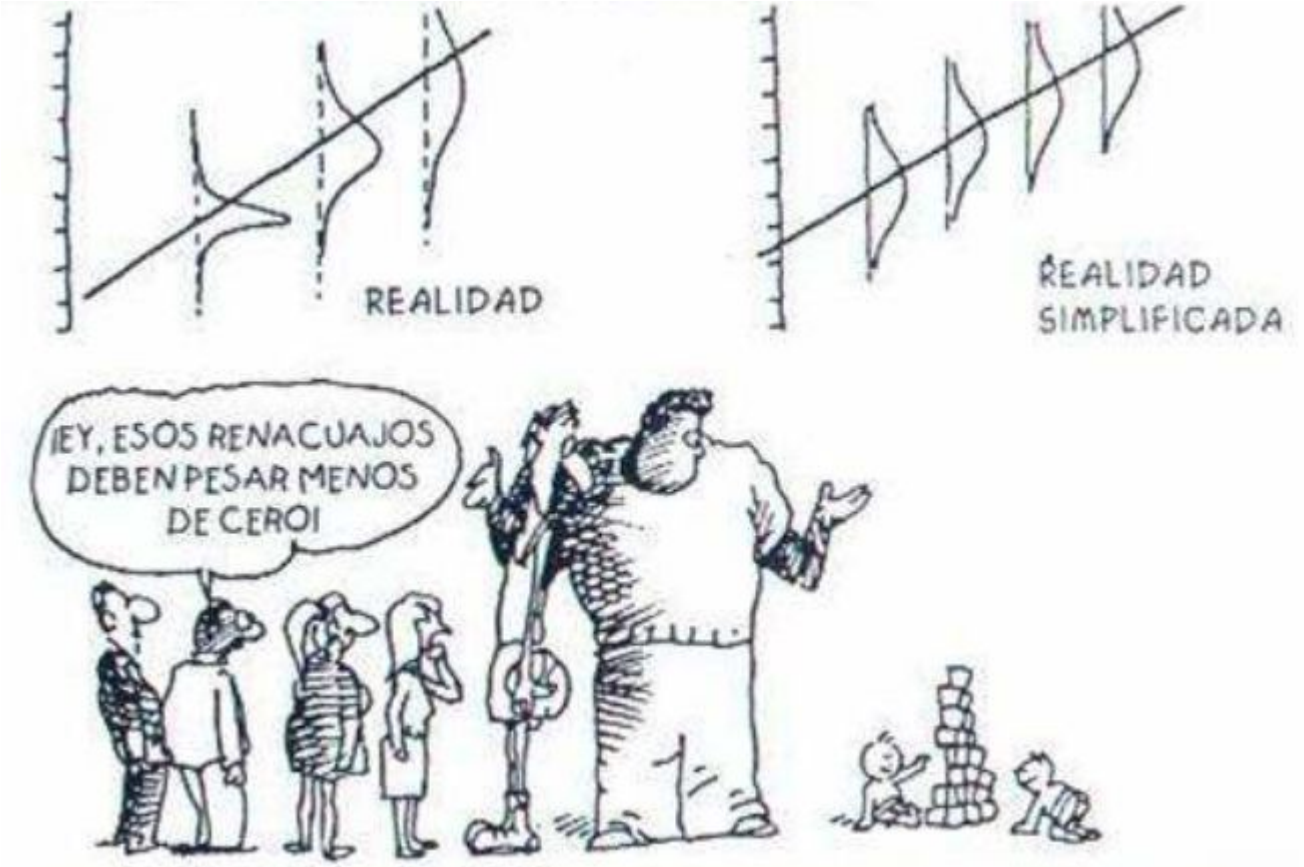
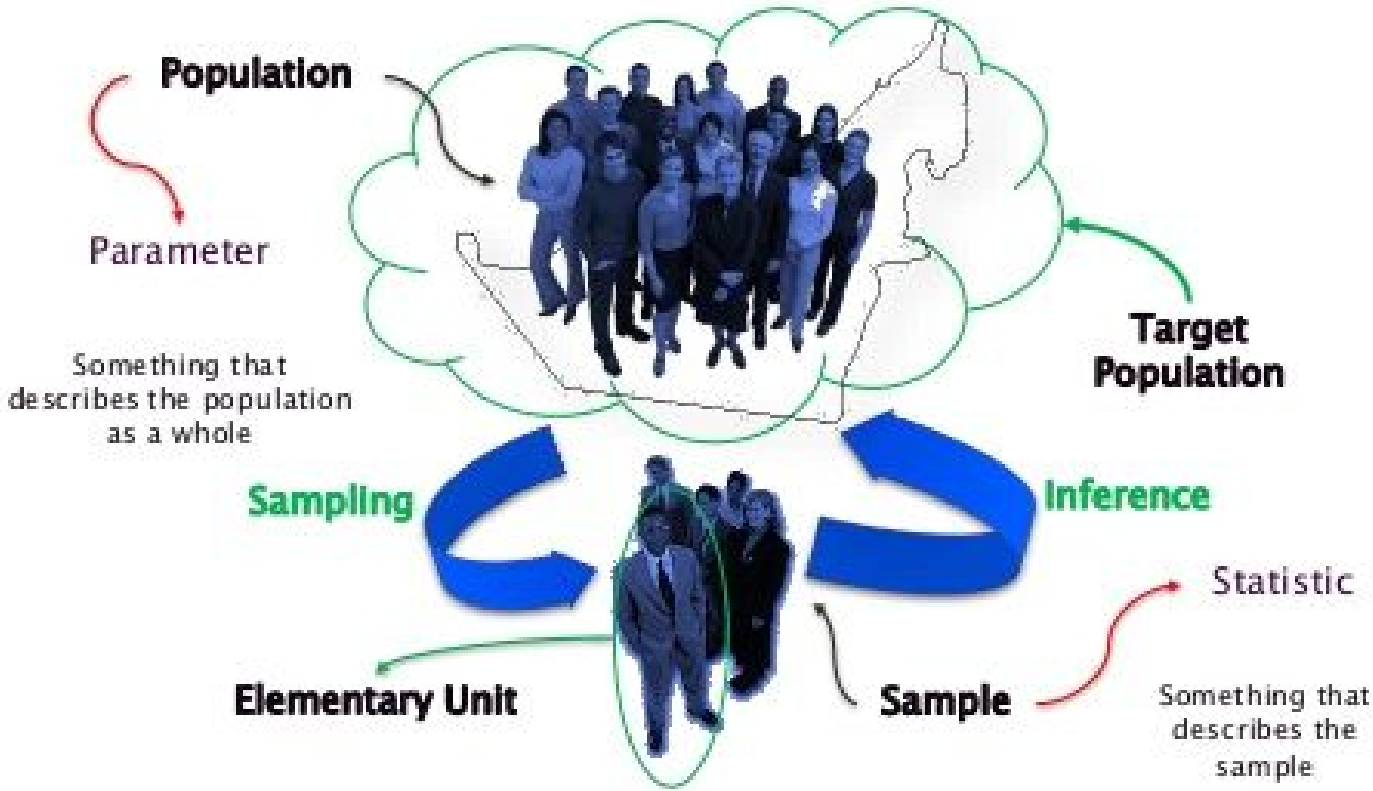
Estadística

