

# Regresión lineal e Incertidumbre

<b>1. Regresión lineal</b>	2
1.1. Análisis de regresión lineal simple	2
1.1.1. Obtención de los coeficientes a, b, y r	2
1.1.2. Cálculo de las desviaciones típicas y covarianzas	9
1.1.3. Confianza en las estimaciones del análisis de regresión	14
1.1.4. Transformaciones	21
1.2. Funciones de distribución de probabilidad	23
1.2.1. Distribución normal	23
1.2.2. Distribución t-Student	25
1.3. Funciones estadísticas en las hojas de cálculo	28
<b>2. Incertidumbre en las medidas experimentales</b>	31
2.1. Distinción entre error e incertidumbre	31
2.2. Ley de propagación de incertidumbres	32
2.3. Apéndices	38
2.3.1. Deducción de la fórmula RSS y de la suma algebraica	38
2.3.2. Aplicabilidad comparada de las fórmulas RSS y de la suma algebraica	39
2.3.3. Sobre la covarianza	41
<b>3. Presentación de resultados numéricos</b>	41
3.1. Reglas para el redondeo de números	42
3.2. Reducción del número de cifras significativas en las incertidumbres	42
3.3. Reducción del número de cifras significativas en valor de la magnitud estimado	43

---

Apuntes de regresión lineal, incertidumbre, reducción de cifras significativas y redondeo.  
Fundamentos Tecnológicos de los Computadores  
Ingeniería Técnica de Informática de Gestión.

---

Octubre de 2006  
Granada, ESPAÑA  
<http://ftcgestion.iespana.es>

*Ignacio Melchor Ferrer*

Reg\_Inct.pdf

---

# 1. Regresión lineal

## 1.1. Análisis de regresión lineal simple

El análisis de regresión lineal simple, nos permitirá analizar la relación entre dos variables ( $x$  e  $y$ ), bajo la hipótesis de que el valor de la variable  $y$  (variable dependiente) depende linealmente sólo de  $x$  (variable independiente). Si dependiese de más variables, entraríamos en el análisis de regresión lineal múltiple, y ya no usaríamos el adjetivo “simple”. El objetivo será encontrar esa relación lineal entre las dos variables, para ello, tomaremos suficientes datos experimentales de las variables  $x$  e  $y$ . Esos datos experimentales los llamaremos  $x_i$  e  $y_i$ . Debemos tener cuidado de que  $y_i$  sea la medida de la variable  $y$  cuando la variable  $x$  tome el valor  $x_i$ . Con el tratamiento de dichos datos experimentales obtendremos los coeficientes  $a$  y  $b$  de la ecuación lineal del ajuste:

$$y = a + b x$$

En realidad los coeficientes  $a$  y  $b$  son sólo estimaciones de los coeficientes teóricos, pero con el análisis de regresión podemos también conocer la bondad de esas aproximaciones.

Siempre debemos comprobar, si existe realmente una correlación lineal significativa entre las dos variables ( $x$  e  $y$ ), ya que aunque tengamos calculados los coeficientes  $a$  y  $b$ , pueden ser producto del azar. Las comprobaciones sobre la correlación lineal (coeficiente  $r$ ) se dejan para más adelante, pero en el trabajo, sería realmente la primera tarea a realizar, puesto que si no existe correlación lineal entre las variables, no tiene ningún sentido calcular los coeficientes  $a$  y  $b$ .

En una aproximación elemental al análisis de regresión lineal, no se tienen en cuenta los errores sistemáticos (sólo los aleatorios) sobre las medidas de la variable dependiente. Tampoco se suelen considerar errores (sistemáticos o aleatorios) sobre las medidas de la variable independiente. El análisis de regresión lineal, también suele hacerse suponiendo que la muestra de datos ( $x, y$ ) es aleatoria, y que la distribución de los valores de la variable dependiente es una distribución normal.

### 1.1.1. Obtención de los coeficientes $a$ , $b$ , y $r$

Nuestro objetivo ahora, es obtener con los datos experimentales los coeficientes  $a$  y  $b$  de la recta:

$$y = a + b x$$

Supongamos que tenemos  $n$  datos experimentales obtenidos como parejas de valores ( $x_i, y_i$ ), y los errores en las medidas de la variable dependiente se distribuyen normalmente. Si deseamos obtener los mejores estimadores de los coeficientes  $a$  y  $b$  de la recta de ajuste, deberíamos buscar los valores de  $a$  y  $b$  que minimizan el error total. Para ello introducimos la función  $L(y_i, a, b)$  que es el productorio de todas las funciones densidad de probabilidad de error:

$$L(y_i, a, b) = \prod_i \frac{1}{\sqrt{2\pi} \sigma} e^{-\frac{1}{2} \left[ \frac{y_i - (a + b x_i)}{\sigma} \right]^2}$$

Necesitamos que la función  $L(y_i, a, b)$  sea mínima, ya que es una medida de lo buena que es la elección de los coeficientes  $a$  y  $b$  de la recta de regresión lineal. Para obtener el mejor estimado para los coeficientes  $a$  y  $b$ , habría que hacer:

$$\frac{\partial L}{\partial a} = 0 \quad ; \quad \frac{\partial L}{\partial b} = 0$$

A veces se usa la función  $W(y_i, a, b) = \ln [L(y_i, a, b)]$ , con lo que el productorio se convertiría en un sumatorio, y para obtener los estimados de  $a$  y  $b$ , habría que despejar de:

$$\frac{\partial W}{\partial a} = 0 \quad ; \quad \frac{\partial W}{\partial b} = 0$$

que se reduciría al siguiente par de ecuaciones:

$$\frac{\partial \sum [y_i - (a + bx_i)]^2}{\partial a} = 0 \quad ; \quad \frac{\partial \sum [y_i - (a + bx_i)]^2}{\partial b} = 0$$

Antes de seguir, hemos de remarcar que en la definición de  $L(y_i, a, b)$  hemos supuesto que todas las medidas  $y_i$  tienen la misma desviación típica ( $\sigma$ ). Las fórmulas de regresión lineal de estos apuntes no servirían, por ejemplo, en el caso de que se midiesen las  $y_i$  con dos rangos de medida distintos en un polímetro (o cualquier otro instrumento de medida).

El error o residuo ( $e_i$ ) es la diferencia entre el dato experimental de la variable dependiente ( $y_i$ ) y el valor predicho por el modelo lineal buscado ( $y(x_i) = a + bx_i$ ) (notado en las fórmulas como  $y$  y circunflejo).

$$e_i = y_i - \hat{y}_i = y_i - y(x_i) = y_i - (a + bx_i)$$

La “suma residual de cuadrados” o “suma de los cuadrados de los errores” (SCE o SSE) también llamado error cuadrático de la recta de ajuste, es la suma a todos los datos experimentales del cuadrado de los errores o residuos ( $e_i$ ).

$$SCE = \sum e_i^2 = \sum (y_i - \hat{y}_i)^2 = \sum (y_i - y(x_i))^2 = \sum [y_i - (a + bx_i)]^2$$

Si la suma residual de cuadrados (SCE) es pequeña, tendremos la seguridad de que el error cuadrático medio de la recta de regresión  $y=a+bx$  es pequeño, y por tanto la recta  $y=a+bx$  es la “mejor” posible. Si la suma residual de cuadrados es mínima, el error cuadrático medio de la recta  $y=a+bx$  será mínimo:

$$\frac{\sum e_i^2}{n} = \frac{\sum [y_i - y(x_i)]^2}{n} = \frac{\sum [y_i - (a + bx_i)]^2}{n}$$

( $n$  es el número de pares de valores ( $x_i, y_i$ ) obtenidos experimentalmente)

Para obtener los coeficientes  $a$  y  $b$  que minimizan la suma residual de cuadrados (SCE), buscamos el mínimo haciendo derivadas parciales:

$$\frac{\partial SCE}{\partial a} = 0 \quad \Rightarrow \quad \sum y_i = n a + b \sum x_i$$

$$\frac{\partial SCE}{\partial b} = 0 \quad \Rightarrow \quad \sum y_i x_i = n a \sum x_i + b \sum x_i^2$$

Se obtiene un sistema de dos ecuaciones (ecuaciones normales o canónicas) con dos incógnitas ( $a$  y  $b$ ), que se podrían expresar en forma matricial. Más adelante, para obtener las desviaciones típicas, veremos que el cálculo matricial es más sencillo.

$$\begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} \sum y_i \\ \sum x_i y_i \end{pmatrix} \quad (H) = \frac{1}{\sigma^2} \begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix}$$

Otra ventaja de trabajar con ecuaciones matriciales, es que otros tipos de análisis de regresión (no lineal, o de más de dos variables) pueden resolverse más fácilmente con un formalismo matricial.

La matriz cuadrada ( $H_{ij}$ ) se define en función de la función  $W$ , como:

$$H_{ij} = - \frac{\partial^2 W}{\partial a_i \partial a_j} \quad \text{con } a_1 = a \quad \text{y} \quad a_2 = b$$

Y el determinante de la matriz es:

$$\Delta = [n \sum x_i^2 - \sum x_i \sum x_i] = n S_{xx} = n \sum (x_i - \bar{x})^2$$

Del sistema de ecuaciones normales se puede sacar información adicional:

$$\sum y_i = n a + b \sum x_i \Rightarrow \frac{\sum y_i}{n} = a + b \frac{\sum x_i}{n} \Rightarrow \bar{y} = a + b \bar{x}$$

$$\sum y_i = n a + b \sum x_i \Rightarrow \sum (y_i - (a + b x_i)) = \sum (y_i - \hat{y}_i) = \sum e_i = 0$$

$$\sum y_i x_i = n a \sum x_i + b \sum x_i^2 \Rightarrow \sum (y_i x_i - (a + b x_i) x_i) = \sum (y_i - \hat{y}_i) x_i = \sum e_i x_i = 0$$

Por tanto el punto la recta de ajuste pasa por el punto  $x$ -media,  $y$ -media. En las dos últimas ecuaciones, las últimas igualdades, que involucran al error o residuo ( $e_i$ ) nos serán de utilidad más adelante.

Ahora solucionamos el anterior sistema de dos ecuaciones normales, con dos incógnitas ( $a$  y  $b$ ). De la primera ecuación normal obtenemos:

$$a = \frac{\sum y_i}{n} - b \frac{\sum x_i}{n} = \bar{y} - b \bar{x}$$

Sustituyendo la anterior, en la segunda ecuación normal, para despejar  $b$ , obtenemos:

$$b = \frac{[\sum x_i y_i - \sum x_i \sum y_i / n]}{[\sum x_i^2 - \sum x_i \sum x_i / n]}$$

Ya tenemos la pendiente de la recta ( $b$ ) y la ordenada en el origen ( $a$ ) que es el corte con el eje “ $y$ ” de la recta de regresión. Nos queda por comprobar si existe realmente una correlación lineal, para ello nos quedan por calcular dos coeficientes importantes,  $r$  y  $R^2$ , que están estrechamente relacionados. Pero antes veamos algunas expresiones alternativas para los coeficientes  $a$  y  $b$ . Ocasionalmente aparecen, y aquí se muestran a efectos de inventario:

$$a = \frac{\sum y_i}{n} - b \frac{\sum x_i}{n} = \frac{\sum y_i - b \sum x_i}{n} = \bar{y} - b \bar{x} = \frac{[\sum y_i \sum x_i^2 - \sum x_i \sum x_i y_i]}{n [\sum x_i^2 - \sum x_i \sum x_i / n]} = \frac{\sum y_i}{n} - \frac{S_{xy}}{S_{xx}} \frac{\sum x_i}{n}$$

Aunque la pendiente de la recta ( $b$ ) ya se tiene calculada, también tiene fórmulas alternativas:

$$b = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{\sum (x_i - \bar{x}) y_i}{\sum (x_i - \bar{x})^2} = \frac{[\sum x_i y_i - \sum x_i \sum y_i / n]}{[\sum x_i^2 - \sum x_i \sum x_i / n]} = \frac{S_{xy}}{S_{xx}}$$

$$b = \frac{S_{xy}}{S_{xx}} = r \sqrt{\frac{S_{yy}}{S_{xx}}} = r \frac{s_y}{s_x} = \frac{s_{xy}}{s_x^2} = \frac{[\sum x_i y_i - a \sum x_i]}{\sum x_i^2}$$

De las expresiones mostradas para el coeficiente b, la tercera (o la cuarta) suele llamarse “fórmula rápida”, ya que es más fácil operar con números complicados o grandes conjuntos de datos, es más rápida y eficiente en la programación de rutinas para ordenador, y evita la propagación de errores en el redondeo de la media. La demostración de las últimas expresiones para el coeficiente b es muy simple, una vez que se hayan introducido los nuevos coeficientes que en ellas aparecen.

En las anteriores expresiones hemos utilizado algunas relaciones, que ahora explicamos. A veces aparecen en la literatura los coeficientes  $S_{xx}$   $S_{yy}$   $S_{xy}$ , y que incluimos en algunas fórmulas:

$$S_{xx} = \sum (x_i - \bar{x})^2 = [\sum x_i^2 - \sum x_i \sum x_i / n] \quad S_{yy} = \sum (y_i - \bar{y})^2 = [\sum y_i^2 - \sum y_i \sum y_i / n]$$

$$S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}) = [\sum x_i y_i - \sum x_i \sum y_i / n]$$

Las igualdades entre los dos últimos miembros de las anteriores ecuaciones son especialmente interesantes, pues son la diferencia básica entre las “fórmulas de docencia” y las “fórmulas rápidas”. También hemos utilizado las medias, y los parámetros  $s_{xy}$   $s_x$   $s_y$  que ahora definimos:

Media de  $x$  de la muestra

$$\bar{x} = \frac{\sum x_i}{n}$$

Media de  $y$  de la muestra

$$\bar{y} = \frac{\sum y_i}{n}$$

La media o promedio, son formas abreviadas de llamar a la media aritmética. Tienen las mismas unidades que su variable.

