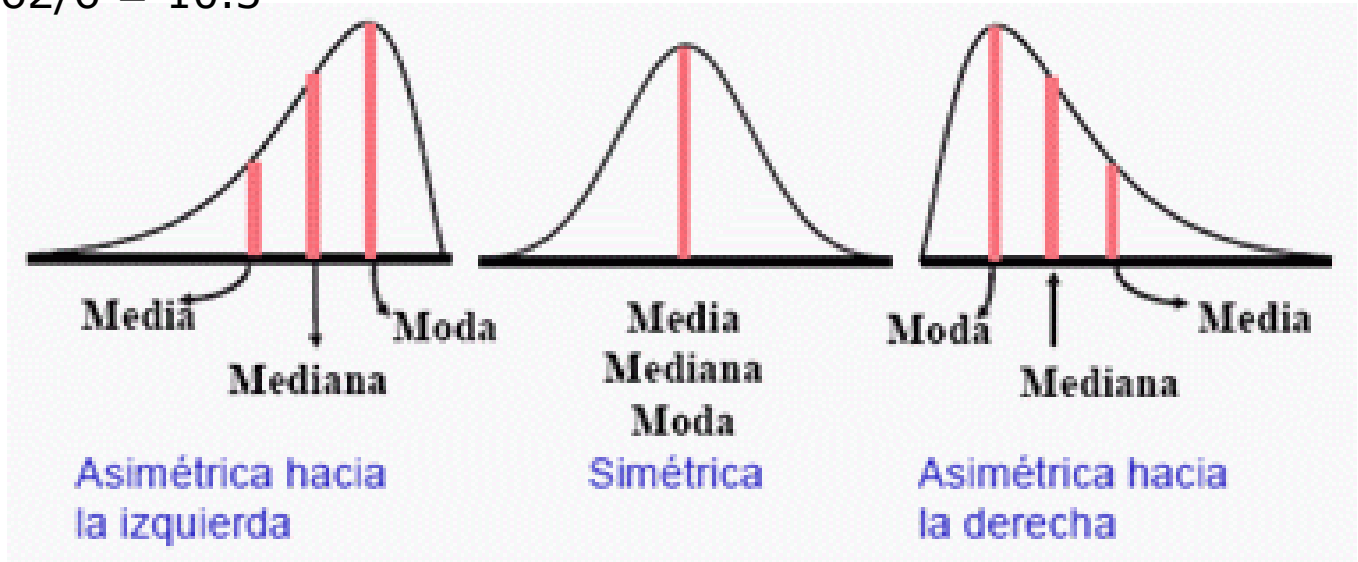
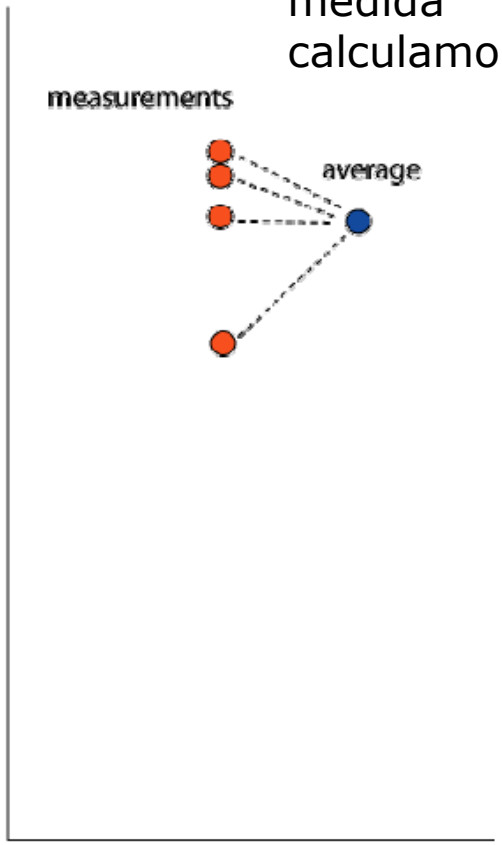


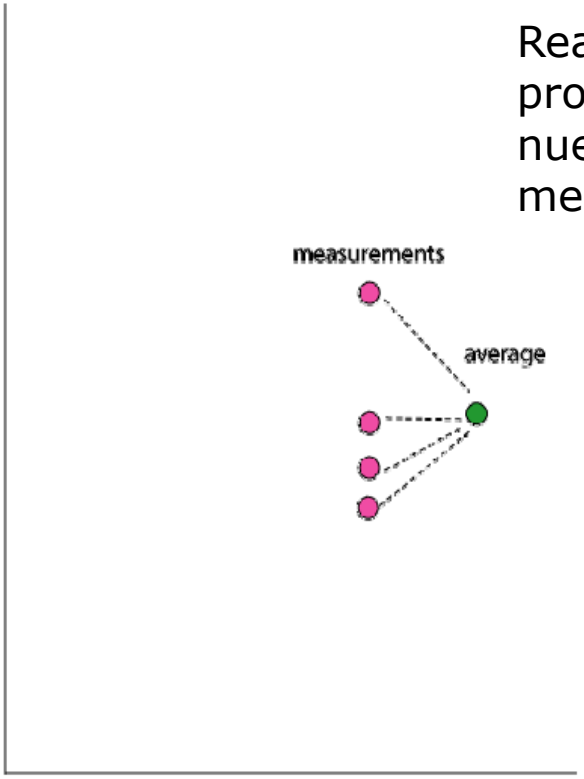
IMAGEN: www.monografias.com

Comparando medias

Cuando necesitamos medir con precisión algo, repetimos la medida varias veces y calculamos un valor medio.

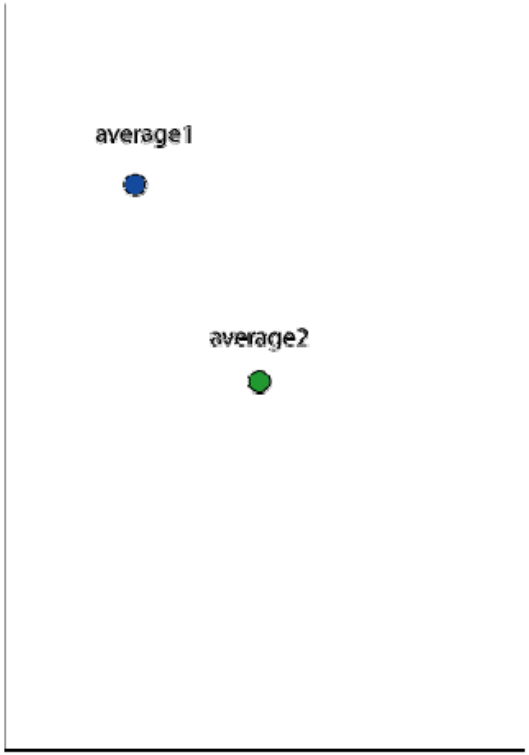


Realizamos el mismo procedimiento para nuestra próxima medida.



Comparando medias

- Una vez que se han obtenido las medias para cada conjunto de datos, los valores pueden ser presentados conjuntamente en una gráfica, para visualizar la relación entre ambos.