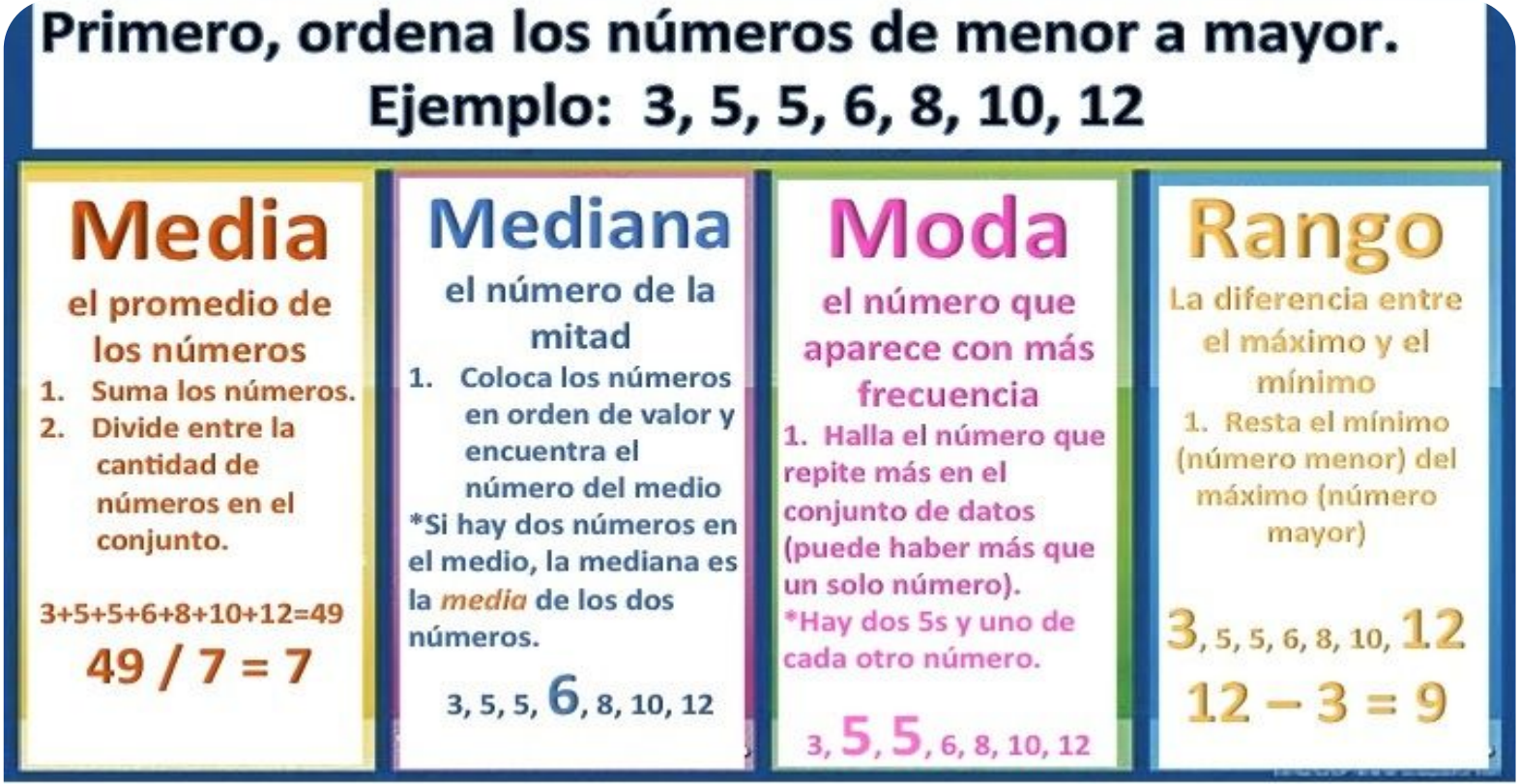
Conceptos generales



Gonick, L; Smith, W. 1993. La Estadistica en Comic.

POPULATION VS. SAMPLING





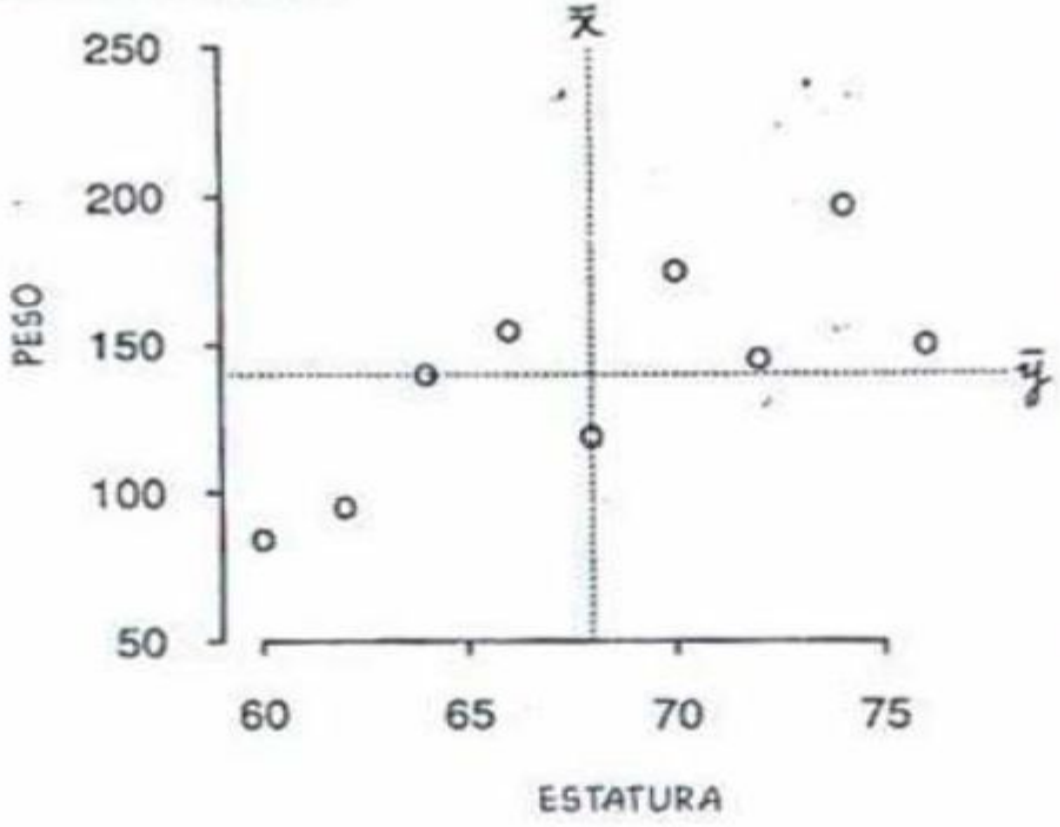
Media - Mediana - Moda - Rango

Población		Muestra
$\mu = \frac{\sum X}{N}$	Media	$\overline{X} = \frac{\sum X}{n}$
$\sigma^2 = \frac{\sum (X-\mu)^2}{N}$	Varianza	$\hat{\sigma}^2 = \frac{\sum (x-\bar{x})^2}{n-1}$
$\sigma = \sqrt{\frac{\sum (X-\mu)^2}{N}}$	Desviación estándar	$\hat{\sigma} = \sqrt{\frac{\sum (x-\bar{x})^2}{n-1}}$
$C.V. = \frac{\sigma}{\mu} * 100$	Coeficiente de variación	$C.V. = \frac{\hat{\sigma}}{\bar{x}} * 100$

Medidas de Variabilidad

PARA ILUSTRAR EL EJEMPLO DE AJUSTE DE LA RECTA, UTILIZAREMOS UN CONJUNTO MÁS REDUCIDO DE DATOS FICTICIOS CON SÓLO NUEVE PAREJAS DE PESOS Y ESTATURAS DE ESTUDIANTES.