Varianza de  $x$  de la muestra

$$s_x^2 = \frac{\sum (x_i - \bar{x})^2}{(n - 1)} = \frac{[\sum x_i^2 - \sum x_i \sum x_i / n]}{(n - 1)}$$

Varianza de  $y$  de la muestra

$$s_y^2 = \frac{\sum (y_i - \bar{y})^2}{(n - 1)} = \frac{[\sum y_i^2 - \sum y_i \sum y_i / n]}{(n - 1)}$$

Desviación típica de  $x$  de la muestra

$$s_x = \sqrt{\frac{\sum (x_i - \bar{x})^2}{(n - 1)}} = \sqrt{\frac{[\sum x_i^2 - \sum x_i \sum x_i / n]}{(n - 1)}}$$

Desviación típica de  $y$  de la muestra

$$s_y = \sqrt{\frac{\sum (y_i - \bar{y})^2}{(n - 1)}} = \sqrt{\frac{[\sum y_i^2 - \sum y_i \sum y_i / n]}{(n - 1)}}$$

La varianza  $s^2$  es un indicador de la dispersión de los datos respecto al valor medio de la variable, y tiene unidades de su variable elevado al cuadrado. Para expresar la dispersión de la variable en las unidades de ésta, se utiliza la desviación típica o desviación estándar  $s$ .

A veces se llama Var a la varianza. La desviación estándar (o típica) también suele nombrarse como SD o como DE. En las calculadoras suele denominarse a  $s$  (desviación típica de la muestra) como  $\sigma_{n-1}$ , y a la desviación típica de la población como  $\sigma_n$ .

A veces para comparar la dispersión de los datos, entre distintas muestras con medias muy distintas, es necesario definir un coeficiente de variación (de Pearson):

$$CV = \frac{s_x}{\bar{x}}$$

Covarianza de  $x$  e  $y$  de la muestra

$$s_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{(n - 1)} = \frac{[\sum x_i y_i - \sum x_i \sum y_i / n]}{(n - 1)}$$

La covarianza permite cuantificar numéricamente la relación entre las variables  $x$  e  $y$ . Si la función es creciente,  $s_{xy}$  es positivo (los dos paréntesis del numerador son positivos o negativos simultáneamente). Si la función es decreciente,  $s_{xy}$  es negativa. Si existe una relación muy fuerte entre las variables  $x$  e  $y$  entonces el módulo de  $s_{xy}$  es muy grande, y finalmente, si no existiera relación alguna entre las dos variables  $s_{xy}$  sería 0.

Las varianzas, desviaciones típicas, y covarianzas, se han expresado de dos maneras distintas. A la segunda suele llamarse “fórmula rápida” o “fórmula abreviada”, ya que es más fácil operar con números complicados o grandes conjuntos de datos, es más rápida y eficiente en la programación de rutinas para ordenador, y evita la propagación de errores en el redondeo de la media. La primera usualmente llamada “fórmula de desviaciones” suele usarse en ámbito de la docencia, para resaltar el concepto de medida de variación.

Con estas definiciones:

$$s_x^2 = \frac{S_{xx}}{(n - 1)} \quad s_y^2 = \frac{S_{yy}}{(n - 1)} \quad s_{xy} = \frac{S_{xy}}{(n - 1)}$$

Como acabamos de ver, la covarianza nos mide la correlación entre las variables  $x$  e  $y$ , pero tiene el problema de que puede llegar a ser muy grande en valor absoluto, no es comparable como medida de correlación al cambiar el experimento, el factor de escala, o simplemente las unidades. Interesa definir un coeficiente de correlación lineal simple, que no esté afectado por problemas de escala o de unidades escogidas para las variables:

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

El coeficiente de correlación lineal simple ( $r$ ) es una medida adimensional de la intensidad de la relación lineal entre las variables  $x$  e  $y$ . Toma valores comprendidos entre -1 y +1. Cuando  $r$  está próximo a sus extremos indica una fuerte relación lineal (que sería perfecta en los casos de  $r=-1$  o  $r=+1$ ). Cuando  $r$  está próxima a 0, indica una débil relación lineal. Finalmente si  $r$  es positivo indica relación lineal creciente, y decreciente para  $r$  negativo).

Coeficiente de correlación lineal simple  $r$

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} = \frac{[\sum x_i y_i - \sum x_i \sum y_i / n]}{\sqrt{[\sum x_i^2 - \sum x_i \sum x_i / n][\sum y_i^2 - \sum y_i \sum y_i / n]}} = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$$

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = \frac{s_{xy}}{s_x s_y} = b \sqrt{\frac{S_{xx}}{S_{yy}}} = b \frac{s_x}{s_y}$$

De las expresiones mostradas para  $r$ , la segunda (o la tercera) suele llamarse “fórmula rápida”, ya que es más fácil operar con números complicados o grandes conjuntos de datos, es más rápida y eficiente en la programación de rutinas para ordenador, y evita la propagación de errores en el redondeo de la media.

A veces se le llama a  $r$  “momento producto de Pearson” o “coeficiente de correlación lineal de Pearson”. Suele omitirse normalmente el adjetivo de simple, que se refiere a sólo dos variables.

Hay que advertir que si  $r$  es cero, no significa que no exista correlación entre las variables  $x$  e  $y$  sólo que no existe relación lineal, pudiera no existir ninguna relación, o por ejemplo que la relación fuese no lineal (p.ej. parabólica).

Antes de meternos en la definición del coeficiente de determinación  $R^2$ , demostraremos una relación que será útil para ver la equivalencia de dos definiciones distintas de  $R^2$ .

En la bibliografía, suelen aparecer distintas denominaciones:

$SCT = SST = S_{yy} =$  Variación total (Total variation)

$SCR = SSR = SS_{REG} =$  Variación explicada por el modelo de regresión lineal (Explained variation) = Suma de cuadrados de la regresión.

$SCE = SSE = SS_{RES} =$  Variación no explicada por el modelo de regresión lineal (Unexplained variation) = Suma de cuadrados de residuos.

$$\begin{aligned} SCT &= SCR + SCE \\ \Sigma(y_i - \bar{y})^2 &= \Sigma(\hat{y}_i - \bar{y})^2 + \Sigma(e_i - \bar{e})^2 \end{aligned} \quad (41)$$

SCT, también llamada SST, es la “suma total de cuadrados” y representa la variabilidad total de la variable dependiente  $y$ . SCR, también llamada SSR, es la “suma de regresión de cuadrados” y representa la variabilidad de la variable dependiente que puede ser explicada por el modelo de regresión lineal. SCE, también llamada SSE, es la “suma residual de cuadrados” o “suma de cuadrados de error” y representa la variabilidad de la variable dependiente que no puede ser explicada por el ajuste de regresión lineal.

$$SST = SCT = S_{yy}$$

$$SSR = SCR = b S_{xy} = b^2 S_{xx} = \frac{S_{xy}^2}{S_{xx}}$$

$$SSE = SCE = SST - SSR = \left[ S_{yy} - \frac{S_{xy}^2}{S_{xx}} \right]$$

SST representa la variación total de los datos  $y$  obtenidos experimentalmente; una parte de esa variación está justificada por la recta de regresión (SSR), y el resto (SSE) es variación no explicada por el modelo de regresión lineal (por errores aleatorios, o por variables no controladas).

Además,  $\hat{y}(x_i)$  es el valor de la variable  $y$  estimado por el modelo lineal:

$$\hat{y} = \hat{y}(x_i) = a + b x_i$$

Y realmente la ecuación 41 se cumple, puesto que con unas pocas operaciones, y recordando que las siguientes relaciones se cumplen:

$$\begin{aligned}\Sigma y_i &= \Sigma \hat{y}_i \Rightarrow \bar{y}_i = \bar{\hat{y}}_i \\ \Sigma e_i &= 0 \Rightarrow \bar{e} = 0 \Rightarrow \Sigma(e_i - \bar{e})^2 = \Sigma e_i^2 \\ \Sigma e_i &= 0 \Rightarrow \Sigma e_i \bar{y} = 0 \\ \Sigma e_i x_i &= 0 \Rightarrow \Sigma e_i \hat{y}_i = \Sigma e_i (a + b x_i) = a \Sigma e_i + b \Sigma e_i x_i = 0 + 0 = 0\end{aligned}$$

Además necesitaremos más adelante, la “varianza residual”:

$$s_e^2 = \frac{\Sigma e_i^2}{(n - 2)} = \frac{\Sigma(e_i - \bar{e})^2}{(n - 2)} = \frac{SCE}{(n - 2)}$$

La desviación típica de los errores en la muestra ( $s_e$ ) es la mejor estimación de la desviación típica (o estándar) de los errores en la población, y tiene (n-2) grados de libertad (n = número de datos) en el análisis de regresión lineal con dos coeficientes (a, b).

El coeficiente de determinación ( $R^2$ ) sirve para analizar si la función lineal  $y=a+bx$  procedente del análisis de regresión lineal simple, representa correctamente la nube de puntos experimentales ( $x_i, y_i$ ).  $R^2$  puede definirse como el cociente entre la varianza de la variable dependiente estimada y la varianza de la variable dependiente ( $s_y^2$ ).

$$R^2 = \frac{SCR}{SCT}$$

De tal forma que si las varianzas de y-estimado ( $\hat{y}$ ) y de la variable dependiente (y) fueran exactamente iguales, significaría un ajuste perfecto, y  $R^2$  sería igual a uno. Ya que  $R^2$  es un cociente de varianzas, siempre es positiva.

Otra forma de ver esta definición de  $R^2$  es que si casi toda la variabilidad de la variable dependiente (SCT) se puede explicar por el ajuste lineal, significa que SCR será casi igual a SCT y por tanto  $R^2$  será casi igual a uno.

$R^2$  también puede definirse como la unidad menos el cociente de la varianza del error ( $s_e^2$ ) y la varianza de la variable dependiente ( $s_y^2$ ).

$$R^2 = 1 - \frac{SCE}{SCT}$$

(Recordar de antes que  $SCT=SCR+SCE$  por lo tanto es una definición equivalente a la anterior)

De tal forma que si la varianza del error ( $s_e^2$ ), fuera muy pequeña (buen ajuste) el coeficiente  $R^2$  tendería a uno menos un cociente, que tiende a cero cuando el error de la recta de ajuste desaparece, es decir  $R^2$  tendería a uno. Si la varianza del error fuese tan grande como la de la varianza de la variable dependiente (y), (ajuste pésimo), entonces  $R^2$  tendería a 0.

Como  $R^2$  es 1 menos un cociente de varianzas, eso hará que  $R^2$  sea siempre menor que 1.

Otra forma de ver esta definición es que si  $R^2$  es casi igual a uno entonces SCE, es decir la variabilidad no explicada por el ajuste lineal, es casi cero, lo cual significa que casi toda la variabilidad de la variable dependiente se puede explicar por el ajuste lineal.

Finalmente si  $R^2$  fuese 0,75, significaría que el 75% del valor de los  $y_i$  es influenciado por los  $x_i$ , y el 25% restante es debido a otras causas.  $R^2 \times 100$  es el porcentaje de variaciones observadas de la variable dependiente (y) que quedan explicadas por el modelo de estimación lineal.



Las dos definiciones vistas para  $R^2$  son equivalentes, (utilizando 41):

$$R^2 = \frac{SCR}{SCT} = \frac{SCT - SCE}{SCT} = 1 - \frac{SCE}{SCT} = 1 - \frac{\sum e_i^2}{\sum (y_i - \bar{y})^2} = 1 - \frac{\sum (y_i - (a + bx_i))^2}{\sum (y_i - \bar{y})^2}$$

Operando sobre  $R^2$  (Coeficiente de determinación) se obtiene la fórmula más habitual (la primera de la izquierda):

$$R^2 = \frac{[\sum (x_i - \bar{x})(y_i - \bar{y})]^2}{[\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2]} = \frac{[\sum x_i y_i - \sum x_i \sum y_i / n]^2}{[\sum x_i^2 - \sum x_i \sum x_i / n][\sum y_i^2 - \sum y_i \sum y_i / n]} = \frac{S_{xy}^2}{S_{xx} S_{yy}}$$

De las tres expresiones mostradas para  $R^2$ , la segunda (y la tercera) suele llamarse “fórmula rápida”, por las mismas razones ya mencionadas anteriormente. También puede expresarse  $R^2$  como:

$$R^2 = \frac{s_{xy}^2}{s_x^2 s_y^2} = b^2 \frac{s_x^2}{s_y^2} = \frac{S_{xy}^2}{S_{xx} S_{yy}} = b \frac{S_{xy}}{S_{yy}} = b^2 \frac{S_{xx}}{S_{yy}} = b^2 \frac{\sum (x_i - \bar{x})^2}{\sum (y_i - \bar{y})^2} = b^2 \frac{[\sum x_i^2 - \sum x_i \sum x_i / n]}{[\sum y_i^2 - \sum y_i \sum y_i / n]}$$

$$R^2 = \frac{S_{xy}^2}{S_{xx} S_{yy}} = b \frac{S_{xy}}{S_{yy}} = \frac{SCR}{S_{yy}} = \frac{SCR}{SCT} = b^2 \frac{S_{xx}}{S_{yy}}$$

### 1.1.2. Cálculo de las desviaciones típicas y covarianzas

Hay que recordar que los valores obtenidos de los coeficientes del ajuste lineal ( $a$  y  $b$ ) son sólo una estimación de los valores reales. Los valores reales son desconocidos, y sólo se puede decir que el resultado obtenido experimentalmente es más o menos significativo. Los resultados obtenidos mediante experimento para  $a$  y  $b$  serán muy significativos, si tienen la propiedad de que al repetir muchas veces el experimento tenemos más posibilidades de que estén cerca de los  $a$  y  $b$  reales (desconocidos).

En el análisis de regresión lineal simple, se tienen como grados de libertad, el número de datos experimentales, menos 2 restricciones (las del sistema de ecuaciones del que se despejan  $a$  y  $b$ ). Antes de meternos en el cálculo de intervalos y nivel de confianza, es preferible calcular los “errores estándar” de  $a$ ,  $b$  y algunos otros. Primero obtendremos el error estándar  $s_e$  que es el mejor estimado que tendremos del error de la recta de regresión. Luego, deduciremos las desviación típicas (o error estándar) de los coeficientes  $a$  y  $b$  ( $s_a$  y  $s_b$ ) de la regresión lineal que serán las raíces de las respectivas varianzas. Como uno de los objetivos de la recta de regresión es predecir el valor de la variable dependiente  $y$ , lo que haremos será obtener el error estándar para la media predicha por la recta de regresión, para una predicción individual, y para la media de un número  $q$  de predicciones.