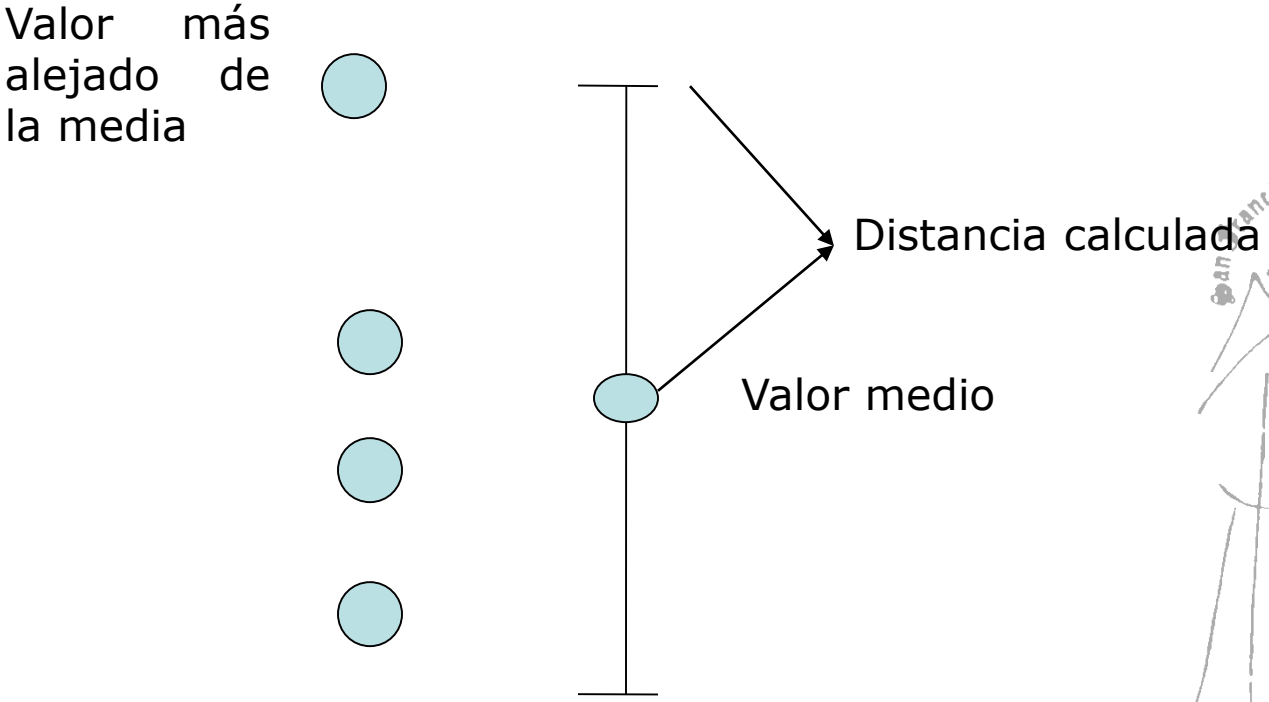


Pero, ¿cómo podemos estar seguros de que estos valores son realmente diferentes?

Tal vez, nuestras medidas no son muy precisas....

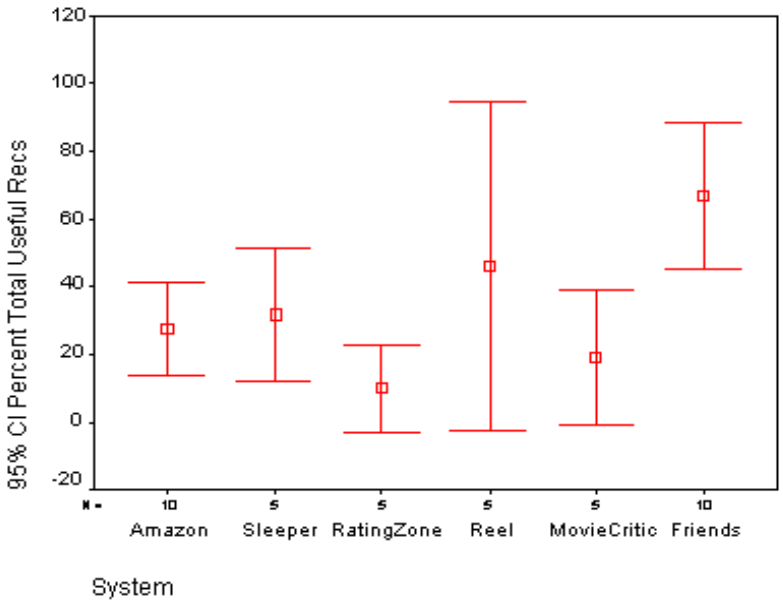
Representando barras de error

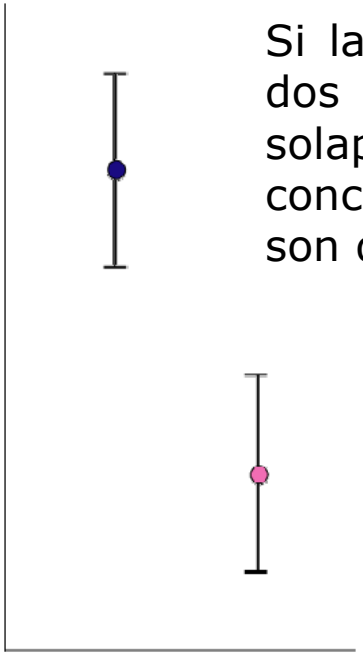
- Las barras de error muestran la distancia entre el valor medio y la medida más alejada de éste. Es decir, **representan la variabilidad de los datos** o la desviación típica de los mismos.



¿Qué sugieren las barras de error?

- Si las barras de error se solapan, es probable que no haya una diferencia significativa entre los valores medios medidos.





Si las barras de error para dos valores medios no se solapan, es justificable concluir que dichas medias son diferentes.

- **Las barras de error representan el valor calculado de la desviación estándar (SD).**

¿Cómo puede representarse gráficamente la longitud de las hojas?

