ANEXO 1. ESTADÍSTICA

1. TABLA RESUMEN CON LAS PRINCIPALES MEDIDAS ESTADÍSTICAS DE UNA DISTRIBUCIÓN UNIDIMENSIONAL

Media Aritmética	Es la suma de todos los valores de la distribución dividida por el número total de observaciones.	$\bar{x} = \frac{1}{N}\sum_{i=1}^{r} x_i n_i$
Media Geométrica	Es la raíz N-ésima del producto de los N valores observados.	$G = \sqrt[n]{x_1^{n_1} x_2^{n_2} \cdots x_r^{n_r}}$
Media Armónica	Es la media aritmética de los inversos de los valores de la variable.	$H = \frac{N}{\frac{n_1}{x_1} + \frac{n_2}{x_2} + \ldots \frac{n_r}{x_r}} = \frac{N}{\sum_{i=1}^{r} \frac{n_i}{x_i}}$
Mediana	En una distribución de frecuencias con los valores ordenados de menor a mayor, se denomina Mediana al valor de la variable que deja a su izquierda y a su derecha el mismo número de frecuencias, es decir aquel valor cuya frecuencia acumulada es N/2.	Cuando el nº de valores de la variable es impar: $M_e = N/2$ Cuando el nº de valores de la variable es par: M_e = media aritmética de los 2 términos centrales de la distribución. En las distribuciones agrupadas en intervalos: $M_e = L_{i-1} + \frac{\frac{N}{2} - N_{i-1}}{n_i} c_i$
Moda	Es el valor de una variable que se repite más veces, es decir, aquel que tiene mayor frecuencia absoluta	Para distribuciones agrupadas en intervalos: $M_o = L_{i-1} + \frac{n_{i-1}}{n_{i-1} + n_{i+1}} c_i$
Cuartiles	Son los tres valores que dividen la distribución en cuatro partes iguales; cada parte incluye, pues, el 25 % de los valores de la distribución.	C_1 = es el valor que ocupa el lugar N/4 C_2 = es el valor que ocupa el lugar 2N/4 C_3 = es el valor que ocupa el lugar 3N/4 Para distribuciones agrupadas en intervalos (k=4; r=1,2..3): $Q_{r/k} = L_{i-1} + \frac{\frac{r}{k}.N - N_{i-1}}{n_i} . c_i$
Deciles	Son los nueve valores que dividen la distribución en diez partes iguales; cada parte incluye, pues, el 10 % de los valores de la distribución.	D_1 = es el valor que ocupa el lugar N/10 D_2 = es el valor que ocupa el lugar 2N/10 . . D_9 = es el valor que ocupa el lugar 9N/10 Para distribuciones agrupadas en intervalos se emplearía la misma fórmula de los cuartiles en la que k=10 y r=1,2..9
	Son los 99 puntos o valores que	P_1 = es el valor que ocupa el lugar N/100

Percentiles	dividen la distribución en cien partes iguales.	P_2 = es el valor que ocupa el lugar 2N/100 . . P_{99} = valor que ocupa el lugar 99N/100 Para distribuciones agrupadas en intervalos se emplearía la misma fórmula de los cuartiles en la que k=100 y r=1,2..99
Recorrido o rango	Es la diferencia entre el mayor y el menor valor de una distribución.	$R_e = x_r - x_1$ = máx $\{x_i\}$ - mín $\{x_i\}$ para $1 \leq i \leq r$
Recorrido o intervalo intercualítico	Es la diferencia existente entre el tercer y el primer cuartil	$R_I = C_3 - C_1$
Recorrido o intervalo semiintercua-lítico	Es la media de la diferencia existente entre el tercer y el primer cuartil. Otros autores lo definen como una medida de dispersión relativa dada por el cociente entre el recorrido intercualítico y la suma del primer y tercer cuartil	$Rs_I = (C_3 - C_1) / 2$ $Rs_I = (C_3 - C_1) / (C_3 + C_1)$
Desviación absoluta media respecto a la media	Es la media aritmética de los valores absolutos de las diferencias entre los valores de la variable y la media aritmética	$D = \sum_{i=1}^{n} \lvert x_i - \bar{x} \rvert \frac{n_i}{N}$
Desviación absoluta media respecto a la mediana	Es la media aritmética de los valores absolutos de las diferencias entre los valores de la variable y la mediana	$D_{M_e} = \sum_{i=1}^{n} \lvert x_i - M_e \rvert \frac{n_i}{N}$
Varianza [1]	Es una medida de dispersión que se define como la media aritmética de los cuadrados de las desviaciones de los valores de la variable respecto a la media aritmética o el momento de segundo orden respecto a la media.	$S^2 = \sum_{i=1}^{n} (X_i - \overline{X})^2 . \frac{n_i}{N}$
CuasiVarianza	Es una medida de dispersión, cuya única diferencia con la varianza es que dividimos por N-1	$S^2_{N-1} = \sigma^2_{N-1} = \frac{N}{N-1} S^2_x$
Desviación típica	Es la raíz cuadrada de la varianza	$S = \sqrt{\sum_{i=1}^{n} (X_i - \overline{X})^2 . \frac{n_i}{N}}$
Coeficiente	Es la relación entre el mayor y el	

[1] Se suele designar con la letra S^2 cuando se refiere a la varianza de una muestra y con σ^2 cuando se refiere a la varianza de toda la población.