Una forma de obtener estos errores es utilizando el operador varianza  $V()$ . La propiedad más útil de este operador es:

$$V(ax - by) = a^2 V(x) + b^2 V(y)$$

donde  $a$  y  $b$  serían constantes, y las variables  $x$  e  $y$  no están correlacionadas, es decir su covarianza es cero.

### - Obtención de $s_e$

$s_e$  es la desviación típica del error (muestral), también llamado error estándar de la regresión lineal:

$$s_e^2 = \frac{\Sigma(e_i - \bar{e})^2}{(n - 2)} = \frac{SCE}{(n - 2)} = \frac{S_{yy} - b S_{xy}}{(n - 2)} = \frac{S_{yy} - b^2 S_{xx}}{(n - 2)} = \frac{S_{yy} [1 - R^2]}{(n - 2)} = \frac{1}{n - 2} \left[ S_{yy} - \frac{S_{xy}^2}{S_{xx}} \right]$$

En realidad  $s_e^2$  tendría más formas alternativas para expresarse, ya que SCE puede calcularse como:

$$SCE = \Sigma e_i^2 = \Sigma (y_i - \hat{y}_i)^2 = \Sigma (y_i - y(x_i))^2 = \Sigma (y_i - (a + b x_i))^2 = \Sigma [(y_i - \bar{y}) - b(x_i - \bar{x})]^2$$

y también:

$$SCE = S_{yy} - 2 b S_{xy} + b^2 S_{xx} = SCT - 2 SCR + SCR = SCT - SCR$$

### - Obtención de $s_b$

$s_b$  es la desviación típica (o error estándar) del coeficiente  $b$ .  $s_b$  es la raíz de la varianza  $s_b^2$ . Obtención de  $s_b^2 = V(b)$  aplicando el operador varianza  $V()$ . Partimos del coeficiente  $b$ :

$$b = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2} = \frac{\Sigma(x_i - \bar{x})y_i}{\Sigma(x_i - \bar{x})^2}$$

y aplicamos el operador varianza:

$$s_b^2 = V(b) = V\left(\frac{\Sigma(x_i - \bar{x})y_i}{\Sigma(x_i - \bar{x})^2}\right) = \frac{\Sigma(x_i - \bar{x})^2}{[\Sigma(x_i - \bar{x})^2]^2} V(y_i) = \frac{1}{\Sigma(x_i - \bar{x})^2} s_e^2 = \frac{s_e^2}{S_{xx}}$$

Hemos utilizado la propiedad (que es válida si todas las  $y_i$  tienen la misma varianza  $V(y)$ , y la covarianza entre cada par de ellas es cero):

$$V(b) = V(\Sigma a_i y_i) = \Sigma a_i^2 V(y)$$

$$s_b^2 = \frac{s_e^2}{\Sigma(x_i - \bar{x})^2} = \frac{s_e^2}{[\Sigma x_i^2 - \Sigma x_i \Sigma x_i / n]} = \frac{s_e^2}{S_{xx}} = \frac{1}{n-2} \left[ \frac{S_{yy}}{S_{xx}} - b^2 \right]$$

$$s_b^2 = \frac{1}{n-2} \frac{S_{yy}}{S_{xx}} [1 - R^2] = \frac{1}{n-2} \left[ \frac{S_{yy}}{S_{xx}} - \frac{S_{xy}^2}{S_{xx}^2} \right] = \frac{1}{n-2} \left[ \frac{S_{yy}}{S_{xx}} - b \frac{S_{xy}}{S_{xx}} \right]$$

### - Obtención de $s_a$

$s_a$  es la desviación típica (o error estándar) del coeficiente  $a$ .  $s_a$  es la raíz de la varianza  $s_a^2$ . Obtención de  $s_a^2 = V(a)$  aplicando el operador varianza  $V()$ . Partimos del coeficiente  $a$ :

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

y aplicamos el operador varianza:

$$s_a^2 = V(a) = V(\bar{y}) + V(b\bar{x}) = V(\bar{y}) + \bar{x}^2 V(b) = \frac{V(y)}{n} + \bar{x}^2 \frac{V(y)}{S_{xx}} = s_e^2 \left[ \frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right]$$

ya que la covarianza entre el coeficiente  $b$  y la media de  $y$  es cero.

$$s_a^2 = s_e^2 \frac{\sum x_i^2 / n}{\sum (x_i - \bar{x})^2} = s_e^2 \frac{\sum x_i^2 / n}{[\sum x_i^2 - \sum x_i \sum x_i / n]} = s_e^2 \frac{\sum x_i^2 / n}{S_{xx}}$$

$$s_a^2 = s_e^2 \left[ \frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2} \right] = s_e^2 \left[ \frac{1}{n} + \frac{\bar{x}^2}{[\sum x_i^2 - \sum x_i \sum x_i / n]} \right] = s_e^2 \left[ \frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right]$$

$$s_a^2 = s_b^2 \sum x_i^2 / n = \frac{\sum x_i^2 / n}{n-2} \left[ \frac{S_{yy}}{S_{xx}} - \frac{S_{xy}^2}{S_{xx}^2} \right] = s_b^2 \left[ \frac{S_{xx}}{n} + \bar{x}^2 \right]$$

$s_a$  y  $s_b$  son estimadores muestrales de las desviaciones típicas de la población. También suelen llamarse “Error estándar”, y se representan como  $SE(a)$  y  $SE(b)$  respectivamente.

### - Obtención de la covarianza entre $a$ y $b$ ( $s_{a,b}$ )

Como sabemos que el coeficiente  $a$  es:

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

la covarianza entre  $a$  y  $b$ , notada como  $cov(a,b)$ , se puede calcular utilizando que la covarianza entre la  $y$  media y  $b$  es cero:

$$s_{a,b} = Cov(a,b) = Cov(\bar{y} - b\bar{x}, b) = -\bar{x} Cov(b,b) = -\bar{x} V(b) = -\bar{x} s_b^2$$

$$s_{a,b} = Cov(a,b) = -\bar{x} \frac{s_e^2}{\sum (x_i - \bar{x})^2} = -\bar{x} \frac{s_e^2}{S_{xx}} = -s_e^2 \frac{\sum x_i / n}{S_{xx}}$$

**- Obtención de  $s_a$ ,  $s_b$  y la covarianza  $s_{a,b}$  matricialmente.**

Al plantear las ecuaciones normales, cuando hallamos los coeficientes  $a$  y  $b$ , introducimos la matriz  $H_{ij}$ :

$$H_{ij} = - \frac{\partial^2 W}{\partial a_i \partial a_j} \quad \text{con } a_1 = a \quad y \quad a_2 = b \quad (H) = \frac{1}{\sigma^2} \begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix}$$

con el determinante de la matriz igual a:

$$\Delta = [n \sum x_i^2 - \sum x_i \sum x_i] = n S_{xx} = n \sum (x_i - \bar{x})^2$$

La matriz de varianzas ( $\sigma_{ij}^2$ ) se puede calcular, teniendo en cuenta:

$$(H_{ij})(\sigma_{ij}^2) = 1 \quad \rightarrow \quad (\sigma_{ij}^2) = (H_{ij})^{-1}$$

$$(\sigma_{ij}^2) = (H_{ij})^{-1} = \frac{\sigma^2}{\Delta} \begin{pmatrix} \sum x_i^2 & - \sum x_i \\ - \sum x_i & n \end{pmatrix} = \begin{pmatrix} \sigma_a^2 & \sigma_{a,b} \\ \sigma_{a,b} & \sigma_b^2 \end{pmatrix}$$

Como tenemos datos de una muestra, no de la población total, debemos sustituir cada varianza ( $\sigma_a^2$ ,  $\sigma_b^2$ ,  $\sigma_{a,b}$ ,  $\sigma^2$ ) por su mejor estimado ( $s_a^2$ ,  $s_b^2$ ,  $s_{a,b}$ ,  $s_e^2$ ), y tras sustituir el valor de  $\Delta$  obtendríamos las mismas fórmulas presentadas anteriormente, pero con un cálculo más simple, y que además se puede extender fácilmente a análisis de regresión no lineal, o con más de dos coeficientes.

**- Obtención de diversos errores para valores predichos por la recta de regresión.**

La recta de regresión  $y=a+bx$ , permite predecir un valor de la variable dependiente para un  $x_0$  concreto, es decir que serviría para prever el valor de la ordenada.

Error estándar para la predicción por la recta de regresión ( $y=a+bx$ ), sería el error estándar de la media de varias predicciones cuando  $x = x_0$ :

$$y = a + b x \quad ; \quad a = \bar{y} - b \bar{x} \quad \Rightarrow \quad \bar{y}_0 = \bar{y} - b(x_0 - \bar{x})$$

$$s_{\bar{y}_0}^2 = V(\bar{y}) + V(b(x_0 - \bar{x})) = \frac{V(y)}{n} + (x_0 - \bar{x})^2 V(b) = V(y) \left[ \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right]$$

$$s_{\bar{y}_0}^2 = s_e^2 \left[ \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right] = s_e^2 \left[ \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{[\sum x_i^2 - \sum x_i \sum x_i / n]} \right] = s_e^2 \left[ \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right]$$

$$s_{\bar{y}_0}^2 = s_e^2 \left[ \frac{1}{n} + \frac{(x_0 - \bar{x})(y_0 - \bar{y})}{\sum (x_i - \bar{x})(y_i - \bar{y})} \right] = (s_a^2 + x_0^2 s_b^2 - 2 x_0 \bar{x} s_b^2)$$

De la aplicación del error estándar para la predicción de la recta de regresión, se observa que depende del punto en el que se haga la predicción ( $x_0$ ). Cuando  $x_0$  se aleja de la media de  $x$ , el error estándar crece, de tal forma que si se representara gráficamente el error, saldría una banda que se abre en los extremos de la recta de regresión. A esa banda se le llama la “banda de confianza de la recta de regresión”, en el que su grosor depende naturalmente del nivel de confianza que se escoja.

Es fácil comprobar ahora, que el error estándar para la media de varias predicciones cuando  $x = 0$ : es igual al error estándar para el coeficiente  $a$ , lo cual es lógico, pues  $y=a+bx = y_{medio}=a_{medio}+b*0$ .

Por otro lado, podemos estar interesados en el intervalo de predicción, es decir el margen en el cual pueda aparecer un  $y_0$  individual (que no puede prever, por su naturaleza, la recta de regresión). Para calcular el error estándar en una sola predicción individual, se añade un término de error que es justamente la variabilidad de la ordenada  $y$  :  $V(y)=s_e^2$

Error estándar para una predicción individual para  $x = x_0$ :

$$s_{y_0}^2 = s_e^2 \left[ 1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\Sigma(x_i - \bar{x})^2} \right] = s_e^2 \left[ 1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{[\Sigma x_i^2 - \Sigma x_i \Sigma x_i / n]} \right] = s_e^2 \left[ 1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right]$$

$$s_{y_0}^2 = s_e^2 \left[ 1 + \frac{1}{n} + \frac{(x_0 - \bar{x})(y_0 - \bar{y})}{\Sigma(x_i - \bar{x})(y_i - \bar{y})} \right] = (s_e^2 + s_a^2 + x_0^2 s_b^2 - 2 x_0 \bar{x} s_b^2)$$

A la banda de error para una predicción individual se les llama “banda de tolerancia” o “banda de predicción”, en el que su grosor depende del nivel de confianza que se escoja.

Si queremos calcular cómo se reduce el error estándar para la media de  $q$  predicciones individuales para un  $x=x_0$ , las expresiones quedan:

$$s_{y_0}^2 = s_e^2 \left[ \frac{1}{q} + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\Sigma(x_i - \bar{x})^2} \right] = s_e^2 \left[ \frac{1}{q} + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{[\Sigma x_i^2 - \Sigma x_i \Sigma x_i / n]} \right] = s_e^2 \left[ \frac{1}{q} + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right]$$

$$s_{y_0}^2 = s_e^2 \left[ \frac{1}{q} + \frac{1}{n} + \frac{(x_0 - \bar{x})(y_0 - \bar{y})}{\Sigma(x_i - \bar{x})(y_i - \bar{y})} \right] = \left( \frac{s_e^2}{q} + s_a^2 + x_0^2 s_b^2 - 2 x_0 \bar{x} s_b^2 \right)$$

Reduciendose el error cuanto mayor sea el número de predicciones a las que se le hace la media.

### 1.1.3. Confianza en las estimaciones del análisis de regresión

Ya tenemos las expresiones del coeficiente de correlación ( $r$ ), de los coeficientes del ajuste lineal ( $a$  y  $b$  de  $y=a+bx$ ). Podría pensarse que se pueden aplicar las fórmulas y se tiene ya hecho el ajuste de regresión lineal. No es así, ya que hay que recordar que  $r$ ,  $a$  y  $b$  se han obtenido experimentalmente y están sometidos a un error, sabemos que  $r$ ,  $a$  y  $b$  son estimaciones de los valores  $r$ ,  $a$  y  $b$  teóricos de la población (de las infinitas medidas posibles). A esos valores  $r$ ,  $a$  y  $b$  teóricos les llamaremos  $\rho$ ,  $\alpha$  y  $\beta$  respectivamente, y en principio son desconocidos, pero en este apartado veremos que podemos al menos saber con qué nivel de confianza los valores experimentales  $r$ ,  $a$  y  $b$  están cercanos a  $\rho$ ,  $\alpha$  y  $\beta$ .

El nivel de confianza lo llamaremos  $(1-\alpha)$  ( $\alpha$  como nivel de significancia, que no tiene nada que ver con el  $\alpha = a_{\text{teórico}}$ ) y se suele expresar en porcentaje. Un nivel de confianza de 95% (0,95) indica que la posibilidad de que nuestras medidas experimentales estén “engañándonos” (por azar) es de sólo un 5%. Personas más exigentes querrán reducir el riesgo al 1%, en ese caso sólo admitirán resultados con un nivel de confianza del 99%. No es recomendable trabajar con niveles de confianza menores de 95%, de hecho desde hace casi un siglo, los trabajos con  $\alpha$  menor de 0,95 se suelen rechazar en las publicaciones. También se puede trabajar con el nivel de significancia ( $\alpha$ ), así para  $(1-\alpha)=0,95$ , el nivel de significancia será 0,05.