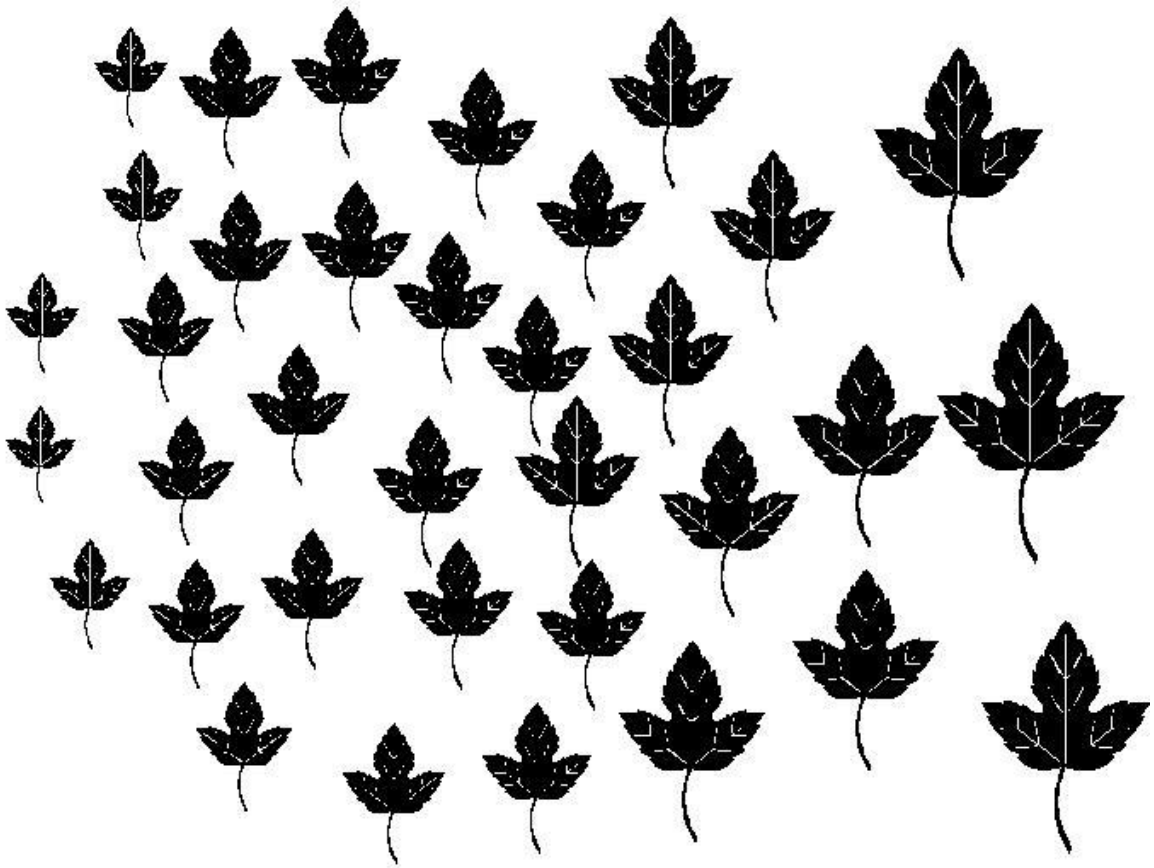


IMAGEN: ime.usp.br/~rvicente/IntroEstatistica.pdf

Distribución normal

- Simplemente midiendo la longitud de cada una y representando cuantas pertenencen a cada longitud.

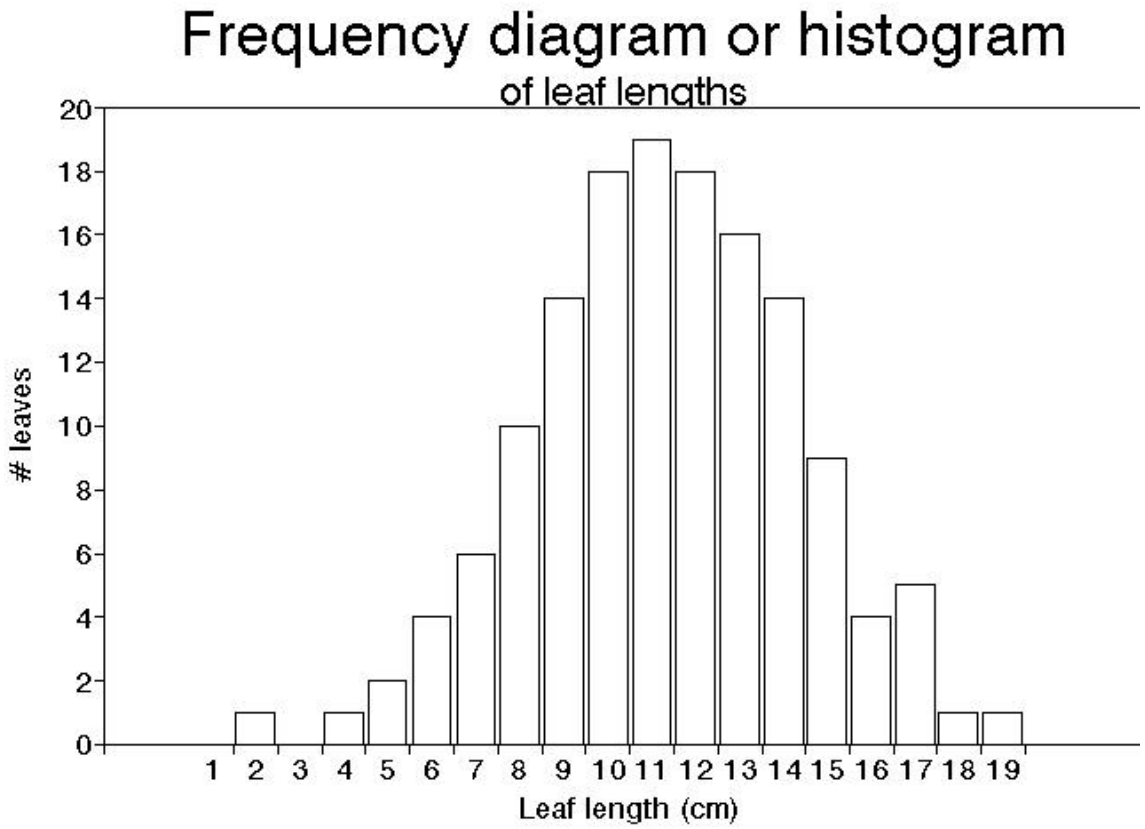
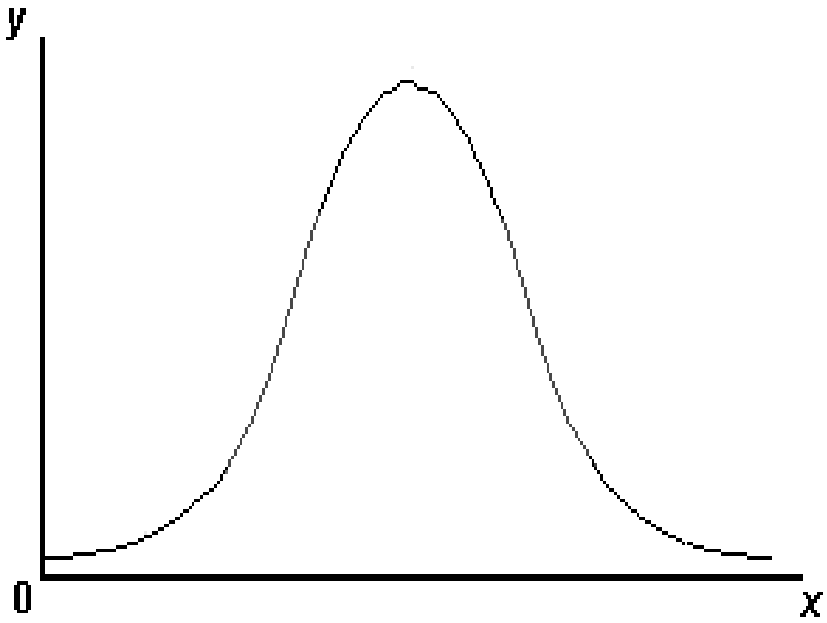


IMAGEN: ime.usp.br/~rvicente/IntroEstatistica.pdf

Distribución normal

- Si lo alisamos, el histograma adquiere esta forma de campana o curva de distribución gaussiana.