ESTATURA	PESO
60	84
62	95
64	140
66	155
68	119
70	175
72	145
74	197
76	150



LA IDEA CONSISTE EN **MINIMIZAR** LA DISTANCIA TOTAL DE LOS VALORES Y A LA RECTA. IGUAL QUE CUANDO DEFINÍAMOS LA VARIANZA, BUSCAMOS LAS **DISTANCIAS AL CUADRADO** DE y CON LA RECTA Y LAS SUMAMOS PARA OBTENER LA **SUMA DE LOS ERRORES CUADRÁTICOS** (*SSE*):

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

ES UNA MEDIDA AGREGADA DE CUÁNTO PUEDEN DIFERIR LAS "PREDICCIONES y_i", LLAMADAS $\hat{y}_i$, CON RESPECTO A LOS VALORES REALES y_i.

La recta de **regresión** o recta de **mínimos cuadrados**

ES LA RECTA CON LA MÍNIMA *SSE*

$$y = a + bx$$

DONDE

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

Y

$$a = \bar{y} - b\bar{x}$$

AQUÍ $\bar{x}$ E $\bar{y}$ SON LAS MEDIAS DE $\{x_i\}$ Y $\{y_i\}$ RESPECTIVAMENTE.

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

$$SS_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

LA SUMA DE LOS CUADRADOS ALREDEDOR DE LA MEDIA MIDE LA DISPERSIÓN DE x_i Y DE y_i.

$$SS_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

EL PRODUCTO CRUZADO DETERMINA (CON SS_{xx}) EL COEFICIENTE b.

ESTE ES EL CÁLCULO TOTAL DE LOS VALORES FICTICIOS:

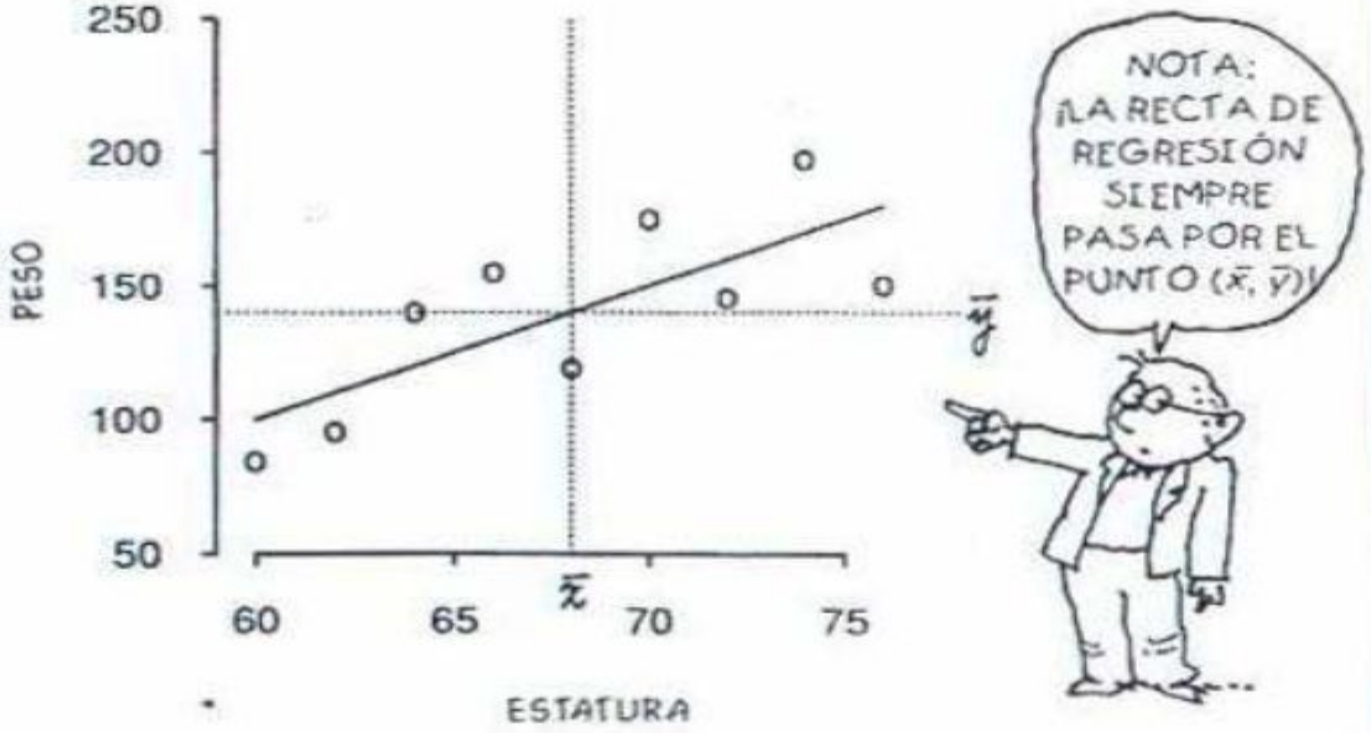
x_i	y_i	$(x_i-\bar{x})$	$(y_i-\bar{y})$	$(x_i-\bar{x})^2$	$(y_i-\bar{y})^2$	$(x_i-\bar{x})(y_i-\bar{y})$
60	84	-8	-56	64	3136	448
62	95	-6	-45	36	2025	270
64	140	-4	0	16	0	0
66	155	-2	15	4	225	-30
68	119	0	-21	0	441	0
70	175	2	35	4	1225	70
72	145	4	5	16	25	20
74	197	6	57	36	3249	342
76	150	8	10	64	100	80
SUMA = 612	1.260			$SS_{xx} = 240$	$SS_{yy} = 10.426$	$SS_{xy} = 1.200$