de apertura	menor valor de la distribución	$C_a = \frac{x_n}{x_1}$
Recorrido relativo	Es el cociente entre el recorrido y la media; mide, pues, el número de veces que el recorrido contiene a la media aritmética.	$R_r = \frac{R_e}{\overline{x}}$
Intervalo intercualítico relativo	Es el recorrido intercualítico dividido por la mediana de la distribución	$R_{ir} = (C_3 - C_1) / M_e$
Coeficiente de variación de Pearson	Es el cociente entre la desviación típica y la media; mide, pues, el número de veces que la desviación típica contiene a la media aritmética.	$C_v = \frac{s}{\overline{x}}$
Medidas de asimetría	**Coeficiente de asimetría de Fisher** $g_1 = 0$ simétrica $g_1 > 0$ asimétrica a la derecha $g_1 < 0$ asimétrica a la izquierda **Coeficiente de asimetría de Bowley** $A_b = 0$ simétrica $A_b > 0$ asimétrica a la derecha $A_b < 0$ asimétrica a la izquierda **Medida de asimetría de Pearson** $A_p = 0$ simétrica $A_p > 0$ asimétrica a la derecha $A_p < 0$ asimétrica a la izquierda	$g_1 = \frac{m_3}{S_3} = \frac{\frac{1}{N}\sum_{i=1}^{n}(X_i - \overline{X})^3 \cdot n_i}{(\sum_{i=1}^{n}(X_i - \overline{X})^2 \cdot \frac{n_i}{N})^{\frac{3}{2}}}$ $A_b = C_3 + C_1 - 2M_e / C_3 - C_1$ $A_p = \frac{(\overline{X} - M_e)}{s}$

2. EJEMPLO PRÁCTICO DE APLICACIÓN DE LOS PRINCIPALES CONCEPTOS DE ESTADÍSTICA DESCRIPTIVA.

Supongamos que disponemos de la siguiente información sobre el peso, en kilogramos, de 21 alumnos de un aula cualquiera:

74, 54, 60, 42, 38, 41, 50, 62, 48, 62, 46, 46, 62, 62, 70, 54, 62, 62, 50, 48, 41

Llamaremos X a nuestra variable observada, en este caso el peso en kilogramos. Se trata, pues, de una variable continua que, al observarla en una unidad de medida concreta, la hemos convertido en variable discreta.

Cada uno de los valores observados se denota por X_i, de forma tal que X_1 =65; X_2 = 54, etc.

Para analizar la distribución procedemos a ordenarla, en orden creciente o decreciente de sus valores; operando con la ordenación creciente y agrupando los valores similares, obtenemos la siguiente distribución:

X_i	n_i
38	1
41	2
42	1
46	2
48	2
50	2
54	2
60	1
62	6
70	1
74	1
	21

La primera columna recoge los valores agrupados y ordenados, mientras que la segunda columna refleja las frecuencias absolutas (n_i) con las que aparece cada valor X_i de la distribución; en la última fila de esta columna N = 21, se indica la suma de las frecuencias absolutas observadas.

Si tuviésemos un mayor número de valores sería aconsejable proceder a una agrupación de los mismos en intervalos, tal como, por ejemplo, la siguiente:

L_{i-1}-L_i	n_i
30-40	1
41-50	9
51-60	3
61-70	8

En la tabla siguiente reflejamos los valores obtenidos para las frecuencias absolutas y relativas de la distribución y para las frecuencias acumuladas ascendentes.

Frecuencias absolutas y relativas del peso de 21alumnos

X_i	n_i	f_i	$N_i^{\uparrow}$	$F^{\uparrow}_i$
38	1	0.05	1	0.05
41	2	0.10	3	0.14
42	1	0.05	4	0.19
46	2	0.10	6	0.29
48	2	0.10	8	0.38
50	2	0.10	10	0.48
54	2	0.10	12	0.57
60	1	0.05	13	0.62
62	6	0.29	19	0.90
70	1	0.05	20	0.95
74	1	0.05	21	1.00
	21			

En esta distribución, los estadísticos indicados en la Tabla 1.3. vendrían dados de la siguiente forma:

Media aritmética: $\overline{X} = \frac{1}{N}\sum_{i=1}^{N} X_i n_i$ = 1134/21= 54 Kgr.

Media geométrica: $G = \sqrt[n]{x_1^{n_1} x_2^{n_2} \ldots x_r^{n_r}} = \sqrt[20]{1.6833E+34} = 53.1$

Media armónica: $H = \frac{N}{\frac{n_1}{x_1} + \frac{n_2}{x_2} + .. \frac{n_r}{x_r}} = 52.2$

Mediana: Como tenemos un número impar de valores (21), la mediana será $M_e = 21/2 = 10,5$; esta frecuencia corresponde al valor 54 de la distribución.

Moda: 62 es el valor más veces repetido (6 veces).

Cuartiles:

C_1 = 1er cuartil: N/4 = 21/4= 5,5	Esta frecuencia corresponde al valor 46 de la distribución.
C_2 = 2º cuartil: 2N/4 = 42/4= 10,5	Esta frecuencia corresponde con la mediana (valor 54 de la distribución).
C_3 = 3er cuartil: 3N/4 = 63/4= 15,75	Esta frecuencia corresponde al valor 62 de la distribución.
C_4 = 4º cuartil: 4N/4 = 84/4= 21	Esta frecuencia corresponde al valor 74 de la distribución.