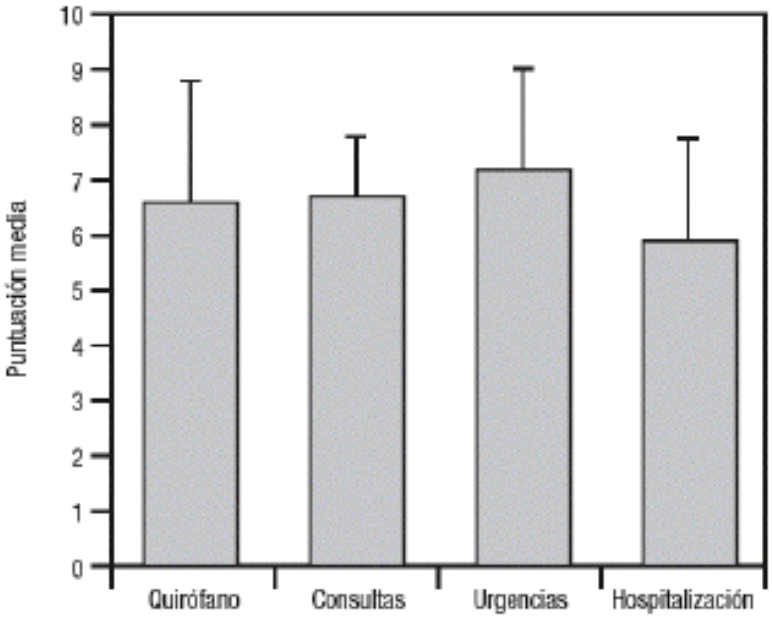


©Copyright Robert Niles, http://www.robertniles.com/stats/stdev.shtml

- Exencialmente, significa que en todos los estudios **con un número suficiente de datos medidos**, un número significativo de los resultados tenderán a estar cercanos a la media.

Desviación estándar (SD)

- Es una medida de cómo las observaciones individuales medidas de un conjunto de datos se distribuyen en torno a la media.
- **La SD** es un indicador más sofisticado **para indicar la precisión de un conjunto de datos medidos**, y puede usarse para **dibujar las barras de error en un gráfico**.



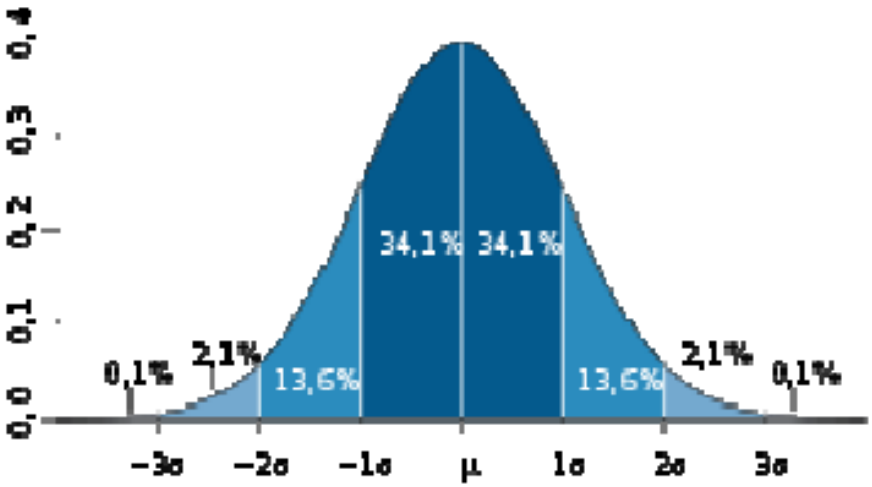
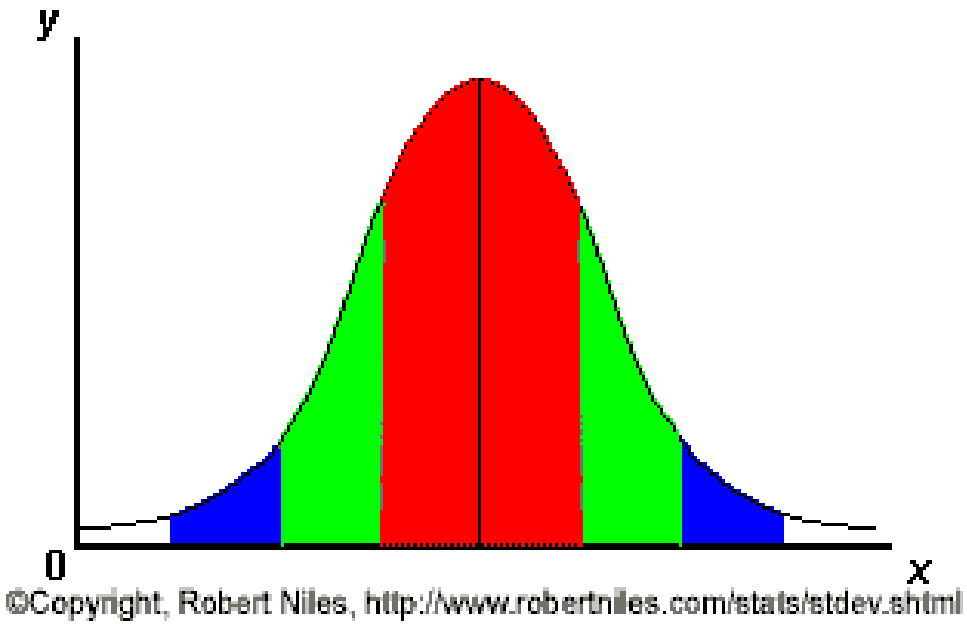
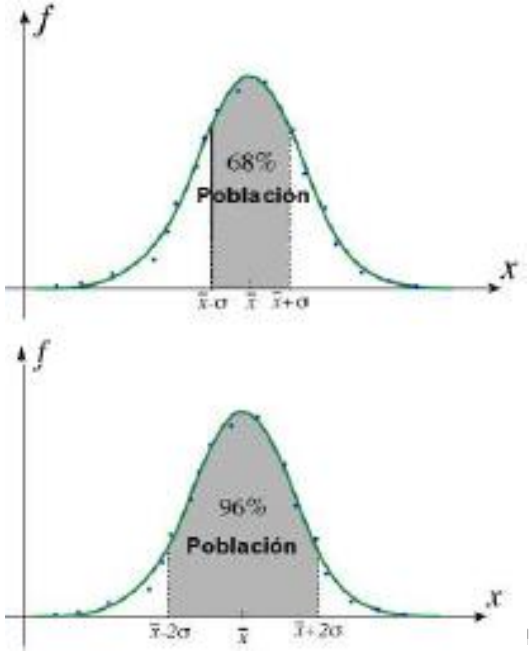