$\bar{x} = 68$ $\bar{y} = 140$

LO CUAL NOS DA VALORES PARA a Y b:

$$b = \frac{1.200}{240} = 5 \qquad a = \bar{y} - b\bar{x} = 140 - 5(68) = -200$$

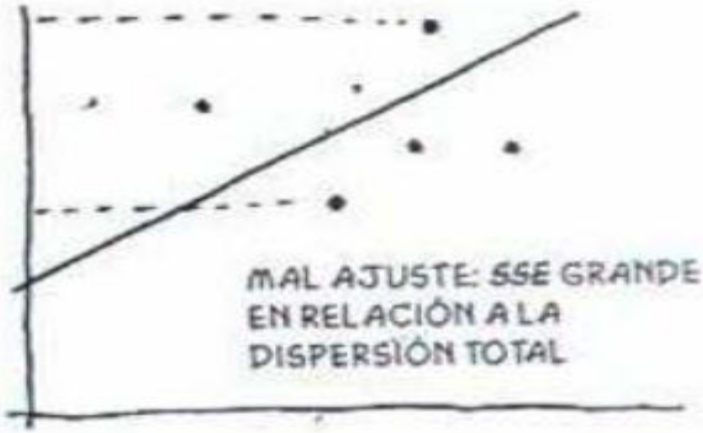
ENTONCES $y = -200 + 5x$



ANOVA

¡COMO HABÍAMOS PROMETIDO, ¡O AMENAZADO!) AHORA NOS PREGUNTAMOS SI ESTE ES EL MEJOR AJUSTE: ¿ES MUY BUENO?


BUEN AJUSTE: SSE PEQUEÑA, AL COMPARARLA CON LA DISPERSIÓN TOTAL


MAL AJUSTE: SSE GRANDE EN RELACIÓN A LA DISPERSIÓN TOTAL


BUEN AJUSTE: SSE MODERADA, PERO DISPERSIÓN TOTAL GRANDE


MAL AJUSTE: SSE GRANDE, INCLUSO EN COMPARACIÓN CON UNA DISPERSIÓN TOTAL GRANDE

VAMOS A CUANTIFICAR ESTO DESGLOSANDO LA VARIABILIDAD DE y. SEGUIREMOS COMO GUÍA EL DIBUJO DE LA DERECHA. TENEMOS

$$\hat{y}_i = a + bx_i$$

ENTONCES, $\hat{y}_i$ SON LOS PESOS PREDICHOS POR LA RECTA DE REGRESIÓN.

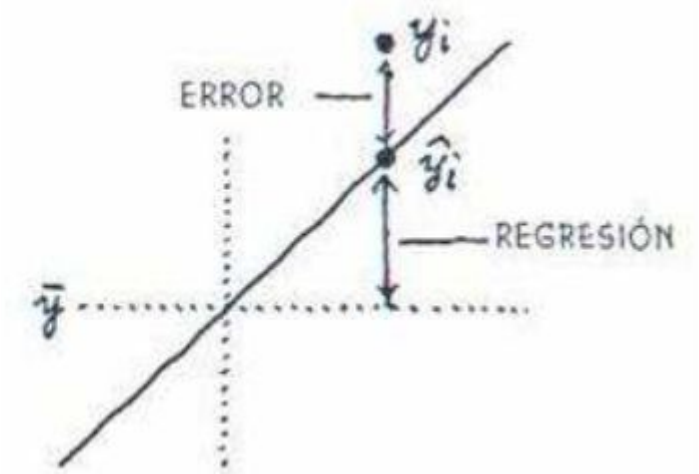


Tabla ANOVA

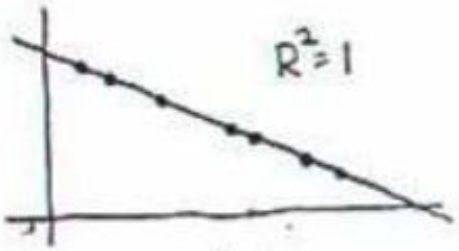
FUENTE DE VARIABILIDAD	SUMA DE CUADRADOS
REGRESIÓN	$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$
ERROR	$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$
TOTAL	$SS_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2$

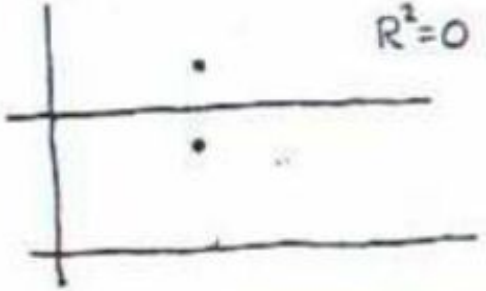
El coeficiente de determinación

ES LA PROPORCIÓN DE TODAS LAS SS_{yy} EXPLICABLES POR LA REGRESIÓN:

$$R^2 = \frac{SSR}{SS_{yy}} = 1 - \frac{SSE}{SS_{yy}}$$

(PORQUE $SSR = SS_{yy} - SSE$), R^2 ES SIEMPRE MENOR QUE 1. CUANTO MÁS SE APROXIMA A 1, MÁS PRECISO ES EL AJUSTE DE LA CURVA. $R^2 = 1$ CORRESPONDE AL AJUSTE PERFECTO.