Para garantizar la “confianza” de los resultados experimentales de  $r$ ,  $a$  y  $b$  pueden utilizarse varios métodos:

- Método de prueba de hipótesis (versión tradicional, bastan las tablas)
- Método de prueba de hipótesis de Valor-P (versión más actual, necesita ordenador)
- Método de intervalos de confianza

El método de intervalos de confianza parece más simple, pero es algo más difícil de interpretar correctamente; nosotros lo usaremos con los coeficientes  $a$  y  $b$ . El método de prueba de hipótesis lo usaremos para el coeficiente  $r$ , en concreto para demostrar que existe realmente una relación lineal entre la variable dependiente  $y$  y la variable independiente  $x$ . El intervalo de confianza para el coeficiente  $r$ , es bastante más complicado, y por eso no se suele emplear.

Supondremos en las explicaciones que:

La muestra puede ser grande o pequeña. Si es grande ( $n>30$ ), habría que usar la distribución normal, si es pequeña ( $n\leq 30$ ) habría que usar la distribución t-Student. Pero como t-Student con muestras grandes ( $n>30$ ) es casi igual a la distribución normal estándar ( $z$ ), usaremos siempre la t-Student.

Como la desviaciones típicas de la población serán desconocidas (ya que sólo tomamos unas pocas medidas experimentales de las infinitas posibles), trabajaremos con desviaciones típicas de muestra (con muestras grandes a veces hay gente que prefiere usar t-student si hace esa sustitución).

La distribución de la población es normal o aproximadamente normal. Esto requeriría estrictamente utilizar los datos experimentales de la muestra y representarlos gráficamente para ver si se distribuyen en forma de campana de Gauss, pero nosotros supondremos que es así sin comprobarlo, especialmente si el número de datos experimentales es muy grande (por teorema del límite central al final la media se va distribuyendo normalmente)

## - Confianza en el coeficiente de determinación $r$ : Test de hipótesis.

Usaremos el método de prueba de hipótesis, primero en versión tradicional, y luego veremos el significado y el cálculo del Valor-P en un paso adicional.

PASOS:

### 1. Definición de las hipótesis

Se parte de la suposición inicial de que no tenemos nada, de que no existe correlación lineal entre las variables  $x$  e  $y$ . es decir que el  $r$  teórico, el de la población, que desconocemos,  $r_{\text{población}}$  es cero (abreviaremos  $r_{\text{población}}$  como  $\rho$ ). Intentaremos demostrar que esa hipótesis es falsa, y si lo conseguimos habremos demostrado que según nuestros datos experimentales existe correlación lineal entre las variables  $x$  e  $y$ .

Hipótesis:

$H_0$ :	$\rho = 0$	No existe correlación lineal	"hipótesis nula"
$H_1$ :	$\rho \neq 0$	Existe correlación lineal	"hipótesis alternativa"

### 2. Escoger el nivel de confianza $(1 - \alpha)$ o el nivel de significancia $(\alpha)$ y cálculo del t-crítico.

No escoger  $1 - \alpha$  menores de 0,95, o lo que es lo mismo  $\alpha$  mayores de 0,05. Y luego, buscando en las tablas t-Student de los libros de estadística, podremos encontrar el t-crítico, para rechazar la hipótesis  $H_0$  y por tanto para aceptar  $H_1$ , es decir demostrar que realmente existe correlación lineal.

### 3. Aplicar la estadística de prueba (o el test-estadístico)

Se aplica con el valor de coeficiente de determinación teórico ( $\rho$ ) con el valor que se da en la hipótesis nula ( $H_0$ ), es decir  $\rho=0$ . Se obtiene para  $r$  experimental, el "valor-t" del experimento.

Estadística de prueba (valor-t):

$$t = \frac{r - \rho}{s_r} \quad ; \quad s_r = \sqrt{\frac{(1 - r^2)(1 - \rho^2)}{n - 2}}$$

Si  $\rho=0$  nos queda

$$t = \frac{r}{\sqrt{\frac{1 - r^2}{n - 2}}} = r \sqrt{\frac{n - 2}{1 - r^2}} = \frac{b}{s_e / \sqrt{S_{xx}}} = \frac{b}{s_b} = \frac{\sqrt{SSR}}{s_e}$$

Es la segunda expresión, la más utilizada para el caso  $\rho=0$ , cuando se quiere calcular el test estadístico. Para los que usan el análisis de la varianza, se suele utilizar el estadístico F de Snedecor, que (para dos grupos) queda relacionado con el valor-t del experimento calculado anteriormente:

$$F = \frac{MS_R}{MS_E} = \frac{MS_R}{s_e^2} = \frac{SSR/1}{SSE/(n - 2)} = \frac{S_{xy}^2}{S_{xx} [S_{yy} - S_{xy}^2/S_{xx}]} \frac{(n - 2)}{(n - 2)} = \frac{r^2}{1 - r^2} (n - 2)$$

$$F = \frac{SSR/1}{SSE/(n - 2)} = \frac{SSR}{s_e^2} = \frac{b^2 S_{xx}}{s_e^2} = \frac{b^2}{s_b^2} = \frac{(b - 0)^2}{s_b^2} = t_b^2 = \left( r \sqrt{\frac{n - 2}{1 - r^2}} \right)^2$$

#### 4. Comprobación de las hipótesis.

- Si el valor absoluto del t obtenido de la estadística de prueba (valor-t) es mayor que el t-crítico (en valor absoluto) se rechaza la hipótesis nula ( $H_0$ ) y se puede decir que:

“Con un nivel de confianza de  $(1-\alpha)\%$ , o con un nivel de significancia de  $\alpha$ , los datos experimentales nos dan suficientes indicios para apoyar  $H_1$ , es decir que existe correlación lineal entre la variable dependiente y la dependiente. La probabilidad de que sin existir realmente correlación lineal, hubiesemos tenido estos datos (que nos llevan a decir que existe) es de menos de un  $\alpha\%$ ”

- Si el t obtenido de la estadística de prueba (valor-t) es menor que el t-crítico no se puede rechazar la hipótesis nula ( $H_0$ ), lo cual no significa que  $H_0$  sea cierta, solamente que no se puede demostrar. Podríamos decir que:

“Con un nivel de confianza de  $(1-\alpha)\%$ , o con un nivel de significancia de  $\alpha$ , los datos experimentales son insuficientes para rechazar  $H_0$ , es decir en ausencia de otras pruebas supondremos que no existe (aunque pudiera existir) correlación lineal entre la variable dependiente y la dependiente”

#### 5. Paso extra para calcular el “valor-P”

Siempre se han tenido tablas con el valor de la abscisa de la distribución t-Student en función de la probabilidad. Con la actual disposición de ordenadores y programas especializados es posible calcular la probabilidad de una cola de la distribución t en función de la abscisa correspondiente, por eso cada vez más, los trabajos suelen incluir el “valor-P” de los datos experimentales.

El experimento aporta indicios más fuertes para rechazar la hipótesis nula  $H_0$  y por tanto apoyar la hipótesis alternativa  $H_1$ , cuanto menor sea el valor-P. Como valores orientativos:

valor-P = 0,01 → Indicios muy fuertes en contra de  $H_0$  y por tanto a favor de  $H_1$ .

valor-P = 0,025 → Indicios fuertes en contra de  $H_0$  y por tanto a favor de  $H_1$ .

valor-P = 0,05 → Indicios razonablemente fuertes en contra de  $H_0$  y por tanto a favor de  $H_1$ .

No admitir nunca valores-P mayores que 0,05.

Una vez obtenido el coeficiente de determinación experimental (r), calculamos en el paso 3 el “valor-t” de nuestro experimento. Con el programa de ordenador adecuado puede calcularse la probabilidad de que siendo cierta  $H_0$  nos salga un valor-t como el del experimento. A la probabilidad de que salga una abscisa mayor que el valor-t se le llama “valor-p”.

#### *Ejemplo:*

Suponga que en laboratorio tuvo mucho cuidado al hacer las medidas experimentales, realizó 12 medidas, y calculó el coeficiente de determinación  $r=0,71$ . El compañero de al lado, tuvo menos cuidado en el trabajo de laboratorio, y sólo realizó 7 medidas, pero al calcular el coeficiente de determinación consiguió  $r=0,80$ . Un tercer compañero obtuvo con sólo 5 medidas un valor “muy bueno”  $r=0,87$

En principio podría pensarse que los demás tuvieron más suerte, ya que tuvo una mejor correlación lineal en sus datos experimentales (r es mayor). Veamos que no es así.

Paso 1 : Definición de hipótesis (las mismas de arriba)

Paso 2 : Escoger nivel de confianza y calcular t-crítico.

Escojo 95%, por tanto nivel de confianza  $(1-\alpha)=0,95$ , y  $\alpha=0,05$ . Por la forma de la hipótesis  $H_0$  se debe calcular el t-crítico con dos colas, ya que admitiré  $H_0$  ( $p=0$ ) si está cerca de  $p=0$  y la rechazaré si está lejos, es decir que puedo rechazar la hipótesis o bien porque  $p$  sea mucho menor que cero (cola izquierda), o bien porque  $p$  sea mucho mayor que cero (cola derecha).

Al ser un problema de 2 colas, busco en las tablas t-Student de  $\alpha/2=0,025$ . En nuestro caso, los grados de libertad de las tablas t-Student serían igual al número de muestras menos 2.



Los resultados para los tres casos se llevan en paralelo en la siguiente tabla.

	caso 1	caso 2	caso 3
número de muestras n	12	7	5
grados de libertad gl=(n-2)	10	5	3
t-crítico con gl=(n-2) y $\alpha/2=0,025$	2,571	2,228	3,182

Paso 3: Aplicar la estadística de prueba.

Con el r obtenido experimentalmente, se calcula el “valor-t”, que es la estadística t resultante de los datos experimentales. Para ello se utiliza la siguiente fórmula.

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = r \sqrt{\frac{n-2}{1-r^2}}$$

El valor-t resultante de los datos experimentales se compara con el t-crítico

	caso 1	caso 2	caso 3
t-crítico con gl=(n-2) y $\alpha/2=0,025$	2,571	2,228	3,182
r-experimental	0,71	0,80	0,87
valor-t obtenido a partir de r-experimental	3,188	2,981	3,056

#### 4. Comprobación de las hipótesis, y conclusiones.

Caso 1 y caso 2: “valor-t” (en valor absoluto) mayor que t-crítico

Por tanto se rechaza la hipótesis nula ( $H_0$ ) y se puede decir que:

“Con un nivel de confianza del 95%, o con un nivel de significancia del 5%, los datos experimentales nos dan suficientes indicios para apoyar  $H_1$ , es decir que existe correlación lineal entre la variable dependiente y la independiente. La probabilidad de que sin existir realmente correlación lineal, hubiésemos tenido estos datos (que nos llevan a decir que existe) es de menos de un 5%”.

Caso 3: “valor-t “ menor que t-crítico

Por tanto, no se puede rechazar la hipótesis nula ( $H_0$ ), lo cual no significa que  $H_0$  sea cierta, solamente que no se ha podido demostrar que fuera falsa. Podríamos decir que:

“Los datos experimentales obtenidos son insuficientes para rechazar  $H_0$ , es decir en ausencia de otras pruebas supondremos que no existe (aunque pudiera existir) correlación lineal entre la variable dependiente y la independiente”.

Es decir que aunque en el caso 3 se tuvo el mejor r, resulta que es insuficiente para probar que existe una relación lineal, por tanto, el trabajo de laboratorio al final no le sirvió de nada, y ni siquiera puede plantear la ecuación de regresión lineal.

#### 5. Paso adicional para calcular el “valor-P”=“probabilidad de significancia”

Con el valor-t obtenido en el paso 3, y utilizando p.ej. la función @TDIST de Quattro-Pro, con dos colas: @TDIST(grados\_libertad,número\_colas,valor-t) Obtendríamos el “valor-p” en los distintos casos.

	caso 1	caso 2	caso 3
valor-t obtenido a partir de r-experimental	3,188	2,981	3,056
valor-p obtenido a partir del valor-t	0,00968	0,03075	0,05516
Nivel de confianza de r-experimental: (1-valor-p)	99,03%	96,93%	94,48%

Como se puede observar la confianza en los datos experimentales, y por lo tanto en los resultados que de ahí se deduzcan, premia al caso 1 (que tomó 12 medidas experimentales). En cambio, en el caso 3, los resultados que pueda deducir de sus 5 medidas experimentales, no sirven para nada, pues el nivel de confianza no llegó al mínimo exigido de 95%.

La interpretación de los datos de la anterior tabla sería (caso 1): “Sólo en 9 de cada 1000 veces podría darse que saliera el resultado de que existe correlación lineal sin existir realmente”. En el caso 2, se diría que en 3 de cada 100 veces que podríamos fallar al decir “existe relación lineal”.

### - Confianza en el coeficiente de correlación r: Intervalo de confianza.

Calcular el intervalo de confianza del coeficiente de correlación r, es bastante complicado, y no se suele usar, aún así, se dan aquí las fórmulas.