IMAGEN: images.cnitblog.com/

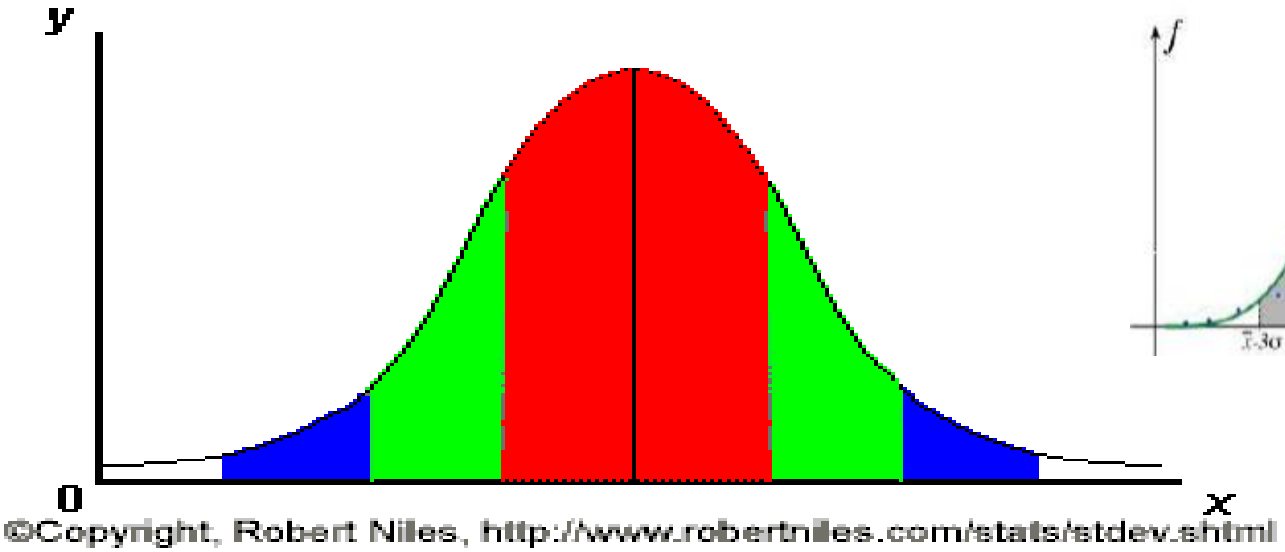
Curva de distribución típica

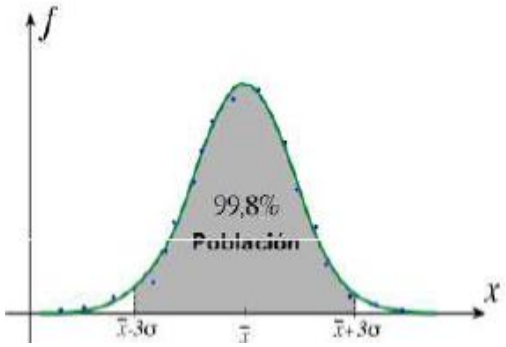




- En una distribución normal, en torno al 68% de todos los datos se encuentran en +/- 1 desviación típica de la media (área roja). Ej: Media=10.3 y SD=1.1; El 68% datos entre 9.2-11.4
- Así mismo, en una distribución normal, en torno al 95% de todos los datos se encuentran en +/- 2 desviaciones típicas de la media (área roja + verde). Ej: Media=10.3 y SD=1.1; El 95% datos entre 8.1-12.5

Curva de distribución típica





-3sd -2sd +/-1sd +2sd +3sd

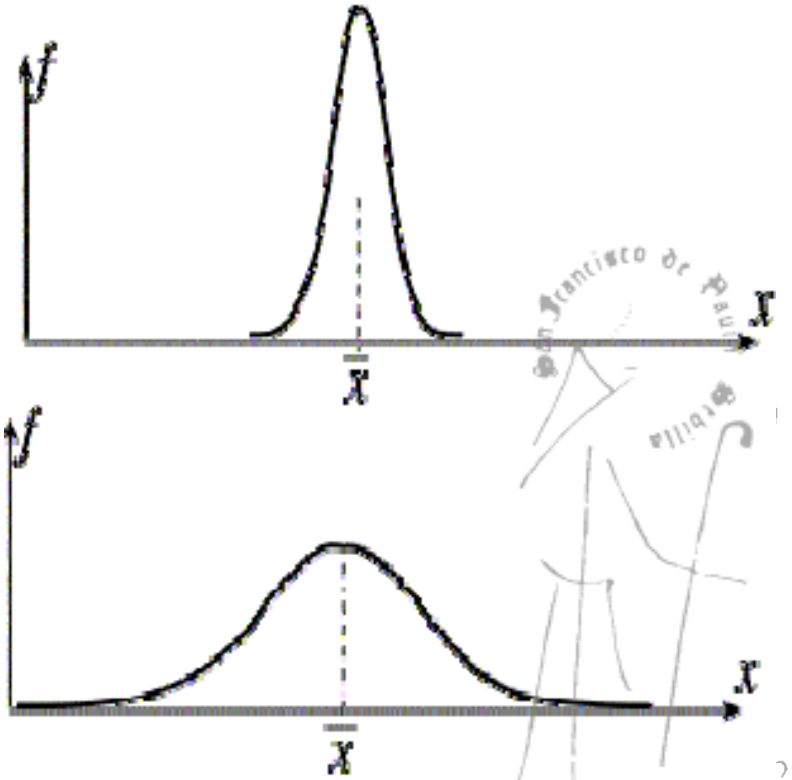
- En una distribución normal, en torno al 99% de todos los datos se encuentran en +/- 3 desviaciones típicas de la media (área roja + verde + azul). Ej: Media=10.3 y SD=1.1; El 99% datos entre 7-13.6

Resumiendo la SD

- La desviación típica es un parámetro estadístico que indica como de cerca se disponen todos los datos en torno a la media en un conjunto de mediciones. **La SD se considera pequeña si es menos del 33% de la media.**

- Cuando los datos se mantienen muy juntos y forman una buena curva en forma de campana, la desviación típica es pequeña (resultados precisos, pequeña SD).

- Cuando los datos se esparcen y forman una curva en forma de campana bastante aplanada, la desviación típica es grande (resultados menos precisos, grande SD).



Contraste de hipótesis

- Aunque el objetivo al llevar a cabo una investigación es, por lo general, conocer las características de los individuos de una **población**, en la práctica suele ser difícil, cuando no imposible.
  - Para obviar estos inconvenientes, se recurre al estudio de una **muestra,** a partir de la cual podemos **inferir o estimar** las características de la población entera de la cual aquella ha sido extraída (**estadística inferencial)**.