Coeficiente de Determinación (R^2)

POR OTRA PARTE, TENEMOS EL

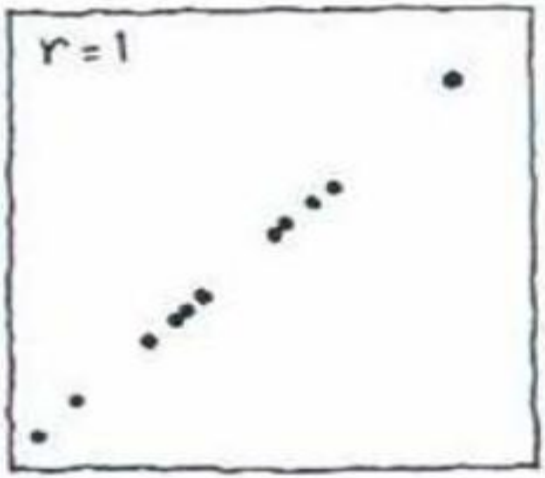
coeficiente de correlación

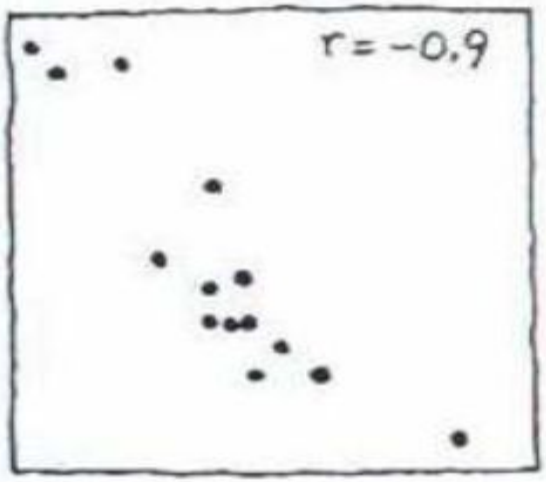
QUE ES LA RAÍZ CUADRADA DE R^2 CON EL SIGNO DE b.

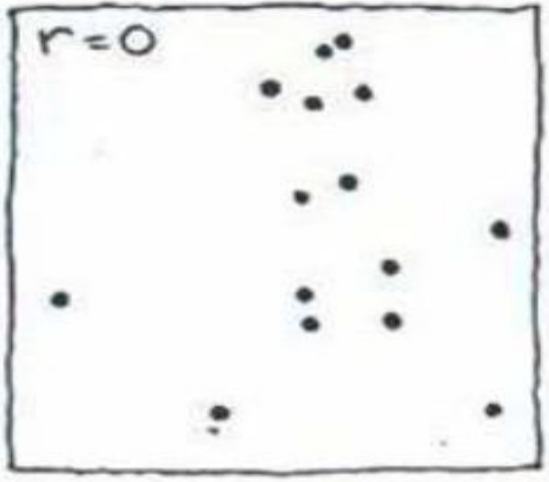
$$r = (\text{SIGNO DE } b)\sqrt{R^2}$$

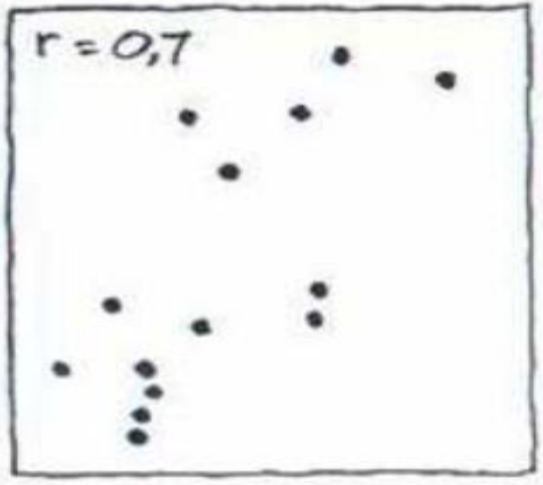
ENTONCES, r ES POSITIVA SI LA RECTA ES ASCENDENTE HACIA LA DERECHA, Y NEGATIVA SI LA RECTA TIENE FORMA DESCENDENTE HACIA LA DERECHA.

Coeficiente de Correlación (r)