La aproximación z de Fisher (z') sigue una distribución normal estándar con media  $\text{tgh}^{-1}(r)$  y varianza  $1/(n-3)^{1/2}$

$$z' = \frac{1}{2} \ln \left( \frac{1+r}{1-r} \right) = \text{tgh}^{-1}(r)$$

Por tanto el intervalo de confianza para  $\text{tgh}^{-1}(\rho)$  se construiría de la siguiente forma.

$$\begin{aligned} \text{tgh}^{-1}(\rho) &= \text{tgh}^{-1}(r) \pm z_{\alpha/2} \frac{1}{\sqrt{n-3}} \\ \frac{1}{2} \ln \left( \frac{1+\rho}{1-\rho} \right) &= \frac{1}{2} \ln \left( \frac{1+r}{1-r} \right) \pm z_{\alpha/2} \frac{1}{\sqrt{n-3}} \end{aligned}$$

El límite inferior ( $w_I$ ) y el límite superior ( $w_D$ ) quedarían de la siguiente forma.

$$\begin{aligned} w_I &= \frac{1}{2} \ln \left( \frac{1+r}{1-r} \right) - z_{\alpha/2} \frac{1}{\sqrt{n-3}} = \text{tgh}^{-1}(r) - z_{\alpha/2} \frac{1}{\sqrt{n-3}} \\ w_D &= \frac{1}{2} \ln \left( \frac{1+r}{1-r} \right) + z_{\alpha/2} \frac{1}{\sqrt{n-3}} = \text{tgh}^{-1}(r) + z_{\alpha/2} \frac{1}{\sqrt{n-3}} \end{aligned}$$

Por tanto, tomando tangente hiperbólica, el intervalo que acota el valor auténtico del coeficiente de correlación ( $\rho$ ) quedaría.

$$\frac{e^{2w_I} - 1}{e^{2w_I} + 1} < \rho < \frac{e^{2w_D} - 1}{e^{2w_D} + 1}$$

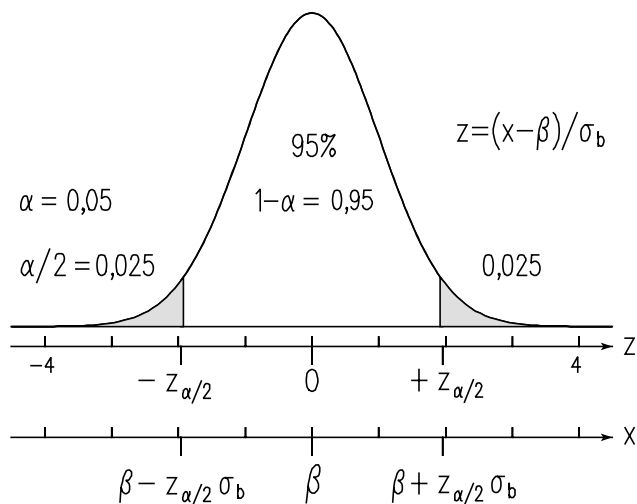
El test estadístico para z' (para probar hipótesis con un  $\rho=\rho_0$ ) quedaría:

$$z' = \frac{\text{tgh}^{-1}(r) - \text{tgh}^{-1}(\rho_0)}{1/\sqrt{n-3}} = \frac{\sqrt{n-3}}{2} \left[ \ln \left( \frac{1+r}{1-r} \right) - \ln \left( \frac{1+\rho_0}{1-\rho_0} \right) \right] = \frac{\sqrt{n-3}}{2} \ln \left[ \frac{(1+r)(1-\rho_0)}{(1-r)(1+\rho_0)} \right]$$

## - Confianza en los coeficientes a y b del análisis de regresión: Intervalos de confianza.

Supongamos que tenemos un coeficiente cuyo valor auténtico es  $\beta$  y queremos obtener experimentalmente su valor. Cualquier experimento está afectado por fuentes de error variadas y por tanto no obtendremos el valor auténtico  $\beta$ , sino una estimación  $b$ . Los intervalos de confianza nos dirán cuán bueno es el valor estimado  $b$ .

Supondremos que los errores experimentales (aleatorios o debido a variables no controladas) se distribuyen normalmente, es decir en campana de Gauss. En la experimentación, habitualmente utilizaremos la distribución t-Student, en vez de la distribución normal (de Gauss).



Si deseamos mostrar los datos experimentales con un nivel de confianza del 95%,  $(1-\alpha)$  será igual a 0,95, y el nivel de significancia ( $\alpha$ ) sería 0,05. El nivel de significancia indica la probabilidad de que nos alejemos del valor medio ( $\beta$ ), y como nos podemos alejar por arriba y por abajo, estaremos en un problema de dos colas (ver gráfica) y por tanto cada cola representaría una probabilidad  $\alpha/2 = 0,025$ . Buscando en las tablas de la distribución normal estándar encontraríamos que  $z_{\alpha/2} = 1,959\ 964 \approx 1,96$ .

La Estadística nos dice que si realizamos un experimento para obtener el valor de  $\beta$ , el 95% de los experimentos nos resultará un valor  $b$  comprendido en el intervalo  $(\beta - z_{\alpha/2} \sigma_b, \beta + z_{\alpha/2} \sigma_b)$ . Ese intervalo se puede escribir como  $\beta - z_{\alpha/2} \sigma_b < b < \beta + z_{\alpha/2} \sigma_b$  o bien como  $b = \beta \pm z_{\alpha/2} \sigma_b$ .

Pero este intervalo no nos resulta útil, ya que casi nunca conocemos el valor auténtico de  $\beta$ , los experimentos sólo nos dan estimaciones  $b$  del coeficiente  $\beta$ . Tras un experimento podemos tener el coeficiente  $b$ , pero no el valor auténtico  $\beta$  (que es el que queremos conocer). Por tanto despejamos del intervalo de confianza anterior (escrito en forma de inecuación) y obtenemos el intervalo de confianza para el coeficiente  $b$  experimental:

$$b - z_{\alpha/2} \sigma_b < \beta < b + z_{\alpha/2} \sigma_b, \text{ que también se puede escribirse como } \beta = b \pm z_{\alpha/2} \sigma_b$$

Con lo que nuestro coeficiente  $\beta$  auténtico estaría acotado, dentro de un intervalo alrededor del coeficiente  $b$  experimental, con un determinado nivel de confianza (que influye a través de  $z_{\alpha/2}$ ). Los resultados experimentales se suelen dar siempre de esta manera  $\beta = b \pm z_{\alpha/2} \sigma_b$ , añadiendo siempre una referencia al nivel de confianza utilizado en la construcción del intervalo de confianza.

Si el experimento se ha hecho con pocas medidas experimentales ( $n \leq 30$ ) entonces en vez de usar la distribución normal estándar ( $z_{\alpha/2}$ ) se usa la distribución t-Student ( $t_{\alpha/2}$ ). También se usa la distribución t-Student, si se desconoce la desviación típica de la población ( $\sigma_b$ ), y tenemos que usar la desviación típica muestral ( $s_b$ ). Lo habitual en trabajos experimentales es que se desconozca  $\sigma_b$  y sólo dispongamos de  $s_b$ .

En un análisis de regresión lineal (con dos coeficientes), el número de grados de libertad (de la distribución t-Student) es igual al número de muestras, menos dos. Para  $n=12$  muestras, el número de grados de libertad  $gl = (n-2) = 10$ , y nivel de confianza del 95% resultaría  $t_{\alpha/2} = 2,228$ , lo cual es lógico, ya que al tener pocas muestras experimentales, la dispersión final resulta ser mayor que en la distribución normal ( $z_{\alpha/2} = 1,96$ ).

Para pocas muestras experimentales hay que usar la distribución t-Student, y para muchas muestras se puede usar también la t-Student, ya que ésta se aproxima asintóticamente a la distribución normal estándar. Por tanto, para nosotros será lo más simple usar siempre la distribución t-Student.

$\beta - t_{\alpha/2} \sigma_b < b < \beta + t_{\alpha/2} \sigma_b$ , esto nos acota el valor de  $b$ , pero es interesante remarcar que es justamente  $b$  (el valor experimental obtenido) el que conocemos, y no tiene sentido acotarlo, y menos con  $\beta$  (valor auténtico y desconocido) y  $\sigma_b$  (desviación típica de la población) también desconocido.

Resumiendo, para construir un intervalo de confianza para un coeficiente  $\beta$  desconocido:

1. Tomamos  $n$  muestras experimentales.
2. Con el análisis de regresión calculamos el coeficiente experimental  $b$ , que es el mejor estimado de  $\beta$ .
3. Con las muestras experimentales calculamos el error estándar (desviación típica o incertidumbre estándar) del coeficiente  $b$ , que llamamos  $s_b$ , y que es el mejor estimado para  $\sigma_b$ .
4. Elegimos el nivel de confianza  $(1-\alpha)$ .
5. Conocido el valor de  $\alpha$  (nivel de significancia), y con los grados de libertad  $gl=(n-2)$  buscamos el valor de  $t_{\alpha/2}$  de la distribución t-Student.

El intervalo de confianza para el coeficiente  $\beta$  se puede escribir de cualquiera de las tres formas siguientes, aunque en física, se suele preferir la última.

$$\begin{aligned}\beta &\in (b - t_{\alpha/2} s_b, b + t_{\alpha/2} s_b) \\ b - t_{\alpha/2} s_b &< \beta < b + t_{\alpha/2} s_b \\ \beta &= b \pm t_{\alpha/2} s_b\end{aligned}$$

Remarquemos, que  $\beta$  no tiene por qué estar en el intervalo de confianza, no hay seguridad absoluta de que esté dentro de el intervalo. Lo que realmente nos dice el intervalo de confianza (95%) es que si repitiésemos el experimento 100 veces, en 95 ocasiones el intervalo así construido contendría el valor auténtico  $\beta$ , pero en ningún caso nos dice que no haya podido suceder la “desgracia” (5%) de que  $\beta$  se quede fuera del intervalo de confianza.

Existe una relación entre intervalos de confianza y prueba de hipótesis. En prueba de hipótesis se emplea un test estadístico  $t_b$  construido como:

$$t_b^2 = \left( \frac{b - \beta}{s_b} \right)^2$$

donde  $t_b$  sería el valor que debería compararse con  $t_{\alpha/2}$ . Si  $t_b$  fuera mayor que  $t_{\alpha/2}$  entonces el valor-t experimental estaría en la cola derecha y fuera del intervalo de confianza. Si en la anterior fórmula hacemos la raíz (con su signo  $\pm$ ) y forzamos el test estadístico  $t_b$  igual a  $t_{\alpha/2}$ , obtenemos que el valor auténtico y desconocido de  $\beta$  está comprendido en el intervalo de confianza mostrado anteriormente. En este apartado se han explicado el cálculo del intervalo de confianza para un coeficiente  $b$ , y su valor auténtico  $\beta$  desconocido; esta discusión se puede aplicar a los coeficientes  $a$  y  $b$  del ajuste de regresión lineal simple.

### 1.1.4. Transformaciones

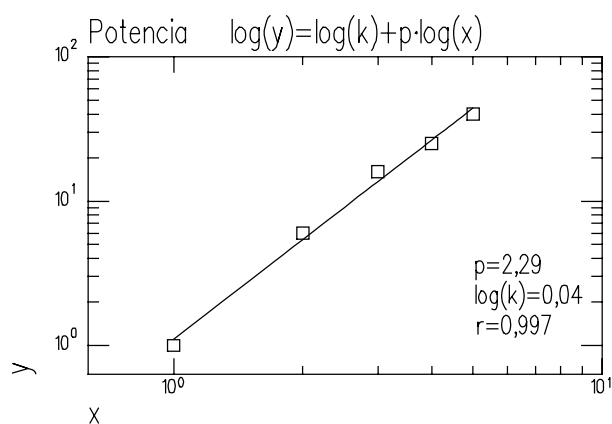
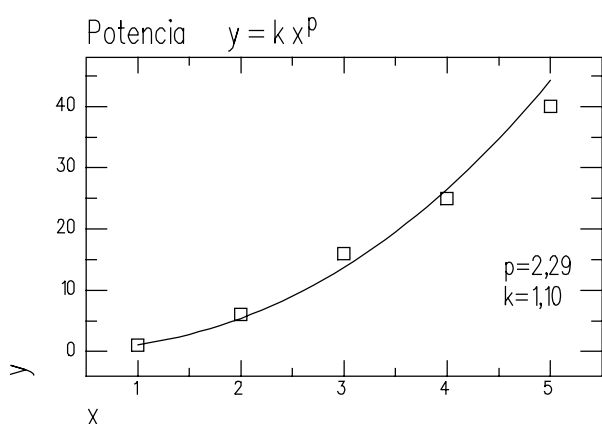
Una vez dominado el análisis de regresión lineal, nos planteamos si podemos hacer un análisis de regresión no lineal, o al menos, adaptar la regresión lineal a casos no lineales.

Utilizaremos transformaciones de variables que reduzcan el ajuste no lineal a un caso de ajuste lineal; para ello haremos el ajuste de regresión lineal sobre unas variables transformadas (X,Y) distintas de las originales (x,y).

Veamos qué transformaciones son útiles para ajustar funciones de tipo logarítmico, de tipo potencia, y exponencial:

	Transformación	Ajuste: $Y = A + B X$	
$y = a + b \log(x)$	$Y = y$ ; $X = \log(x)$	$Y = a + b X$	$a = A$ ; $b = B$
$y = k x^p$	$Y = \log(y)$ ; $X = \log(x)$	$Y = \log(k) + p X$	$k = 10^A$ ; $p = B$
$y = k b^{rx}$	$Y = \log(y)$ ; $X = x$	$Y = \log(k) + r \log(b) X$	$k = 10^A$ ; $r \log(b) = B$

Se han utilizado logaritmos decimales, pero se puede utilizar cualquier otra base (con los cambios pertinentes). Como ejemplo, se puede ver en las siguientes gráficas, que el ajuste lineal se emplea con la función  $y=k x^p$  usando la transformación  $X=\log(x)$   $Y=\log(y)$ , que da una recta en la gráfica logarítmica.



En el caso de función exponencial de la forma  $y = k b^{rx}$  puede suceder que sepamos la base (b), en ese caso podemos despejar la constante multiplicativa del exponente (r); si se conociese r, entonces se puede calcular la base. Si la base b fuera el número de euler (e), entonces sería más elegante utilizar logaritmos neperianos.

Siempre hay que tener en cuenta, que las fórmulas del ajuste de regresión lineal se desarrollaron suponiendo que los errores se distribuían según la distribución normal. Esto significa que, estrictamente, cuando hacemos una transformación, también habría que transformar la función de distribución de errores, y a partir de ahí, rehacer todas las fórmulas del ajuste de regresión. El trabajo necesario es normalmente tan grande, que se prefiere utilizar las transformaciones, simplemente añadiendo una frase en descargo, del tipo “suponemos que los errores de la función transformada se distribuyen también normalmente”.

Una advertencia sobre las funciones de tipo potencia ( $y = k x^p$ ), sólo se pueden ajustar funciones con  $x$  positivo, y se tiene la restricción de tener que pasar por el punto  $x=0$  e  $y=0$ . Esto nos excluiría el ajuste de parábolas arbitrarias. Veamos que también se puede hacer un ajuste, si usamos otras transformaciones.

Con la transformación de la primera fila de la siguiente tabla, podemos ajustar cualquier parábola, y vemos que el mismo procedimiento sirve para utilizar lo que ya sabemos sobre ajuste lineal a nuevas funciones no lineales.