Deciles:

Primer decil: N/10 = 21/10= 2,1	Esta frecuencia corresponde al valor 41 de la distribución.
Segundo decil: 2N/10 = 42/10= 4,2	Esta frecuencia corresponde con el valor 46 de la distribución.
.....	.....
Noveno decil: 9N/10 = 189/10= 18,9	Esta frecuencia corresponde al valor 62 de la distribución.
Décimo decil: 10N/10 = 210/10= 21	Esta frecuencia corresponde al valor 74 de la distribución.

Recorrido: $R_e = 74 - 38 = 36$

Recorrido intercualítico: $R_I = C_3 - C_1 = 15,75 - 5,5 = 10,25$

Recorrido semi-intercualítico:

$Rs_I = (C_3 - C_1)/2 = (15,75 - 5,5)/2 = 5,125$ ó alternativamente
$Rs_I = (C_3 - C_1)/(C_3 + C_1) = (15,75 - 5,5)/(15,75 + 5,5) = 10,25/21,25 = 0,48$

Varianza[2]:

$$s^2 = \sum_{i=1}^{N} (X_i - \overline{X})^2 \frac{n_i}{N} = \frac{2046}{21} = 97.4$$

[2] Si el lector utiliza para efectuar estas operaciones la hoja de cálculo Excel y aplica la función VARP (varianza poblacional) obtendrá este valor, sí, por el contrario aplica la función VAR (varianza muestral), obtendrá para la varianza el valor 102,3, resultado de dividir la expresión de la varianza por "*n-1*" y no por "*n*" (en nuestro ejemplo por 20 y no por 21); este valor suele también conocerse por la cuasivarianza, y es un estimador muy utilizado en inferencia estadística.

Desviación típica: $s = 9,87$

Coeficiente de apertura: $C_a = X_n / X_1 = 74/38 = 1,95$

Coeficiente de Pearson: 9,87/54= 0,18; su valor próximo a cero indica una distribución con escasa dispersión, es decir, cuyos valores están bastante próximos a la media.

3. PRINCIPALES MEDIDAS ESTADÍSTICAS DE UNA DISTRIBUCIÓN BIDIMENSIONAL

Decimos que una serie de datos es multidimensional cuando para cada individuo recogemos información acerca de más de una variable.

Si solamente estudiamos 2 variables X, Y, podemos representar los datos en una tabla de doble entrada, de modo que, en la cabecera de las filas ponemos las modalidades de una de las variables y en la cabecera de las columnas las de la otra. En las celdillas que se forman se anota el número de observaciones que presentan a la vez las características de la fila y la columna en la que se encuentran.

A esta estructura se le denomina distribución conjunta o, en el caso de variables cualitativas, tabla de contingencia.

Una tabla tipo de distribución conjunta de frecuencias para dos variables es la siguiente:

	Variable					
Individuo	X_1	X_2		X_j		X_m
1	x_{11}	x_{12}		x_{1j}		x_{1m}
2	x_{12}	x_{22}		x_{1j}		x_{1m}
i	x_{i1}	x_{i2}		x_{ij}		x_{im}
n	x_{n1}	x_{n2}		x_{nj}		x_{nm}

Las distribuciones que aparecen en la última fila y en la última columna de la tabla son, en realidad, las distribuciones *univariables* de X y de Y consideradas por separado, y se denominan distribuciones marginales.

Se denomina distribución condicionada de Y para $X = x_i$ a la distribución que aparece en la columna i.
Se denomina distribución condicionada de X para $Y = y_i$ a la distribución que aparece en la fila i.

Para múltiples variables los conceptos son similares a los presentados en el caso de dos variables, pero con la diferencia de que no es posible tabular los datos de la misma forma, ya que la única posibilidad es detallar los resultados para cada individuo:

La estructura resultante es una matriz de n filas por m columnas, denominada matriz de datos.

Si consideramos cada una de las variables por separado (DISTRIBUCIÓN MARGINAL), podemos tratarla como una distribución univariable, calculando su media. Con el resultado obtenido, podemos construir un vector de dimensión m, que se denomina VECTOR DE MEDIAS y que contiene la media de cada variable:

MATRIZ DE VARIANZAS Y COVARIANZAS. COEFICIENTES DE CORRELACIÓN.

VARIANZA.

Igual que sucede con la media, podemos calcular la varianza de cada variable por separado, denotando S_i^2 la varianza de X_i.

COVARIANZA.

Se define la covarianza entre dos variables X e Y como:

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

Nótese que $S_{xy} = S_{yx}$.

La covarianza no tiene porque ser necesariamente positiva, cosa que sí sucede con la varianza.

Expresa el grado de variación conjunta de dos variables. En este sentido puede suceder que:

1) COVARIANZA > 0 ===> Cuando aumenta una de ellas, también aumenta la otra.
2) COVARIANZA < 0 ===> Cuando aumenta una, la otra disminuye
3) COVARIANZA = 0 ===> No hay relación entre los aumentos de una y otra.

Estas relaciones pueden ser de mayor o menor intensidad, según la magnitud de la COVARIANZA. El mayor inconveniente de la covarianza es que su magnitud no sólo depende del grado de variación conjunta de las variables, sino también de la dispersión de cada variable. Para eliminar la influencia de este último factor, se utiliza el denominado coeficiente de correlación.