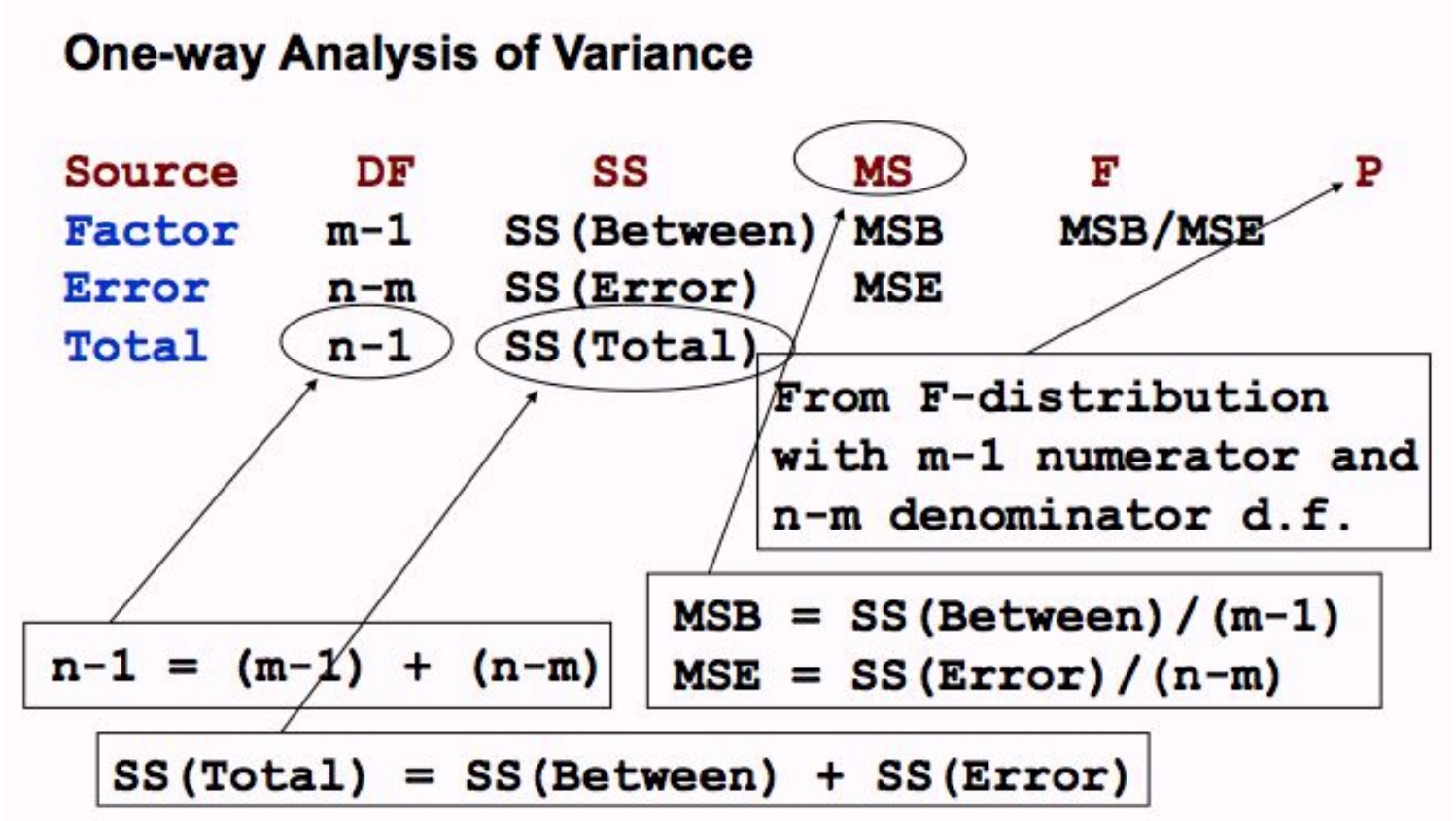








PERO SEAMOS SINCEROS: NADIE (BUENO, CASI NADIE) HACE YA ESTOS CÁLCULOS A MANO. CON EL ORDENADOR TODO ESTE TRABAJO PUEDE REALIZARSE ESCRIBIENDO UNA SOLA LÍNEA DE CÓDIGO...


One-way Analysis of Variance
Source DF SS MS F P
Factor m-1 SS(Between) MSB MSB/MSE
Error n-m SS(Error) MSE
Total n-1 SS(Total)
From F-distribution with m-1 numerator and n-m denominator d.f.
MSB = SS(Between)/(m-1)
MSE = SS(Error)/(n-m)
n-1 = (m-1) + (n-m)
SS(Total) = SS(Between) + SS(Error)