	Transformación	Ajuste: $Y = A + B X$	
$y = a + b x^2$	$Y = y$ ; $X = x^2$	$Y = a + b X$	$a = A$ ; $b = B$
$y = a + b x^3$	$Y = y$ ; $X = x^3$	$Y = a + b X$	$a = A$ ; $b = B$
$y = \frac{1}{a + b x}$	$Y = \frac{1}{y}$ ; $X = x$	$Y = a + b X$	$a = A$ ; $b = B$
$y = a + b \sqrt[3]{x}$	$Y = y$ ; $X = x^{1/3}$	$Y = a + b X$	$a = A$ ; $b = B$
$y = \frac{1}{a + b/x^2}$	$Y = \frac{1}{y}$ ; $X = \frac{1}{x^2}$	$Y = a + b X$	$a = A$ ; $b = B$

Este método de las transformaciones nos sirve también en el caso de tener que ajustar una función lineal cuyos valores de las variables ( $x$  o  $y$ ) son tan grandes, que se produce desbordamiento al calcular las sumas de cuadrados.

Si los valores de la variable  $x$  son muy grandes, haremos la transformación  $X=x/100$  (primera fila), y si además los valores de la variable  $y$  fuesen muy grande podríamos hacer la transformación  $Y=y/100$  (segunda fila):

	Transformación	Ajuste: $Y = A + B X$	
$y = a + b x$ ; $x \gg$	$Y = y$ ; $X = x/100$	$Y = A + B \frac{x}{100}$	$a = A$ ; $b = B / 100$
$y = a + b x$ ; $x, y \gg$	$Y = y/100$ ; $X = x/100$	$\frac{y}{100} = A + B \frac{x}{100}$	$a = 100 A$ ; $b = B$

## 1.2. Funciones de distribución de probabilidad

### 1.2.1. Distribución normal

Aunque la expresión funcional de la distribución normal fue introducida por De Moivre, fueron Gauß y Laplace quienes desarrollaron y ampliaron los conceptos de curva y probabilidad normal. La distribución normal, es la forma más habitual en la que se distribuyen los errores aleatorios en los experimentos. El que la distribución de probabilidad se llame “normal” no significa que todos los errores se distribuyan según la campana de Gauss (o distribución normal).

La distribución normal es simétrica con respecto a la media ( $\mu$ ) y tiene una desviación típica nombrada como  $\sigma$ . En el caso particular de que la media fuese 0 ( $\mu=0$ ) y la desviación típica fuese 1 ( $\sigma=1$ ), entonces se denomina distribución normal estándar o tipificada.

En las siguientes formulas tenemos la expresiones para la distribución normal, y su variante estándar.

$$f(x) = \frac{1}{\sqrt{2\pi} \sigma} e^{-\frac{1}{2} \left[ \frac{x-\mu}{\sigma} \right]^2} \qquad f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2} z^2}$$

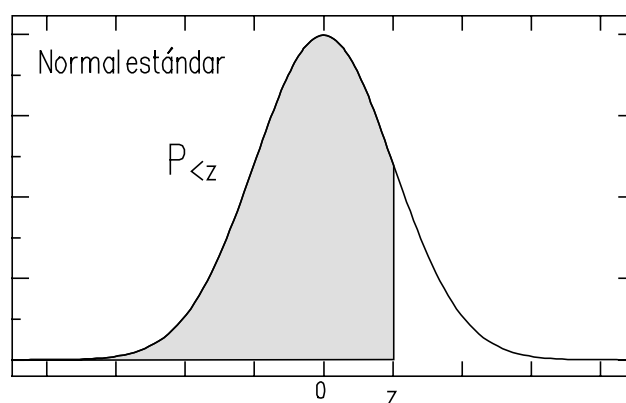
En los libros de estadística, sólo se tabula la distribución normal estándar (o tipificada), ya que basta un cambio de variable para pasar de una a otra:

$$z^2 = \left[ \frac{x-\mu}{\sigma} \right]^2 \Rightarrow \pm z = \left[ \frac{x-\mu}{\sigma} \right] \Rightarrow x = \mu \pm \sigma z \Rightarrow \mu = x \pm \sigma z$$

La probabilidad acumulada (de  $-\infty$  a  $x$ ) de la función de distribución normal (o de la distribución normal estándar) es la integral desde  $-\infty$  a  $x$  de la función densidad de probabilidad respectiva.

$$P_{<z}(x) = \frac{1}{\sqrt{2\pi} \sigma} \int_{-\infty}^x e^{-\frac{1}{2} \left[ \frac{t-\mu}{\sigma} \right]^2} dt \qquad P_{<z}(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{1}{2} t^2} dt$$

En la gráfica se representa la función densidad de probabilidad (la curva) de la distribución normal, y la probabilidad acumulada en la cola izquierda (de  $-\infty$  a  $x$ ) representada como un área oscura:



Ejemplo de tabla de la distribución normal estándar en los libros de Estadística. En el cuerpo de la tabla está la probabilidad acumulada ( $P_{<z}$ ) desde  $-\infty$  hasta  $z$  :

z	0,00	0,02	0,04	0,06	0,08
-3,0	0,0013	0,0013	0,0012	0,0011	0,0010
-2,0	0,0228	0,0217	0,0207	0,0197	0,0188
-1,0	0,1587	0,1539	0,1492	0,1446	0,1401
-0,0	0,5000	0,4920	0,4840	0,4761	0,4681
0,0	0,5000	0,5080	0,5160	0,5239	0,5319
1,0	0,8413	0,8461	0,8508	0,8554	0,8599
2,0	0,9772	0,9783	0,9793	0,9803	0,9812
3,0	0,9987	0,9987	0,9988	0,9989	0,9990

Así la probabilidad de que  $z$  sea menor que  $-1,02$  es de  $0,1539$  y por tanto la probabilidad de que  $z$  sea mayor que  $-1,02$  es de  $(1 - 0,1539) = 0,8461$ .

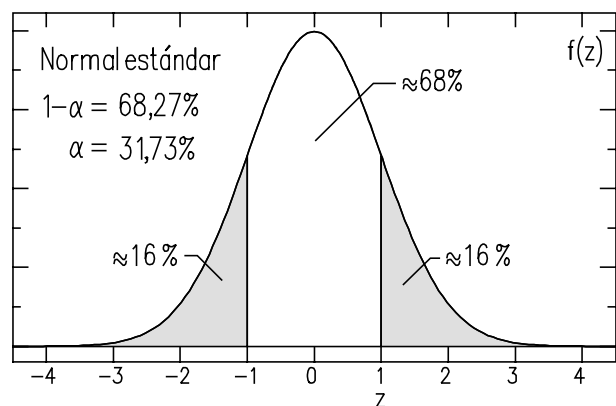
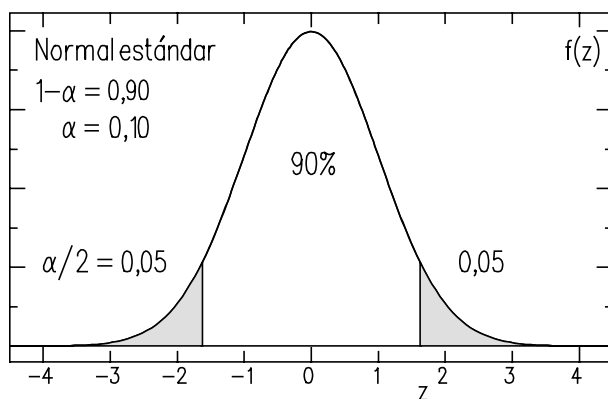
Otros casos:  $P_{<z}(-0,02) = 0,4920$      $P_{<z}(+0,02) = 0,5080$      $P_{<z}(2,04) = 0,9793$

Es habitual recurrir a niveles de confianza  $(1-\alpha)$  del 95%, o incluso del 99% en aplicaciones que requieren mayor seguridad. Raramente se baja de niveles de confianza del 90%. Son útiles los casos destacados en las dos tablas siguientes, en las que damos la probabilidad de la cola derecha, que es la unidad menos la probabilidad de la cola izquierda ( $P_{<z} + P_{>z} = 1 \Rightarrow P_{>z} = 1 - P_{<z}$ ).

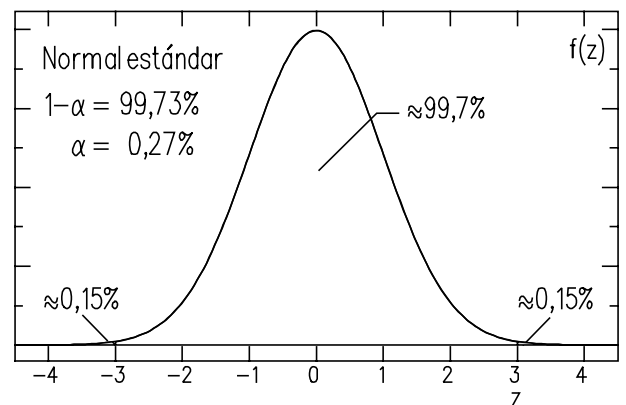
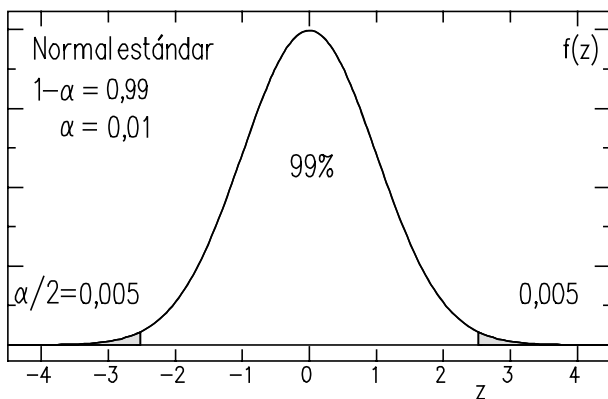
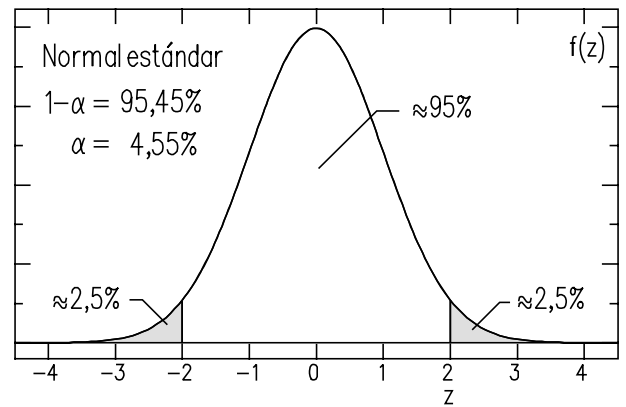
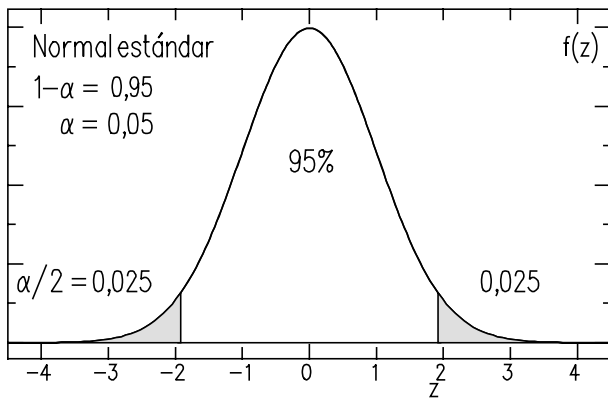
z	$\alpha/2$	$\alpha$	$1-\alpha$
0,674 489 75	0,25	0,5	50%
1,644 853 63	0,050	0,10	90%
1,959 963 99	0,025	0,05	95%
2,575 829 30	0,005	0,01	99%

z	$\alpha/2$	$\alpha$	$1-\alpha$
1	15,866%	31,731%	68,269%
2	2,275%	4,550%	95,450%
3	0,135%	0,270%	99,730%

Por último tenemos unas gráficas de la distribución normal estándar, con niveles de confianza  $(1-\alpha)$  distintos. En ellas se aprecia cómo al aumentar el nivel de confianza, se reducen las áreas de las dos colas. Se ve también, que aumentar el nivel de confianza, aumenta el ancho del intervalo (la incertidumbre).







### 1.2.2. Distribución t-Student

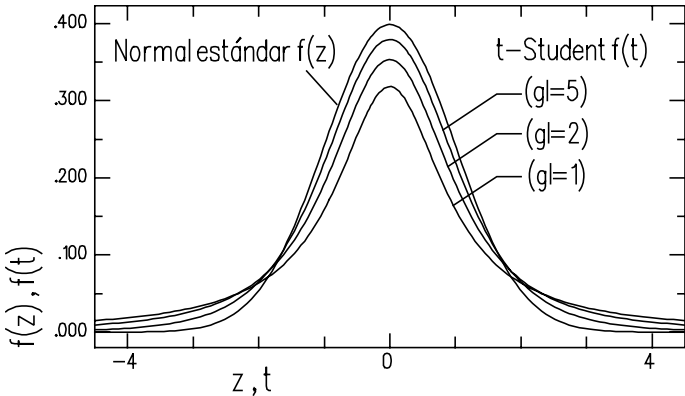
Un requisito imprescindible para usar la distribución normal, es conocer la desviación estándar, pero esto requiere conocer toda la población, es decir, que se necesitaría hacer infinitas medidas experimentales. Por tanto, en experimentación, se usa casi siempre la distribución t-Student. Cuando las medidas experimentales que tenemos son numerosas ( $n \geq 30$ ) puede usarse la distribución normal estándar, pero si hay pocas muestras ( $n \leq 30$ ) debe utilizarse la distribución t-Student, ya que es de esperar una distribución más dispersa que la distribución normal de la población de partida.

La razón, es que si se toma una muestra de datos grande, el teorema del límite central (de Estadística) nos dice que podemos utilizar la distribución normal para la media de los valores, aunque la distribución original de la población no fuese normal, pero cuando se hacen pocas medidas experimentales, la distribución no tiene por qué ser normal.