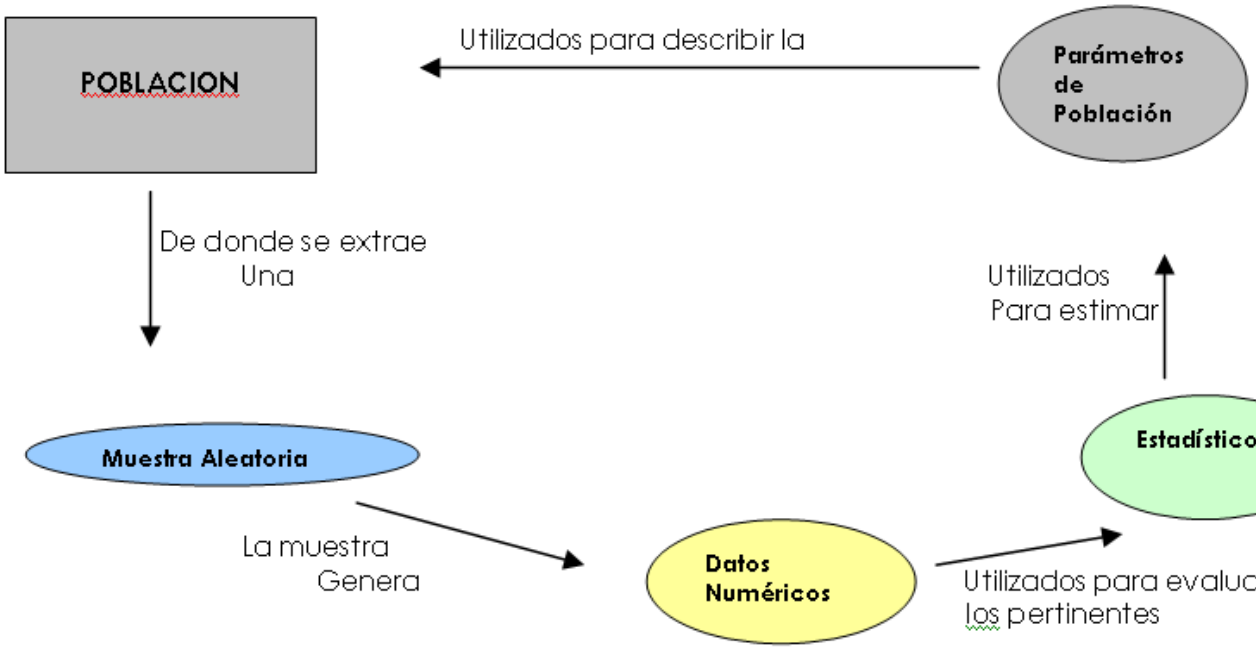


IMAGEN: virtual.unal.edu.co

Contraste de hipótesis

- La estadística inferencial comprueba hipótesis estadísticas, por lo que son diferentes de las hipótesis experimentales. Para entender su significado, veamos el siguiente ejemplo.
- **EJEMPLO**: Se realiza un experimento para comprobar el "efecto del nitrógeno sobre el crecimiento del maíz". La hipótesis experimental es que sí lo favorece, al ser un bioelemento necesario para síntesis de proteínas y ácidos nucleicos. Para comprobar esta hipótesis, se cultiva una parcela de plantas de maíz con nitrógeno y otra sin el mismo. Al cabo de un mes, se mide el ancho de la hoja de 10 plantas de cada parcela, obteniéndose una media de 56.2 cm y 48.5 cm, respectivamente.

¿Indica este resultado que hay una diferencia entre las dos poblaciones y que el nitrógeno promueve el crecimiento vegetal? ¿O se debe la diferencia observada entre ambas medias al simple azar?

IMAGEN: pastaoro.com.ve

- **Se require un test estadístico para discriminar entre ambas posibilidades.**

Contraste de hipótesis

- Tras realizar un experimento o investigación, hay que determinar si los resultados son **estadísticamente significativos** o no. En una investigación es erróneo contrastar datos sin una hipótesis previa.
- El **contraste de hipótesis** consiste en evaluar una población a través de una muestra aleatoria de la misma, con objeto de determinar si existen evidencias para que una hipótesis formulada sobre la población deba ser rechazada o no.

IMAGEN: 1.bp.blogspot.com

- El propósito de un contraste de hipótesis es rechazar la **hipótesis nula (Ho)**, aquella en la que no existe relación entre ambas variables de estudio o que dos medias no son diferentes.
- Únicamente rechazando la Ho podremos quedarnos con la **hipótesis alternativa (H1)** o hipótesis formulada por el investigador, aquella en la que sí existe una relación entre ambas variables de estudio o que dos medias sí son diferentes.

Contraste de hipótesis

- Para poder comprobar (rechazar) la hipótesis nula, hay que escoger un test estadístico concreto, como el **t-test** o el **test de chi-cuadrado**, ya que cada uno tiene una función diferente.
- El resultado de dicho test será un valor de **p** que indique la **probabilidad de que los resultados se deban al azar**, es decir, no sean estadísticamente significativos.
- El valor de p solo puede estar en 0 y 1. Cuanto menor sea el valor de p obtenido, mejor.
- La comunidad científica acepta un **nivel de significación estadística** o valor de **p ≤ 0.05** (5%) para rechazar la hipótesis nula, lo que indica que existe una probabilidad del 95% de que los resultados sean significativos y no debidos al azar.

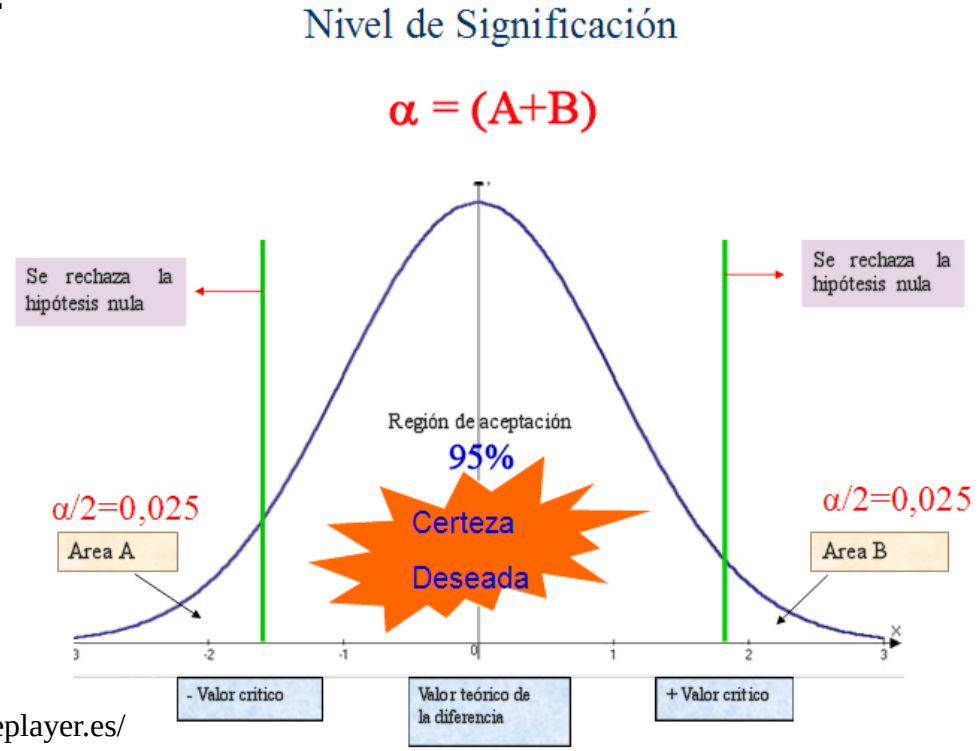


IMAGEN: images.slideplayer.es/

¿Cómo saber si la diferencia entre dos conjuntos de datos es significativa?

- En nuestro ejemplo del efecto del nitrógeno sobre el maíz, encontramos que existe una diferencia entre las medias de nuestras dos muetras. Sin embargo, necesitamos determinar si esta diferencia es significativa, es decir, si indica una **diferencia real en las medias** de las poblaciones de ambas áreas debido a la concentración de nitrógeno.
- Para determinar si la diferencia existente entre las medias de dos conjuntos de datos es significativa (real), se utiliza el **t-test** o **t-Student**.

IMAGEN: pastaoro.com.ve

t-test

Se aplica cuando los datos:

- Son independientes.
- Tienen una distribución normal (curva gausiana).
- Tienen un tamaño mínimo de muestra de 10 valores cada conjunto de datos.
- Tienen una variabilidad similar.
- El valor de *t* se calcula usando la fórmula:

$$t_{obt} = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Where:

t_{obt} = obtained t

$\overline{X}_1$ and $\overline{X}_2$ = means for the two groups

s_1^2 and s_2^2 = variances of the two groups

n_1 and n_2 = number of participants in each of the two groups

t-test

- Una vez calculado, dicho valor se busca en una tabla estandarizada de valores de *t*.
- En dicha tabla, hay dos columnas denominadas "grados de libertad" y "nivel de significación o probabilidad (valores de p)".
- Los grados de libertad se calculan como $(n_1 + n_2)-2$.
- El nivel de significación de referencia es $p \leq 0.05$.
- Si el valor de t obtenido (t_{obt}) > $t_{0.05}$, la hipótesis nula se rechaza.