COEFICIENTE DE CORRELACIÓN:

$$r = \frac{S_{xy}}{S_x S_y}$$

Es un coeficiente adimensional cuyo valor es siempre mayor o igual que -1 y menor o igual que 1.

Cuando su valor es 1 indica que la variación conjunta es máxima, de modo que existe una relación lineal perfecta entre las variables, que puede expresarse mediante una ecuación del tipo Y = a + bX, por lo que podemos prescindir de una de ellas.

VARIANZA GENERALIZADA: MATRIZ DE VARIANZAS Y COVARIANZAS.

Dada una distribución *multidimensional* con m variables, se define la varianza generalizada, *S*, como:

$$\mathbf{S} = \begin{bmatrix} S_{11}^2 & S_{12}^2 & S_{13}^2 & S_{1n}^2 \\ S_{21}^2 & S_{22}^2 & S_{23}^2 & S_{2n}^2 \\ S_{31}^2 & S_{32}^2 & S_{33}^2 & S_{3n}^2 \\ & & & \\ S_{m1}^2 & S_{m2}^2 & S_{m3}^2 & S_{mn}^2 \end{bmatrix}$$

Donde S es una matriz *simétrica semidefinida positiva*.

4. CORRELACIÓN Y REGRESIÓN ENTRE DOS VARIABLES.

Bastante frecuentemente se aprecia que ciertas variables macroeconómicas varían entre sí con una sincronización más o menos intensa. Para expresar esta variación conjunta se emplea el término de covariación. Esta variación conjunta puede ser de distintos tipos:

1º Dependencia causal unilateral.

Esto se da cuando una variable influye en la otra, pero no al contrario. Por ejemplo la cantidad de lluvia (X) influye en el rendimiento de la cosecha (Y), pero el rendimiento de la cosecha no influye sobre la cantidad de lluvia. La variable X se denomina *independiente explicativa* la variable Y se denomina *dependiente explicada*.

2º Interdependencia.

En este caso la influencia entre X e Y es recíproca y se produce por lo tanto en dos direcciones. Hay pues dependencia bilateral o independencia. Un ejemplo puede ser la relación entre el aumento del PIB y el aumento del empleo. A mayor nivel de empleo mayor crecimiento del PIB y a mayor PIB mayor aumento del empleo.

3º Dependencia indirecta.

Dos variables X e Y pueden mostrar una covariación conjunta, debido a la influencia de una tercera variable Z que influye sobre las dos.

Por ejemplo si medimos en una variable X la longitud del Pie de un niño y en una variable Y su capacidad lectora, observaremos que a un mayor tamaño de pie corresponde una mejor capacidad lectora. Esto es debido que la variable Z = edad del niño influye tanto en el crecimiento del pie como en su capacidad para leer.

4º Concordancia.

A veces sabemos que dos variables X e Y, son independientes. No obstante, se desea saber si en sus variaciones existe una cierta concordancia.

Por ejemplo dos profesores corrigen un mismo examen a un mismo grupo de alumnos, de manera independiente de tal forma que uno no conoce las calificaciones del otro. Sería interesante conocer si existe concordancia entre las variaciones que se producen.

5º Covariación Causal.

Hay veces que entre dos variables se observa una sincronización de la que pudiera deducirse una asociación o dependencia entre dichas variables. No obstante dicha covariación o dependencia puede ser accidental o casual. A esta conclusión se llega cuando se sabe a ciencia cierta que no existe dicha relación entre dichas variables.

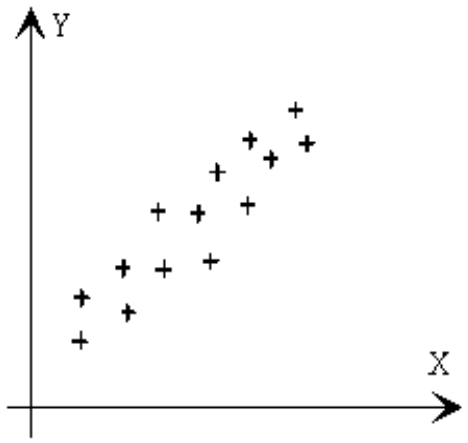
Es conveniente resaltar que el tipo de relación observada entre dos variables no se deduce de los datos estadísticos de que se disponga; el decir a cuál l de los cinco tipos de covariación pertenece el caso que estemos estudiando depende del conocimiento previo que tengamos de ambas variables.

La estadística lo que hace es por medio de técnicas numéricas cuantificar y formalizar matemáticamente la relación si existe, para poder explicarla, y poder realizar predicciones.

Vamos a estudiar la covariación que existe entre dos variables cuantitativas.

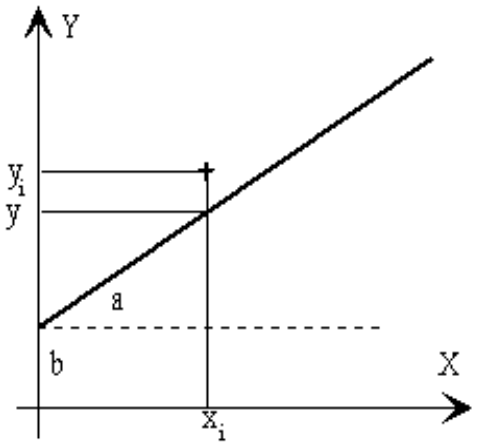
Si utilizamos un sistema de coordenadas cartesianas para representar la distribución bidimensional, obtendremos un conjunto de puntos conocido como el diagrama de dispersión, cuyo análisis permite estudiar cualitativamente, la relación entre ambas variables tal como se ve en la figura. El siguiente paso, es la determinación de la dependencia funcional que mejor se ajusta a la distribución bidimensional entre las dos variables x e y. Si suponemos una función lineal esta dependencia se obtiene mediante la regresión lineal, es decir, mediante la determinación de dos parámetros: la pendiente y la ordenada en el origen de la recta de regresión, $y=ax+b$.