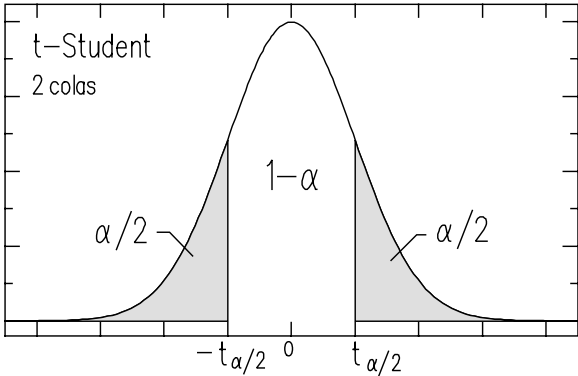
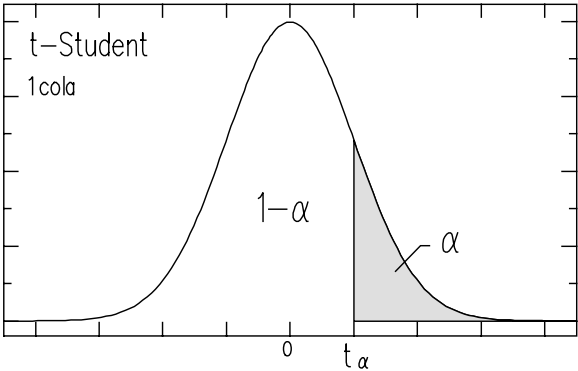
Conforme el tamaño de la muestra crece, la distribución t-Student tiende a coincidir con la distribución normal estándar, de tal forma que con muestras grandes ( $n > 30$ ) suele ser indiferente usar la distribución normal estándar o la distribución t.

En la siguiente gráfica se comparan las dos distribuciones, y se comprueba que con sólo 5 grados de libertad, ya son muy parecidas (gl=5 corresponde a 7 muestras experimentales en un análisis de regresión lineal con dos coeficientes). Con 10 grados de libertad (gl) en la escala de la gráfica siguiente, no se distinguiría la t-Student de la distribución normal estándar.

La distribución t-Student tiene una desviación típica  $\sigma$  mayor que uno, y además depende de los grados de libertad (gl).



En la distribución normal, es más habitual encontrar las tablas con la probabilidad acumulada desde  $-\infty$  hasta  $z$ , pero en la distribución t-Student, lo más habitual es dar la probabilidad de la cola derecha (ver gráfica de la izquierda) es decir con una sola cola. Algo menos habitual es encontrar las tablas con la t-Student de dos colas (ver gráfica de la derecha).



Como se puede apreciar, lo único a tener en cuenta para calcular la abscisa, es ver si se utiliza  $\alpha$  (en una cola) o  $\alpha/2$  (en dos colas). En la siguiente tabla se muestra la relación entre el nivel de confianza  $(1-\alpha)$  y el nivel de significancia ( $\alpha$ ) según sea el problema de una o dos colas.

UNA COLA $t_\alpha$	$1 - \alpha$	99,9%	99,5%	99%	97,5%	95%	90%	75%
	$\alpha$	0,001	0,005	0,01	0,025	0,05	0,1	0,25
DOS COLAS $t_{\alpha/2}$	$\alpha / 2$	0,001	0,005	0,01	0,025	0,05	0,1	0,25
	$\alpha$	0,002	0,01	0,02	0,05	0,1	0,2	0,5
	$1 - \alpha$	99,8%	99%	98%	95%	90%	80%	50%

Se emplea la distribución de una sola cola, cuando interesa saber la probabilidad de que podamos superar un límite por encima de la media. Se emplea la distribución de dos colas, cuando interesa saber la probabilidad de que nos alejemos de la media, por defecto o por exceso.

### Ejemplo

Tenemos latas de refresco con un contenido nominal de 33 cl. A la salida de la cadena de producción se toman 31 latas, se calcula su media, que resulta ser 32 cl y la desviación típica de la muestra, que resulta ser de 3 cl. El número de grados de libertad es 30, y la distribución t-Student ( $gl=30$ ) para  $\alpha=0,05$  es de 1,697. Por tanto, podemos decir que cuando cojamos una lata al azar, a la salida de la cadena de fabricación, tenemos un 5% de posibilidades de que la lata tenga más de 37,091 cl (= 32 cl + 1,697 x 3 cl). La probabilidad de que la lata tenga menos de 26,909 cl (= 32 cl - 1,697 x 3 cl) es también de un 5%, y finalmente tenemos un “nivel de confianza” del 90%, de que el contenido de la lata esté entre 37,091 cl y 26,909 cl.

Si a la salida de la cadena de fabricación tomamos 31 latas y hacemos la media de su contenido, la desviación estándar (de la media de 31 latas) es igual a la desviación estándar (3 cl) partido la raíz cuadrada del número de muestras ( $3\text{cl}/\sqrt{31} = 0,5388$  cl). Por tanto, en el anterior ejemplo se podría calcular la dispersión, utilizando esta desviación típica, en vez de los 3 cl usados anteriormente. Y diríamos, que la probabilidad de que el contenido medio de 31 latas esté entre  $(32 - 1,697 \times 0,5388)$  cl y  $(32 + 1,697 \times 0,5388)$  cl, es de un 90%.

El nivel de confianza es  $CL=(1-\alpha)=90\%$ , y el nivel de significancia ( $\alpha$ ) sería de un 10%, como es un experimento de dos colas (por arriba y por abajo), resulta que cada cola tiene un área de ( $\alpha/2=5\%$ ), por eso en la tabla siguiente hemos buscado  $\alpha = 0,05$ , puesto que es una tabla de una sola cola, que da la abscisa correspondiente al área de la cola derecha, y en nuestro problema, la cola de la derecha tenía un área de 0,05 ( $\alpha/2=5\%$ ).

Distribución t-Student de una cola, en función del número de grados de libertad (columna  $gl$ ), y del area de la cola derecha ( $\alpha$ ):

$gl \setminus \alpha$	0,001	0,005	0,01	0,025	0,05	0,1	0,25
1	318,309	63,657	31,821	12,706	6,314	3,078	1,000
2	22,327	9,925	6,965	4,303	2,920	1,886	0,816
3	10,215	5,841	4,541	3,182	2,353	1,638	0,765
5	5,893	4,032	3,365	2,571	2,015	1,476	0,727
10	4,144	3,169	2,764	2,228	1,812	1,372	0,700
20	3,552	2,845	2,528	2,086	1,725	1,325	0,687
30	3,385	2,750	2,457	2,042	1,697	1,310	0,683
$\infty$	3,090	2,576	2,326	1,960	1,645	1,282	0,675

En los libros de estadística o en las hojas de cálculo como Quattro-Pro o MS-Excel pueden obtenerse los valores de la distribución t-Student. Aquí, se ha dado una tabla con unas pocas entradas. En esta tabla, infinito corresponde a infinitas muestras; en ese caso la t-Student coincide con la distribución normal.

### 1.3. Funciones estadísticas en las hojas de cálculo

Se listan algunas funciones estadísticas utilizadas en Quattro-Pro, versión inglesa (comienzan por @) y en MS-Excel (versión española).

Referencia indirecta a la celda. El contenido de la celda es usado como rango

@@ (celda)

INDIRECTO (celda)

Número de datos n:

@COUNT (rango)

n

Sumatoria de datos:

@SUM (rango)

SUMA (rango)

$$\sum x$$

Suma de los cuadrados de los datos:

@SUMSQ (rango)

SUMAPRODUCTO (rango, rango)

$$\sum x^2$$

Suma del producto de datos x por datos y :

@SUMXY (rango\_X,rango\_Y)

SUMAPRODUCTO (rango\_X, rango\_Y)

$$\sum xy$$

Media de los valores:

@AVG (rango)

PROMEDIO (rango)

$$\sum x / n$$

Suma de cuadrados de las desviaciones:

@DEVSQ (rango)

DESVIA (rango)

$$\sum (x - \bar{x})^2$$

Desviación estándar de la población :

@STD (rango)

DESVESTP (rango)

$$\sqrt{\sum (x - \bar{x})^2 / n}$$

Desviación estándar de la muestra:

@STDS (rango)

DESVEST (rango)

$$\sqrt{\sum (x - \bar{x})^2 / (n - 1)}$$

Varianza de la población:

@VAR (rango)

VARP (rango)

$$\sum (x - \bar{x})^2 / n$$

Varianza de la muestra:

@VARs (rango)

VAR (rango)

$$\sum (x - \bar{x})^2 / (n - 1)$$

Covarianza de dos variables de una magnitud:

@COVAR (rango\_X,rango\_Y)

COVAR (rango\_X,rango\_Y)

$$\sum (x - \bar{x})(y - \bar{y}) / n$$

En las siguientes funciones, la fórmula se da en el apartado teórico correspondiente (salvo que aparezca explícitamente)

## Funciones de regresión lineal:

Coefficiente de correlación lineal (de Pearson) r:

@CORREL (rango\_Y, rango\_X)

@PEARSON (rango\_Y, rango\_X)

COEF.DE.CORREL (rango\_X,rango\_Y)

PEARSON (rango\_X,rango\_Y)

Coefficiente de determinación  $R^2 = r^2$

@RSQ (rango\_Y, rango\_X)

COEFICIENTE.R2 (rango\_X,rango\_Y)

De la recta de regresión  $y=a+bx$ :

Pendiente b:

@SLOPE (rango\_Y, rango\_X)

PENDIENTE (rango\_Y, rango\_X)

Corte de la recta con el eje de ordenadas a:

@INTERCEPT (rango\_Y, rango\_X)

INTERSECCION.EJE (rango\_Y;rango\_X)

Predice un valor y ( $y=a+bx$ ) para el valor de x, según el ajuste lineal ( $y=a+bx$ ) realizado sobre los rangos de variables dependientes (Y) e independientes (X).

@FORECAST (x,rango\_Y,rango\_X)

PRONOSTICO (x,rango\_Y,rango\_X)

## Funciones asociadas a funciones de distribución:

Normaliza los valores de una distribución con media  $\mu$  y desviación típica  $\sigma$

@STANDARDIZE (x, $\mu$ , $\sigma$ )

NORMALIZACION (x, $\mu$ , $\sigma$ )

$$z = \frac{x - \mu}{\sigma}$$

Calcula la probabilidad acumulada (de  $-\infty$  a  $x$ ) de la función de distribución normal estándar (con media 0 y desviación típica 1)

@NORMSDIST (x)

DISTR.NORM.ESTAND (x)

$$\int_{-\infty}^x \frac{e^{-\frac{1}{2}t^2}}{\sqrt{2\pi}} dt$$

La función inversa de @NORMSDIST o DISTR.NORM.ESTAND. Dada la probabilidad acumulada (prob) da la abscisa x correspondiente.

@NORMSINV (prob)

DISTR.NORM.ESTAND.INV (prob)

Calcula la función de distribución normal (con media  $\mu$  y desviación típica  $\sigma$ ):

@NORMDIST (x, $\mu$ , $\sigma$ ,0)

DISTR.NORM (x, $\mu$ , $\sigma$ ,0)

$$\frac{e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}}{\sqrt{2\pi}\sigma}$$

Calcula la probabilidad acumulada (de  $-\infty$  a  $x$ ) de la función de distribución normal (con media  $\mu$  y desviación típica  $\sigma$ ).

@NORMDIST (x, $\mu$ , $\sigma$ ,1)

DISTR.NORM (x, $\mu$ , $\sigma$ ,1)

$$\int_{-\infty}^x \frac{e^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2}}{\sqrt{2\pi}\sigma} dt$$

La función inversa de @NORMDIST o DISTR.NORM: Dada la probabilidad acumulada (prob) calcula la abscisa  $x$  correspondiente.

@NORMINV (prob, $\mu$ , $\sigma$ )

DISTR.NORM.INV (prob, $\mu$ , $\sigma$ )

Probabilidad de la distribución t-Student para el número de colas indicado, los grados de libertad indicados, y la abscisa  $x$ .

Para una cola, daría el área de probabilidad para t-Student desde  $|x|$  hasta  $\infty$ .

Para dos colas, daría el área de probabilidad para t-Student desde  $|x|$  hasta  $\infty$  más el área desde  $-|x|$  hasta  $\infty$  (que es el doble de área de probabilidad que para una t-Student de una cola)

@TDIST (x,gl,colas)

DISTR.T (x,gl,colas)

La función inversa de prob=@TDIST(x,gl,2). Dada la probabilidad de dos colas ( $\alpha$ =prob) se calcula la abscisa  $x$  correspondiente. Sería lo mismo que decir que calcula la abscisa que corresponde sólo a la cola derecha (con prob/2= $\alpha$ /2).

Para utilizarlo con una sola cola de probabilidad prob (=  $\alpha$ ) se utiliza @TINV(prob\*2,gl)

@TINV (prob,gl)

DISTR.T.INV (prob,gl)

@CONFIDENCE ( $\alpha$ , $\sigma$ ,n) Intervalo de confianza para la media de una población, es decir calcula el error de la siguiente forma:  $E_{x\_media} = |z_{\alpha/2}| \cdot \sigma_{x\_media} = |z_{\alpha/2}| \cdot \sigma_x / \sqrt{n}$

(entendiendo  $z_{\alpha/2}$  la abscisa que hace que la probabilidad acumulada desde  $-\infty$  a  $z$  sea  $\alpha/2$ )

@STEC (rango\_Y, rango\_X) Error estándar del coeficiente  $b$  de la recta de regresión, es decir  $s_b$  mostrado en fórmulas anteriores.

Error estándar de la regresión lineal  $s_e$

@STEYX (rango\_Y, rango\_X)

ERROR.TIPICO.XY (rango\_Y,rango\_X)

## 2. Incertidumbre en las medidas experimentales

El trabajo de laboratorio siempre está sometido a errores, que aunque no sean perfectamente cuantificables, influyen en que el resultado de la medida sea solamente, una aproximación o estimación de la magnitud física que estamos midiendo.

Los resultados medidos siempre tendrán una incertidumbre asociada, y el valor de esta incertidumbre debe darse con el resultado de la medida; si no se da la incertidumbre, el resultado no está completo.

### 2.1. Distinción entre error e incertidumbre

Error en una medida, es el resultado de la medida, menos el auténtico valor de la magnitud física que estamos midiendo. Como los valores auténticos nunca se pueden conocer exactamente, no tiene mucho sentido calcular el error, y más aún si tuviéramos el valor auténtico. Es por esa razón que los físicos utilizan como sinónimos las palabras “error” e “incertidumbre”, ya que tienen en cuenta que el error es imposible conocerlo.

La incertidumbre de una medida es el rango de valores en los que se cree que estará el error (valor medido menos valor real) con una determinada probabilidad (el llamado “nivel de confianza”).

Aunque un error lo tuviésemos identificado, y corrijiésemos la medida, siempre quedaría la propia incertidumbre en la medida del error que utiliza para la corrección.

Puede darse el caso de tener una medida sin error (ha salido exactamente el valor real) pero nosotros no podamos saberlo, y por tanto, tengamos incertidumbre en la medida, aunque ésta no tenga error.