	decreasing value of *p* ⟶			
Degrees of freedom (df)	*p* values			
	0.10	0.05	0.01	0.001
1	6.31	12.71	63.66	636.60
2	2.92	4.30	9.92	31.60
3	2.35	3.18	5.84	12.92
4	2.13	2.78	4.60	8.61
5	2.02	2.57	4.03	6.87
6	1.94	2.45	3.71	5.96
7	1.89	2.36	3.50	5.41
8	1.86	2.31	3.36	5.04
9	1.83	2.26	3.25	4.78
10	1.81	2.23	3.17	4.59
12	1.78	2.18	3.05	4.32
14	1.76	2.15	2.98	4.14
16	1.75	2.12	2.92	4.02
18	1.73	2.10	2.88	3.92
20	1.72	2.09	2.85	3.85
22	1.72	2.08	2.82	3.79
24	1.71	2.06	2.80	3.74
26	1.71	2.06	2.78	3.71
28	1.70	2.05	2.76	3.67
30	1.70	2.04	2.75	3.65
40	1.68	2.02	2.70	3.55
60	1.67	2.00	2.66	3.46
120	1.66	1.98	2.62	3.37
∞	1.64	1.96	2.58	3.29

t-Test: Ejemplo práctico

- **Ejemplo**: Un grupo de investigación ha estudiado el uso de hormonas como forma de aumentar la tasa de crecimiento en animales. Para ello, han inyectado la hormona tiroxina (afecta a la tasa metabólica) a 12 ratas, mientras que usaron otras 12 ratas como grupo control.
- Tras cuatro semanas, se volvieron a pesar las ratas y la diferencia en masa observada se recogió en la siguiente tabla:

Grupo experimental/ g ± 0,1 g					Grupo control/ g ± 0,1 g			
	3,1	3,0				3,0	2,7	
	4,1	4,0				2,6	3,0	
	4,0	3,8				2,7	2,8	
	3,5	2,9				2,4	2,5	
	3,5	2,9				2,4	2,5	
	3,0	4,2				2,9	3,0	

¿Cuál sería la hipótesis nula? ¿Puede ser rechazada?

t-Test: Ejemplo práctico

- Para calcular el t-test a partir de la fórmula, tan solo hay que completar la siguiente tabla usando Excel:

$$t_{obt} = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Media del grupo experimental (X_1)	3,5
Media del grupo control (X_2)	2,7
X_1-X_2	0,8
SD del grupo experimental (S_1)	0,5
SD del grupo control (S_2)	0,2
$(S_1)^2$	0,256
$(S_2)^2$	0,054
$(S_1)^2/n_1$	0,021
$(S_2)^2/n_2$	0,004
$(S_1)^2/n_1 + (S_2)^2/n_2$	0,026
$\sqrt{(S_1)^2/n_1 + (S_2)^2/n_2}$	0,16
$t_{obtenido}$	5,22

t-Test: Ejemplo práctico

	decreasing value of p ⟶			
Degrees of freedom (df)	p values			
	0.10	0.05	0.01	0.001
1	6.31	12.71	63.66	636.60
2	2.92	4.30	9.92	31.60
3	2.35	3.18	5.84	12.92
4	2.13	2.78	4.60	8.61
5	2.02	2.57	4.03	6.87
6	1.94	2.45	3.71	5.96
7	1.89	2.36	3.50	5.41
8	1.86	2.31	3.36	5.04
9	1.83	2.26	3.25	4.78
10	1.81	2.23	3.17	4.59
12	1.78	2.18	3.05	4.32
14	1.76	2.15	2.98	4.14
16	1.75	2.12	2.92	4.02
18	1.73	2.10	2.88	3.92
20	1.72	2.09	2.85	3.85
		*		
22	1.72	2.08	2.82	3.79
24	1.71	2.06	2.80	3.74
26	1.71	2.06	2.78	3.71
28	1.70	2.05	2.76	3.67
30	1.70	2.04	2.75	3.65
40	1.68	2.02	2.70	3.55
60	1.67	2.00	2.66	3.46
120	1.66	1.98	2.62	3.37
∞	1.64	1.96	2.58	3.29

- Grados de libertad: (12 + 12) - 2 = 22
- Para un grado de libertad de 22, tenemos que el valor de t crítico es 2.08 (p=0.05) o mayor.
- Como el valor calculado de t_{obt} (5.22) es mayor que el valor crítico para $t_{0.05}$ (2.08), concretamente en torno al valor de $t_{0.001}$, podemos concluir que existe una probabilidad >99% de que la diferencia entre ambas medias sea significativa.
- Por tanto, podemos **rechazar la hipótesis nula**, es decir, que las hormona tiroxina no tiene efecto sobre el crecimiento de las ratas.

t-Test: Ejercicio

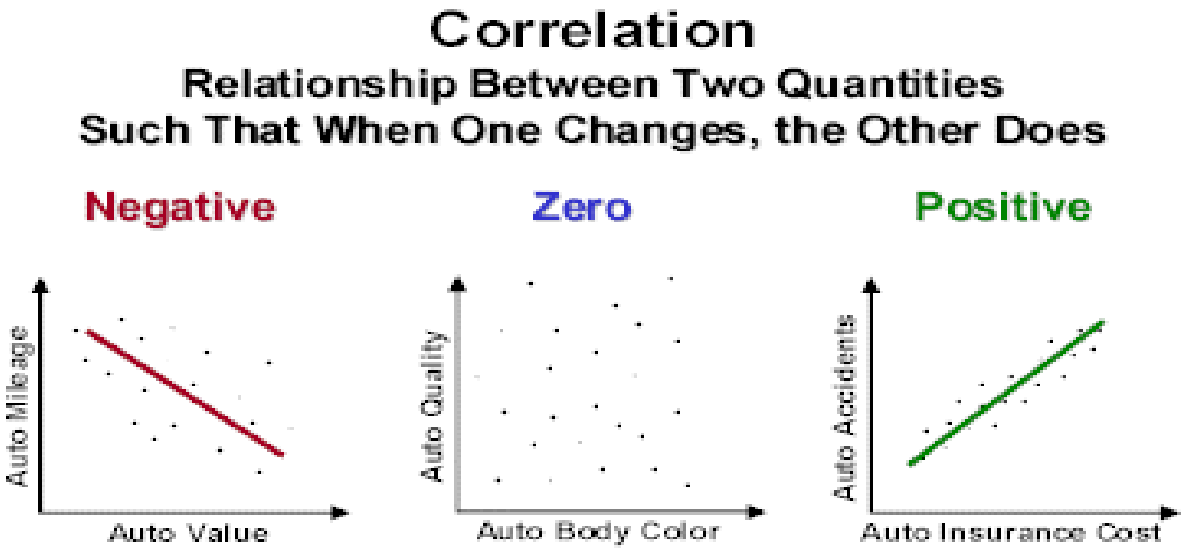
- Se ha investigado si la longitud de la concha de los percebes pudiera estar condicionada por la cercanía de estos al mar. Para ello, se comparó la longitud de la concha de dos grupos de percebes, un grupo 1 que vive más cerca del agua que otro grupo 2. Se ha medido la longitud de 15 percebes de cada grupo, cuyos resultados se encuentran a continuación:

Grupo 1/ cm ± 0,1 cm					Grupo 2/ cm ± 0,1 cm				
4,0	3,5	4,0			3,0	3,7	3,7		
4,0	3,9	3,5			3,8	3,0	3,8		
4,1	4,1	3,8			3,7	3,8	3,0		
3,8	4,0	4,0			3,6	3,9	3,0		
4,0	4,0	3,1			3,5	3,5	3,0		

¿Cuál sería la hipótesis nula en esta investigación? ¿Puede ser rechazada?