Una vez obtenida esta recta de regresión, además de determinar el grado de dependencia de las series de valores X e Y, podemos predecir el valor *y* estimado que se obtendría para un valor *x* que no esté en la distribución.



Utilizando el método de los mínimos cuadrados, se trata, pues, de determinar la ecuación de la recta que mejor se ajusta a los datos representados en la figura. Se denomina error e_i a la diferencia y_i-y, entre el valor observado y_i, y el valor ajustado $y= ax_i+b$, tal como se ve en la figura inferior. El criterio de ajuste se toma como aquél en el que la desviación cuadrática media sea mínima, o lo que es lo mismo que el error cometido al realizar la predicción de Y sea mínimo es decir, debe ser mínima la suma:

$$e=\sum_{i=1}^{n}\left(Y_i-Y_i'\right)^2=\sum_{i=1}^{n}\left(Y_i-a-bX_i\right)^2$$



El extremo de una función, máximo o mínimo, se obtiene cuando las derivadas de *s* respecto de *a* y de *b* sean nulas. Lo que da lugar a un sistema de dos ecuaciones con dos incógnitas del que se despeja *a* y *b*.

Traduciendo el criterio anterior a términos matemáticos nos queda:

$$\frac{\partial}{\partial a}\sum_{i=1}^{n}\left(Y_i - a - bX_i\right)^2 = 0$$

$$\frac{\partial}{\partial b}\sum_{i=1}^{n}\left(Y_i - a - bX_i\right)^2 = 0$$

Operando llegamos al siguiente sistema de ecuaciones:

$$\sum_{i=1}^{n} y_i = Na + b\sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{n} y_i x_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2$$

que desarrollado queda:

$$a = \overline{Y} - bX$$

$$b = \frac{N\sum_{i=1}^{N} X_i Y_i - \sum_{i=1}^{N} X_i \sum_{i=1}^{N} Y_i}{N\sum_{i=1}^{N} X_i^2 - \left(\sum_{i=1}^{N} X_i\right)^2} = \frac{s_{xy}}{s_x^2}$$

Luego la recta obtenida será:

$$Y - \overline{Y} = \frac{s_{xy}}{s_x^2}\left(X - \overline{X}\right)$$

El parámetro b se denomina coeficiente de regresión de Y sobre X.

La correlación será mayor cuánto más se aproximen los puntos a las curva. El coeficiente de correlación nos indica la intensidad o grado de dependencia entre las variables X e Y. Dicho coeficiente *r* es un número que se obtiene mediante la fórmula:

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

y que tiene las siguientes propiedades:

a) $-1 \leq rxy \leq 1$

b) $r_{xy} = r_{yx}$

c) Es invariable bajo transformaciones lineales:

INTERPRETACIÓN:
Los valores extremos no plantean duda:

a) $|rxy|=1$ Indica que hay una relación lineal perfecta, por lo que podemos calcular exactamente que valor de la segunda variable se asocia con cada uno de los de la primera, o viceversa.

b) $rxy = 0$ Indica que no existe ninguna relación entre las variables.

La interpretación de otros valores es muy relativa y depende de cada estudio. En general suele aceptarse la siguiente clasificación:

$0 \leq \|rxy\| < 0.30$	Relación baja entre las variables.
$0.30 \leq \|rxy\| < 0.70$	Relación media.
$0.70 \leq \|rxy\| \leq 1$	Relación alta.

Ejemplo:

Nº de parados según la EPA (miles) Trimestre					**Población mayor de 16 años (miles de personas)**				
Año	1º	2º	3º	4º	Año	1º	2º	3º	4º
1977	630	641	718	749	1977	25681	25776	25856	25923
1978	834	896	962	1013	1978	25993	26098	26170	26243
1979	1058	1107	1178	1263	1979	26330	26416	26513	26585
1980	1381	1509	1557	1663	1980	26680	26780	26887	26992
1981	1759	1877	1964	2047	1981	27086	27156	27252	27322
1982	2105	2168	2261	2337	1982	27438	27521	27609	27726
1983	2408	2435	2514	2586	1983	27799	27880	27972	28064
1984	2640	2724	2795	2912	1984	28170	28243	28335	28432
1985	2932	2975	2983	2991	1985	28522	28642	28724	28820
1986	2987	2967	2946	2938	1986	28872	28950	29041	29148
1987	2964	2968	2972	2916	1987	29246	29362	29456	29550
1988	2915	2913	2886	2693	1988	29645	29740	29836	29932
1989	2661	2568	2501	2509	1989	30028	30124	30221	30319
1990	2477	2453	2423	2409	1990	30363	30408	30452	30496
1991	2390	2403	2512	2550	1991	30575	30652	30728	30806
1992	2598	2703	2822	3033	1992	30881	30954	31027	31099
1993	3257	3416	3586	3673	1993	31170	31238	31307	31375
1994	3740	3782	3739	3694	1994	31452	31530	31608	31686