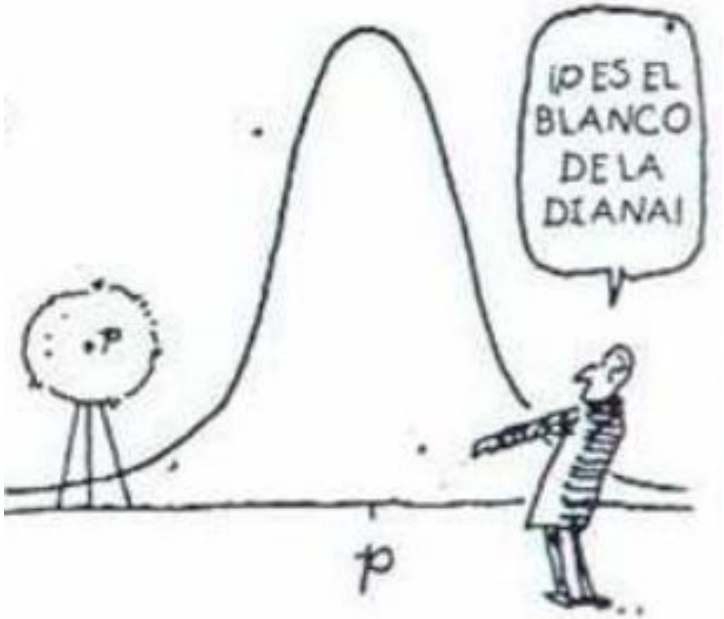
ANOVA table

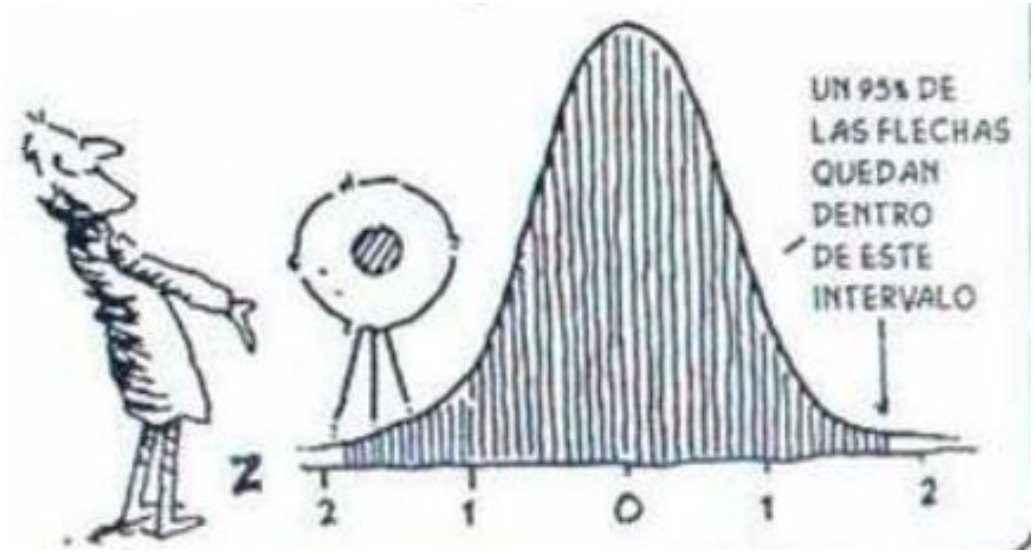
Two-way ANOVA Table

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square	*F*-ratio	*P-value*
Factor *A*	$r-1$	SS_A	MS_A	$F_A = MS_A / MS_E$	Tail area
Factor *B*	$c-1$	SS_B	MS_B	$F_B = MS_B / MS_E$	Tail area
Interaction	$(r-1)(c-1)$	SS_{AB}	MS_{AB}	$F_{AB} = MS_{AB} / MS_E$	Tail area
Error (within)	$rc(n-1)$	SS_E	MS_E		
Total	$rcn-1$	SS_T			

ANOVA table

Intervalo de confianza



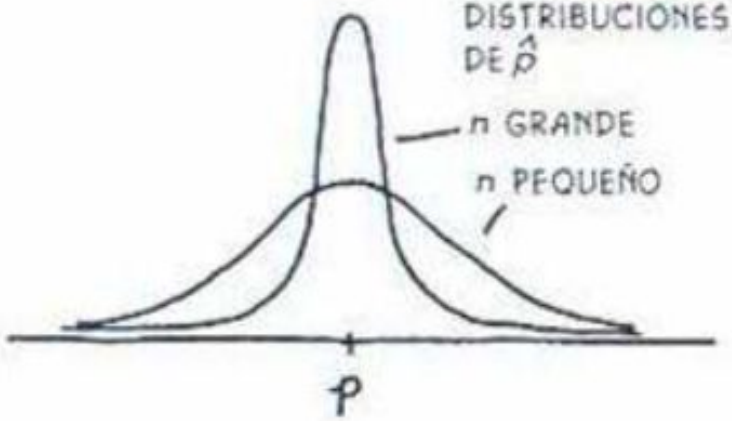


Intervalo de confianza

¿CÓMO PODEMOS CONSEGUIRLO? ¡AUMENTANDO EL TAMAÑO DE LA MUESTRA! LA AMPLITUD DEL INTERVALO DE CONFIANZA DEPENDE DEL TAMAÑO MUESTRAL: EL INTERVALO TIENE LA FORMA $\hat{p} + E$, EN LA QUE E, EL ERROR, VIENE DADO POR

$$E = z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

ASÍ QUE CUANTO MAYOR SEA n, EL ERROR SERÁ MENOR. (ES DECIR, SI MULTIPLICAMOS n POR CUATRO, LA AMPLITUD DEL INTERVALO SE REDUCE A LA MITAD.)



Intervalo de confianza

EN LOS ESTUDIOS CIENTÍFICOS, SE USA CON FRECUENCIA UN VALOR α FIJO DE 0,05 O 0,01. PODEMOS DECIR QUE ESTOS VALORES FIJOS SON RELIQUIAS DE LA ERA PREINFORMÁTICA, CUANDO NOS REFERÍAMOS A TABLAS QUE SÓLO SE PUBLICABAN PARA DETERMINADOS VALORES CRÍTICOS. AÚN HOY, EN ALGUNAS PUBLICACIONES CIENTÍFICAS SÓLO APARECEN LOS RESULTADOS SI EL VALOR $p \leq 0,05$.

Nivel de significancia

	Ho es verdadera	Ho es Falsa	Acusado!!
Aceptamos Ho	✓	 ERROR TIPO II	Ho: Inocente
Rechazamos Ho	 ERROR TIPO I	✓	Ha: Culpable

Teoria de decision