Correlaciones

- Frecuentemente parece que dos variables pudieran estar ligadas. Por ejemplo, la intensidad de luz y la densidad de vegetación.
- Dos conjuntos de datos pueden ser respresentados gráficamente y hallada la línea de mejor ajuste. Esta línea puede mostrar que existe una correlación positiva, negativa o no haberla.



- Existe un test estadístico para la correlación, el **coeficiente de correlación de Spearman (r_s)**, que varía de +1 (correlación perfecta) pasando por 0 (no correlación) hasta –1 (perfecta correlación negativa).

Test de correlación de Spearman

Se aplica cuando las variables:

- Son aleatorias.
- Son continuas.
- El valor de r_s se calcula usando la fórmula:

$$r_s = 1 - \frac{6\sum d^2}{n(n^2-1)}$$

r_s = Coeficiente de correlación por rangos de Spearman

d = Diferencia entre los rangos (X menos Y)

n = Número de datos

Test de correlación de Spearman

- Una vez calculado, dicho valor se busca en una tabla estandarizada de valores de r_s, siendo n el número de datos.
- Como siempre, el nivel de significación de referencia es $p \leq 0.05$, de manera que si el valor de $rs_{obtenido} > rs_{0.05}$, la hipótesis nula se rechaza.

TABLE D. **Values of Spearman r_s for the .05 and .01 Levels of Significance**

N	.05	.01	*N*	.05	.01
6	.886	—	19	.462	.608
7	.786	—	20	.450	.591
8	.738	.881	21	.438	.576
9	.683	.833	22	.428	.562
10	.648	.818	23	.418	.549
11	.623	.794	24	.409	.537
12	.591	.780	25	.400	.526
13	.566	.745	26	.392	.515
14	.545	.716	27	.385	.505
15	.525	.689	28	.377	.496
16	.507	.666	29	.370	.487
17	.490	.645	30	.364	.478
18	.476	.625			

From E. G. Olds, Distribution of sums of squares of rank differences for small numbers of individuals, *Annals of Mathematical Statistics 9*: 133–48 (1938), and E. G. Olds, The 5% significance levels for sums of squares of rank differences and a correction, *Annals of Mathematical Statistics 20*: 117–18 (1949). Copyright 1938 and Copyright 1949 by the Institute of Mathematical Statistics, Hayward, Calif. Reprinted by permission of the publisher.

Test de correlación de Spearman: Ejemplo

- ¿Existe una correlación entre el número de árboles que hay en una determinada zona y el número de animales que pasan por ella?

Zona	Nº animales	rango	Nº árboles	rango	d	d^2
1	40		8			
2	8		2			
3	25		5			
4	60		15			
5	12		7			
6	18		3			
7	19		4			
8	27		8			
9	24		7			
10	21		6			
11	64		19			
12	70		22			

Test de correlación de Spearman: Ejemplo

- El rango se determina de mayor a menor, siendo el 1º la zona con más animales o árboles, y el 12º la que menos, pudiendo o no coincidir.

Zona	Nº animales	rango	Nº árboles	rango	d	d^2
1	40	4	8	4,5	-0,5	0,25
2	8	12	2	12	0	0
3	25	6	5	9	-3	9
4	60	3	15	3	0	0
5	12	11	7	6,5	4,5	20,25
6	18	10	3	11	-1	1
7	19	9	4	10	-1	1
8	27	5	8	4,5	0,5	0,25
9	24	7	7	6,5	0,5	0,25
10	21	8	6	8	0	0
11	64	2	19	2	0	0
12	70	1	22	1	0	0

Test de correlación de Spearman: Ejemplo

- Sustituyendo en la fórmula:

$6\Sigma d^2$	192
$n(n^2-1)$	1716
$6\Sigma d^2/n(n^2-1)$	0,111
r_s	0,888

$$r_s = 1 - \frac{6\sum d^2}{n(n^2-1)}$$

- Para n=12, tenemos que el valor de rs crítico es 0.591 (p=0.05) o mayor.
- Como el valor calculado de rs (0.888) es mayor que ese valor crítico y de p = 0.01, podemos rechazar la hipótesis nula, ya que existe una probabilidad >99% de que la correlación entre ambas variables sea estadísticamente significativa.

TABLE D. Values of Spearman r_s for the .05 and .01 Levels of Significance

N	.05	.01	N	.05	.01
6	.886	—	19	.462	.608
7	.786	—	20	.450	.591
8	.738	.881	21	.438	.576
9	.683	.833	22	.428	.562
10	.648	.818	23	.418	.549
11	.623	.794	24	.409	.537
12	.591	.780	25	.400	.526
13	.566	.745	26	.392	.515
14	.545	.716	27	.385	.505
15	.525	.689	28	.377	.496
16	.507	.666	29	.370	.487
17	.490	.645	30	.364	.478
18	.476	.625			

Correlaciones

- Otra forma de comprobar cómo de buena es la correlación entre dos variables es el **coeficiente de determinación (R^2)**, que mide la bondad del ajuste realizado y que permite decidir si el ajuste lineal es suficiente o se deben buscar modelos alternativos.
- También se le denomina **bondad del ajuste**, ya que es una medida de la proximidad o de ajuste de la recta de regresión a la nube de puntos.
- Varía de 0 (el modelo no explica nada de la variable *y* a partir de la variable *x*) a 1 (ajuste perfecto donde una variable depende funcionalmente de la otra.

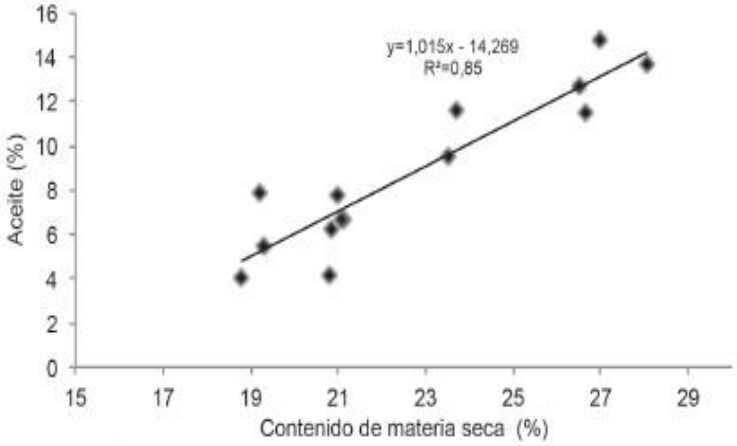


Fig. 1. Contenido de aceite (%) y materia seca (%) en aguacate Hass cosechado a partir de la semana 31 después del cuaje del fruto.
Medias con letras iguales no difieren entre sí según prueba DMS al 5%.

- La R^2 del gráfico indica que el 85% de la variabilidad en el contenido en aceite del aguacate puede atribuirse a una relación lineal con su contenido en materia seca.

Correlación vs relación causal

- Es importante resaltar que aunque el análisis estadístico de los datos indique una correlación entre la variable independiente y la dependiente, esto no prueba que exista una relación causal. Solo posteriores investigaciones revelarán el efecto causal entre las dos variables.
- **La correlación no implica relación causal**. Ejemplo: Las ventas de helados y el número de ataques por tiburones a los bañistas están correlacionados.
- Claramente se observa que no hay una interacción real entre los factores implicados, solo una simple coincidencia de los datos.
- Una vez que una correlación entre dos factores ha sido establecida a partir de los datos experimentales, es necesario avanzar en la investigación para determinar cual pudiera ser la relación causal entre ambos.