Suma de y = 171536
Suma de x = 2066802
n = 72
Suma de xy = 5015812736
Suma de x^2 = 59564571954

El sistema de ecuaciones normales será:

171536 = 72 a + 2066802 b
5015812736 = 2066802 a + 59564571954 b

Cuya solución es:

-354530947872 = -148809744 a - 4271670507204 b
361138516992 = 148809744 a + 4288649180688 b
6607569120 = 16978673484 b

b = 0,389
a = - 8788,872

Con lo que la recta ajustada será

y_j * = - 8788,872 + 0,389 xi

Veamos ahora cuál será el nº de parados cuando la población mayor de 16 años sea igual a:

Población	Valor Real	Valor Estimado
25993	834	1327
30575	2390	3110
31452	3740	3451

5. DISTRIBUCIÓN NORMAL. USO DE TABLAS

La distribución normal, o de Gauss, es la más importante de las distribuciones estadísticas. Fue descubierta por Gauss cuando estudiaba errores de medición, aunque ya antes había sido establecida por De Moivre como límite de la distribución Binomial *B (n, p)*, cuando n $\to\infty$

Es una distribución de tipo continuo, que representa multitud de sucesos naturales; ejemplos de variables que suelen adaptarse a esta distribución son la altura de las personas de una edad y sexo determinados, los errores de medición de un calibrador, los coeficientes de inteligencia de las personas que se presentan a una prueba de selección de personal, etc..

La función de densidad de esta distribución adopta la siguiente forma:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \text{ para } -\infty < x < \infty$$

Donde:

μ representa la esperanza matemática
σ la desviación típica, y
π = 3,1415..
e = 2,718281..

Abreviadamente se dice que $y \to N(\mu,\sigma)$.

Recordemos que la función de densidad de una distribución cualquiera cumple dos condiciones, a saber:

$$f(x) \geq 0, \text{ para } -\infty \leq x \leq +\infty$$

$$\int_{-\infty}^{+\infty} f(x) = 1$$

La primera indica que la probabilidad es siempre positiva y la segunda que el conjunto de toda la probabilidad, es decir el área existente bajo la curva de la función de densidad es la unidad.

Conocida una función de densidad, se define la función de distribución acumulativa de una variable aleatoria X, como la probabilidad de que la variable aleatoria continua X tome valores menores o iguales a *x*, es decir:

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(x)$$

En el caso de la distribución normal, se tiene que:

$$F(x) = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} .dx$$

Tanto la función de densidad como la función de distribución de una normal quedan, pues, perfectamente definidas si se conoce su esperanza μ y su varianza y, consiguientemente su desviación típica σ, representándose abreviadamente como N(μ, σ).

Puede demostrarse que la distribución normal es simétrica respecto al valor de μ; la forma de la campana que la representa es más o menos apuntada o plana en función del valor de σ.
Decimos que una distribución normal es de tipo reducido o tipificado cuando μ=0 y σ=1, de forma que su función de densidad será:

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

y su función de distribución:

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}^2} .dx$$

Para tipificar una distribución normal procedemos a restar a todos los valores de X su media μ y a dividir el resultado por su desviación típica σ, de forma tal que:

$$Z = \frac{X - \mu}{\sigma}$$

Los valores que toma la función de distribución de la N(0,1) se presentan habitualmente tabulados para diferentes valores de X, de forma que el usuario no tenga necesidad de calcular continuamente el valor de la integral más arriba indicada.

Los principales valores que toma esta integral son:

x	*F(X)*
0	0,5
0,1	0,5398
..	.....
1	0,8413
1,5	0,9332
..	.....
2	0,97725
..	...
3	0,99865

Dichos valores indican que el área o probabilidad existente a la izquierda de x =3, es 0,99865, valor muy próximo a 1 que indica que la "cola de probabilidad" a la derecha de x = 3 es prácticamente insignificante.

Explicamos a continuación el uso que puede hacerse de esta tabla.

Consideremos a tal fin una variable aleatoria X que se distribuye como una normal de media 2 y desviación típica 4 y de la que deseamos conocer la probabilidad de que un valor de X sea menor que 6.

Realizamos en primer lugar el proceso de tipificación:

$$P(x < 6) = P(\frac{X-2}{3} < \frac{6-2}{3}) = P(Z < 1)$$

donde Z = (X-2)/3 es ahora una variable tipificada del tipo N (0,1).

Mirando dicho valor en la segunda columna de la tabla obtenemos que la probabilidad pedida para Z es 0,8413. Es decir, P(X< 6) = P (Z< 1)= 84,13 %.

La función de distribución de una normal es simétrica, de forma que los valores de F(x) correspondientes a "x" negativas, se pueden obtener por simetría, sabiendo que P (Z≤0) = 0,5, ya que en esta distribución el valor z = 0, además de ser la media coincide con la mediana.

Los casos que pueden darse en la utilización de la tabla son los siguientes:

P ($Z \leq x_1$)
P ($Z \geq x_1$)
P ($Z \leq -x_1$)
P ($x_1 \leq Z \leq x_2$)
P ($-x_1 \leq Z \leq -x_2$)
P ($-x_1 \leq Z \leq x_2$)

Las reglas básicas que deben aplicarse son las siguientes:

$P(Z \leq x_1) = F(x_1)$, cuando $x_1 > 0$

Ejemplo: P(Z≤1,72) = F(1,72)= 0,95728 (4ª columna de la tabla).

$P(Z \geq x_1) = 1- F(x_1)$

Ejemplo: P(Z≥1) = 1- F(1)= 1- 0,8413= 0,1587

$P (Z \leq -x_1) = F (-x_1) = 1- F(x_1)$

Ejemplo: P(Z ≤-1,45) = 1-F(1,45)= 1- 0,92647= 0,07353

$P (x_1 \leq Z \leq x_2) = F (x_2) - F (x_1)$

Ejemplo: P(2 ≤ Z ≤ 3) = F(3)- F(2) = 0,998650 - 0,97725 = 0,0214

$P (-x_1 \leq Z \leq -x_2) = F(-x_2) - F(-x_1)=[1- F(x_2)]- [1- F(x_1)]= F(x_2)- F(x_1)]$

Ejemplo: P(-2 ≤ Z ≤ -3) = F(3) - F (2) = 0,998650 - 0,97725 = 0,02115

$P (-x_1 \leq Z \leq x_2) = F(x_2)- F(-x_1)$

Ejemplo: $P(-3 \leq Z \leq 3) = F(3) - F(-3) = F(3) - 1 + F(3) = 2\,F(3) - 1 = 2.\ 0{,}99865 - 1 = 0{,}9973$

El lector puede comprobar los siguientes datos fundamentales:

$P(-1 \leq Z \leq 1) = 0{,}6826$
$P(-2 \leq Z \leq 2) = 0{,}9544$
$P(-1{,}96 \leq Z \leq 1{,}96) = 0{,}95$

Indicativos de que:

- Entre la media (0) y ± una vez la desviación típica (1), se encuentran comprendidos el 68,26 % de los valores de la distribución.
- Entre la media y ± dos veces la desviación típica se encuentran comprendidos el 95,44 % de los valores de la distribución.
- Entre la media y ± tres veces la desviación típica se encuentran comprendidos el 99,73 % de los valores de la distribución.
- Entre la media y ± 1,96 veces la desviación típica se encuentran comprendidos el 95 % de los valores de la distribución.

6. TABLA DE LA DISTRIBUCIÓN NORMAL

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
.7	.7580	.7611	.7642	.7673	.7703	.7734	.7764	.7794	.7823	.7852
.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	. 8508	.8531	.8554	.8577	.8599	.8661
1.1	.8643	.8665	.8686	.8708	. 8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	. 8925	.8944	.8962	.8980	.8997	.90147
1.3	.90320	.90490	.90658	.90824	.90988	.91149	.91309	.91466	.91621	.91774
1.4	.91924	.92073	.92220	.92364	.92507	.92647	.92785	.92922	.93056	.93189
1.5	.93319	.93448	.93574	.93669	.93822	.93943	.94062	.94179	.94295	.94408
1.6	.94520	.94630	.94738	.94845	.94950	.95053	.95154	.95254	.95352	.95449
1.7	.95543	.95637	.95728	.95818	.95907	.95994	.96080	.96164	.96246	.96327
1.8	.96407	.96485	.96562	.96638	.96712	.96784	.96856	.96926	.96995	.97062
1.9	.97128	.97193	.97257	.97320	.97381	.97441	.97500	.97558	.97615	.97670
2.0	.97725	.97778	.97831	.97882	.97932	.97982	.98030	.98077	.98124	.98169
2.1	.98214	.98257	.98300	.98341	.98382	.98422	.98461	.98500	.98537	.98574
2.2	.98610	.98645	.98679	.98713	.98745	.98778	.98809	.98840	.98870	.98899
2.3	.98928	.98956	.98983	.990097	.990358	.990613	.990863	.991106	.991344	.991576
2.4	.991802	.992024	.992240	.992451	.992656	.992857	.993053	.993244	.993431	.993613
2.5	.993790	.993963	.994132	.994297	.994457	.994614	.994766	.994915	.995060	.995201
2.6	.995339	.995473	.995604	.995731	.995855	.995975	.996093	.996207	.996319	.996427
2.7	.996533	.996636	.996736	.996736	.996928	.997020	.997110	.997197	.997282	.997365
2.8	.997445	.997523	.997599	.997673	.997744	.997814	.997882	.997948	.998012	.998074
2.9	.998134	.998193	.998250	.998305	.998359	.998411	.998462	.998511	.998559	.998605
3.0	.998650	.998694	.998736	.998777	.998817	.998856	.998893	.998930	.998965	.998999

7. DISTRIBUCIÓN T DE STUDENT

Es una distribución derivada también de la distribución normal. Definimos la variable *t* de Student con *n* grados de libertad *t(n)* como:

$$t(n) = \frac{\eta}{\sqrt{\frac{\eta_1^2 + \eta_2^2 + \eta_3^2 + \ldots \eta_n^2}{n}}}$$

Siendo $\eta_1^2, \eta_2^2, \eta_3^2, .., \eta_n^2$ variables aleatorias N(0,σ). En concreto, la función de densidad de la distribución *t* es:

$$f(x) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\,\Gamma(\frac{n}{2})}(1 + \frac{x^2}{n})^{-\frac{n+1}{2}}$$

La función es simétrica respecto al eje de ordenadas.

La esperanza matemática de la distribución es igual a cero y la varianza es igual a:

$$\frac{n}{n-2}$$

Finalmente, la función de distribución es igual a:

$$F(x) = P(x \le x_n) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\,\Gamma(\frac{n}{2})} \int_{-\infty}^{xn} (1 + \frac{x^2}{n})^{-\frac{n+1}{2}} dx$$

Manejo de tablas

El manejo de tablas es similar al de los casos anteriores. Se debe mirar los grados de libertad y la probabilidad asociada al valor. Presentamos los siguientes ejemplos:

Ejemplo 1
Determinar la P(t(7)) ≤1,119.

$$P(t(7)) \leq 1{,}119 = 0{,}85.$$

Ejemplo 2

Determinar P(t(8)) ≥2,896

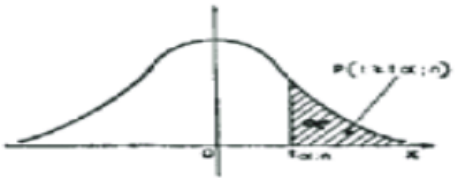
Al ser simétrica, tenemos que:

$$P(t(8) \geq 2{,}896 = 1 - [P(t(8) \leq 2{,}896] = 1 - 0{,}990 = 0{,}001$$

Ejemplo 3 Determinar $P(t(10) \geq 0{,}542$

Al ser una función simétrica tenemos que $P(t(10) \geq 0{,}542 = 1 - [P(t(10) \leq 0{,}542] = 0{,}7$

Distribución t de Student



n \ α	0,40	0,30	0,20	0,10	0,050	0,025	0,010	0,005	0,001	0,0005
1	0,325	0,727	1,376	3,078	6,314	12,71	31,82	63,66	318,3	636,6
2	0,289	0,617	1,061	1,886	2,920	4,303	6,965	9,925	22,33	31,60
3	0,277	0,584	0,978	1,638	2,353	3,182	4,541	5,841	10,22	12,94
4	0,271	0,569	0,941	1,533	2,132	2,776	3,747	4,604	7,173	8,610
5	0,267	0,559	0,920	1,476	2,015	2,571	3,365	4,032	5,893	6,859
6	0,265	0,553	0,906	1,440	1,943	2,447	3,143	3,707	5,208	5,959
7	0,263	0,549	0,896	1,415	1,895	2,365	2,998	3,499	4,785	5,405
8	0,262	0,546	0,889	1,397	1,860	2,306	2,896	3,355	4,501	5,041
9	0,261	0,543	0,883	1,383	1,833	2,262	2,821	3,250	4,297	4,781
10	0,260	0,542	0,879	1,372	1,812	2,228	2,764	3,169	4,144	4,587
11	0,260	0,540	0,876	1,363	1,796	2,201	2,718	3,106	4,025	4,437
12	0,259	0,539	0,873	1,356	1,782	2,179	2,681	3,055	3,930	4,318
13	0,259	0,538	0,870	1,350	1,771	2,160	2,650	3,012	3,852	4,221
14	0,258	0,537	0,868	1,345	1,761	2,145	2,624	2,977	3,787	4,140
15	0,258	0,536	0,866	1,341	1,753	2,131	2,602	2,947	3,733	4,073
16	0,258	0,535	0,865	1,337	1,746	2,120	2,583	2,921	3,686	4,015
17	0,257	0,534	0,863	1,333	1,740	2,110	2,567	2,898	3,646	3,965
18	0,257	0,534	0,862	1,330	1,734	2,101	2,552	2,878	3,611	3,922
19	0,257	0,533	0,861	1,328	1,729	2,093	2,539	2,861	3,579	3,883
20	0,257	0,533	0,860	1,325	1,725	2,086	2,528	2,845	3,552	3,850
21	0,257	0,532	0,859	1,323	1,721	2,080	2,518	2,831	3,527	3,819
22	0,256	0,532	0,858	1,321	1,717	2,074	2,508	2,819	3,505	3,792
23	0,256	0,532	0,858	1,319	1,714	2,069	2,500	2,807	3,485	3,767
24	0,256	0,531	0,857	1,318	1,711	2,064	2,192	2,797	3,467	3,745
25	0,256	0,531	0,856	1,316	1,708	2,060	2,485	2,787	3,450	3,725
26	0,256	0,531	0,856	1,315	1,706	2,056	2,479	2,779	3,435	3,707
27	0,256	0,531	0,855	1,314	1,703	2,052	2,473	2,771	3,421	3,690
28	0,256	0,530	0,855	1,313	1,701	2,048	2,467	2,763	3,408	3,674
29	0,256	0,530	0,854	1,311	1,699	2,045	2,462	2,756	3,396	3,659
30	0,256	0,530	0,854	1,310	1,697	2,042	2,457	2,750	3,385	3,646
40	0,255	0,529	0,851	1,303	1,648	2,021	2,423	2,704	3,307	3,551
50	0,255	0,528	0,849	1,298	1,676	2,009	2,403	2,678	3,262	3,495
60	0,254	0,527	0,848	1,296	1,671	2,000	2,390	2,660	3,232	3,460
80	0,254	0,527	0,846	1,292	1,664	1,990	2,374	2,639	3,195	3,415
100	0,254	0,526	0,845	1,290	1,660	1,984	2,365	2,626	3,174	3,389
200	0,254	0,525	0,843	1,286	1,653	1,972	2,345	2,601	3,131	3,339
500	0,253	0,525	0,842	1,283	1,648	1,965	2,334	2,586	3,106	3,310
∞	0,253	0,524	0,842	1,282	1,645	1,960	2,326	2,576	3,090	3,291