El error total se compone de la suma del **error aleatorio** y del **error sistemático**.

#### **Error aleatorio**

Si se hiciese la media de infinitas medidas experimentales, bajo las mismas condiciones, el error aleatorio de una medida en concreto, sería el valor de esa medida menos aquella media de las infinitas medidas experimentales.

#### **Error sistemático**

Si se hiciese la media de infinitas medidas experimentales, bajo las mismas condiciones, el error sistemático sería la resta de esa media menos el valor real (desconocido) de la magnitud que se mide.

Como evidentemente no se pueden hacer infinitas medidas, ni conseguir el valor auténtico, sólo podremos conseguir unas estimaciones de los errores aleatorios o sistemáticos.

En definitiva, el error aleatorio de una medida se puede hacer tender a cero aumentando la cantidad de medidas realizadas. Eso no sucede para un error sistemático, que siempre tiene media distinta de cero aunque se aumente el número de medidas, por ello siempre que se pueda, hay que corregir los errores sistemáticos, o mejor aún eliminarlos rehaciendo de nuevo el experimento sin errores sistemáticos desde el principio.

### Ejemplo

Tenemos que medir una longitud, y utilizamos una regla a la que le faltan los 2 cm iniciales. El resultado de la medida queda entre 112 y 113 mm. Como queda más cerca de 112 mm, ésta será la medida. El observador apunta  $x = 112 \pm 1$  mm. La incertidumbre debida al error aleatorio de alineación de los extremos inicial y final de la regla es 1 mm. Se descubre posteriormente que a la regla le falta el trozo inicial, y como ya no se dispone del objeto a medir (pero sí de la regla), se mide el trozo que le falta; la corrección resulta ser:  $c = 20 \pm 1$  mm. Hemos descubierto un error sistemático que habrá afectado a todas las medidas realizadas con esa regla. No tiene mucho sentido aumentar el error total y decir que la medida es  $x = 112 \pm 20$  mm. Lo conveniente cuando se descubren errores sistemáticos es corregir la medida:  $x = (112 - 20) \pm (1^2 + 1^2)^{1/2}$  así que el resultado final queda  $x = 92 \pm 1,4$  mm\*. Si se dispone del objeto a medir, lo mejor es volver a repetir la medida (con una regla correcta), si se hiciera, la medida tendría menor error:  $x = 92 \pm 1$  mm.

(\* estrictamente deberíamos haber reducido el número de cifras significativas de 1,4 a 1)

La naturaleza de un error (sistemático o aleatorio) viene condicionada por el uso que se le da. Un error aleatorio puede convertirse en un error sistemático y viceversa. Como ejemplo, una componente de incertidumbre debida a un error aleatorio que sucede en la fabricación de un aparato de medida, se convierte en un error sistemático (fijo) para el usuario de ese aparato de medida. Hay que resaltar que un error es aleatorio o sistemático sólo en un experimento concreto, en otro distinto, su clasificación puede ser otra.

La incertidumbre de una medida puede ser clasificada como:

**de Tipo A:** Cuando las incertidumbres han sido calculadas utilizando métodos estadísticos, sobre varias observaciones; por ejemplo, cuando se utiliza el método de los mínimos cuadrados.

**de Tipo B:** Cuando las incertidumbres no han sido calculadas con métodos estadísticos, sino con razonamientos, deducciones o cualquier otro tipo de información (p.ej. las especificaciones del aparato de medida).

Aunque es usual que los errores aleatorios se evalúen por métodos estadísticos (tipo A), y los sistemáticos no (tipo B), no siempre es así.

## 2.2. Ley de propagación de incertidumbres

En el laboratorio, suele ocurrir que una magnitud física a medir ( $Y$ ) no se pueda medir directamente, sino que tenga que ser calculada con las medidas de otras magnitudes ( $X_i$ ):

$$Y = f(X_1, X_2, \dots, X_N)$$

$X_i$  pueden ser magnitudes físicas, pero también pueden ser correcciones para las medidas de esas magnitudes; también podrían ser cualquier otra fuente de variabilidad, como constantes físicas o numéricas, no utilizadas con su valor exacto. Las medidas tomadas por distintos operadores o con distintos instrumentos se tratan también como magnitudes  $X_i$  distintas.



En ese caso el valor estimado de la medida de Y (y) se obtendrá a través de la relación funcional:

$$y = f(x_1, x_2, \dots, x_N)$$

Todas las magnitudes  $X_i$  las tendremos expresadas como valor estimado ( $x_i$ ) e *incertidumbre estándar* de  $x_i$ , que llamaremos  $u(x_i)$ :

$$X_i = x_i \pm u(x_i)$$

y necesitamos calcular la *incertidumbre estándar combinada* para la medida y de la magnitud Y, que llamaremos  $u_c(y)$ .

La incertidumbre estándar combinada  $u_c(y)$  se calcula con la *ley de propagación de incertidumbres*, también llamada “fórmula RSS” (raíz de suma de cuadrados).  $u_c^2(y)$  representaría el estimado de la varianza de la medida y. Normalmente se calcula  $u_c^2(y)$  con la siguiente fórmula, y luego se hallaría la raíz cuadrada positiva.

$$u_c^2(y) = \sum_{i=1}^N \left( \frac{\partial f}{\partial x_i} \right)^2 u^2(x_i) + 2 \sum_{i=1}^{N-1} \sum_{j=i+1}^N \frac{\partial f}{\partial x_i} \frac{\partial f}{\partial x_j} u(x_i, x_j)$$

despejando, quedaría:

$$u_c(y) = \sqrt{\sum_{i=1}^N \left( \frac{\partial f}{\partial x_i} \right)^2 u^2(x_i) + 2 \sum_{i=1}^{N-1} \sum_{j=i+1}^N \frac{\partial f}{\partial x_i} \frac{\partial f}{\partial x_j} u(x_i, x_j)}$$

donde se observa que la incertidumbre combinada depende de las incertidumbres  $u(x_i)$  y también de las covarianzas estimadas  $u(x_i, x_j)$  entre las medidas  $x_i$  y  $x_j$ .

Si todas las variables  $x_i$  son estadísticamente independientes, es decir no están correlacionadas, sus covarianzas serán cero, y la fórmula RSS antes presentada, se simplifica:

$$u_c^2(y) = \sum_{i=1}^N \left( \frac{\partial f}{\partial x_i} \right)^2 u^2(x_i) = \sum_{i=1}^N c_i^2 u^2(x_i) = \sum_{i=1}^N u_i^2(y)$$

donde hemos utilizado los *coeficientes de sensibilidad*  $c_i$ , que se definen como la derivada parcial de la función  $f$  respecto a cada uno de sus argumentos  $x_i$ , evaluadas en los valores medidos  $x_1, x_2, \dots, x_N$ .

$$c_i = \left. \frac{\partial f}{\partial x_i} \right|_{x_1, x_2, \dots, x_N}$$

Los coeficientes de sensibilidad pueden calcularse derivando analíticamente la función, o bien experimentalmente, calculados como  $c_i = \Delta y / \Delta x_i$ . Estos coeficientes indican también la influencia de cada magnitud medida  $x_i$  en el valor final de la magnitud y calculada.

También utilizamos la *incertidumbre estándar* de la medida  $y$ ,  $u_i(y)$ , que es la influencia de la incertidumbre estándar de cada  $x_i$  ( $u(x_i)$ ) multiplicada por el coeficiente de sensibilidad.

$$u_i(y) = c_i u(x_i) = \frac{\partial f}{\partial x_i} u(x_i)$$

Algunos autores definen la incertidumbre estándar, con un coeficiente de sensibilidad en módulo.

Aunque teniendo el valor medido ( $y$ ) y su incertidumbre estándar combinada  $u_c(y)$  tenemos la medida totalmente especificada, suelen darse las medidas indicando un *nivel de confianza* CL (también llamado probabilidad de cubrimiento  $p$ , o también notado como  $(1-\alpha)$ , siendo  $\alpha$  el nivel de significancia). El nivel de confianza nos indicaría la probabilidad de que el intervalo de la medida contenga (o cubra) al auténtico valor de la magnitud.

Para ello, tras especificar el nivel de confianza se calcula la *incertidumbre expandida*  $U$ , definida como

$$U = k u_c(y)$$

donde  $k$  es el *factor de cubrimiento*, que depende del nivel de confianza elegido, de la distribución de probabilidades de la magnitud  $y$ , (distribución normal, distribución t-Student, etc.) y de los grados de libertad (si procediera). Para calcular  $k$ , basta consultar las tablas estadísticas.

El número de grados de libertad efectivos ( $v_{ef}$ ) de la magnitud  $y$ , se calcularía despejando  $v_{ef}$  de la siguiente fórmula:

$$\frac{u_c^2(y)}{\sqrt{v_{ef}(y)}} = \sum_{i=1}^N \frac{\left(\frac{\partial f}{\partial x_i}\right)^2 u^2(x_i)}{\sqrt{v_i}} + 2 \sum_{i=1}^{N-1} \sum_{j=i+1}^N \frac{\frac{\partial f}{\partial x_i} \frac{\partial f}{\partial x_j} u(x_i, x_j)}{\sqrt{\sqrt{v_i} \sqrt{v_j}}}$$

Afortunadamente, en los casos prácticos esta fórmula no hace falta utilizarla, puesto que suele bastar con dos propiedades muy sencillas:

- Si la variable  $y$  sólo depende de una variable, el número de grados de libertad de  $y$  es igual al número de grados de libertad de la variable  $x$ .
- Si la variable  $y$  depende de varias variables, que tienen todas el mismo número de grados de libertad, entonces el número de grados de libertad de la variable  $y$  es igual al número de grados de libertad que comparten todas las variables.

En el caso de que todas las variables  $x_i$  sean estadísticamente independientes (no están correlacionadas) se simplifica:

$$v_{ef}(y) = u_c^4(y) \left( \sum_{i=1}^N \frac{\left(\frac{\partial f}{\partial x_i}\right)^2 u^2(x_i)}{\sqrt{v_i}} \right)^{-2}$$

En general,  $v_{ef}$  saldrá un número no entero. Normalmente se trunca (al entero inferior), pero también podría interpolarse; aunque esto último sólo se suele hacer cuando  $v_{ef}$  es muy bajo, y se trabaja con niveles de confianza altos.

Cuando se trate de calcular los grados de libertad en incertidumbres de tipo B, (o en el caso de incertidumbres de tipo A si desconocemos los grados de libertad) y si no tenemos otra información se prefiere sobreestimar la incertidumbre; para ellos supondríamos que  $\nu \rightarrow \infty$ , antes de deducir la desviación típica o equivalente.

Así el *intervalo de confianza* (estrictamente sólo se puede hablar de intervalo de confianza cuando sólo se aplican métodos estadísticos, es decir si sólo hubiesen incertidumbres de tipo A) quedaría:

$$Y = y \pm U = y \pm k u_c(y)$$

Niveles de confianza (CL), distribuciones de probabilidad (Dist.), grados de libertad ( $\nu$ ) y factor de cubrimiento ( $k$ ) más usuales se dan en la siguiente tabla. En la distribución normal, no tienen sentido los grados de libertad y en la distribución t-Student, se dan sólo  $\nu$  igual a 10 y a 15, pero sólo como ejemplo (no tienen por qué coincidir con los del experimento).

CL	Dist	$\nu$	k
50%	Normal	-	0,6745
68,27%	Normal	-	1
95,45%	Normal	-	2
99,73%	Normal	-	3
90%	Normal	-	1,6449
95%	Normal	-	1,9600
99%	Normal	-	2,5758
95%	t-Student	10	2,2281
99%	t-Student	15	2,9467

En la presentación de los resultados finales conviene dar la mayor información posible sobre las incertidumbres. Se debe dar el resultado  $Y = y \pm U$ , y aparte se darían:

- el nivel de confianza utilizado (CL)
- la distribución de probabilidad utilizada (Dist.)
- los grados de libertad (si procede)  $\nu$
- el factor de cubrimiento ( $k$ )

En el caso de que no se dieran todos estos datos, la mayoría suele suponer que la distribución de probabilidad es la distribución normal y que el nivel de confianza es CL=95% (factor de cubrimiento  $k=1,960$ ) o que el factor de cubrimiento  $k=2$  (CL≈95%). Pero no siempre es así; hay otros, que en ausencia de información suponen CL=68,27%, otros suponen CL=50%, y en campos especializados en la seguridad de las personas se trabaja habitualmente con niveles de confianza por encima del 99%.

También se recomienda hacer una lista de todas las componentes de incertidumbre estándar, que contribuyan a la incertidumbre estándar combinada.

**Ejemplo 1** Calcular el valor de la tensión térmica  $V_T = k T / q$ .

Datos de las constantes físicas:

magnitud de la carga del electrón  $q = (1,602\ 176\ 53 \pm 0,000\ 000\ 14) 10^{-19}$  C

Constante de Boltzmann  $k = (1,380\ 6505 \pm 0,000\ 0024) 10^{-23}$  J/K

En la tabla de datos físicos consultada, se nos indica que la incertidumbre reflejada es la misma desviación típica. Como no dicen nada de los grados de libertad, suponemos que son infinito para las dos constantes físicas.

La temperatura (K) se midió en el laboratorio 10 veces, y tras el cálculo estadístico resultó:

299,93 K Media de T

0,319896 K Desviación estándar

0,101160 K Desviación estándar de la media

9 Grados de libertad

Calculamos los coeficientes de sensibilidad ( $c_i$ ), derivando la función  $V_T(k,T,q)$ :

$$c_k = \frac{\partial V_T}{\partial k} = \frac{T}{q} \quad ; \quad c_T = \frac{\partial V_T}{\partial T} = \frac{k}{q} \quad ; \quad c_q = \frac{\partial V_T}{\partial q} = -\frac{kT}{q^2}$$

En la siguiente tabla agrupamos los valores numéricos de los valores de las variables ( $x_i$ ), las incertidumbres de las variables ( $u(x_i)$ ), el tipo de incertidumbre, los grados de libertad y los coeficientes de sensibilidad ( $c_i$ ).

$X_i$	$x_i$	$u(x_i)$	tipo	$\nu$	$c_i$	$c_i u(x_i)$
k (J/K)	$1,380\ 6505\ 10^{-23}$	$2,4\ 10^{-29}$	B	$\infty$	$1,872\ 016\ 10^{21}$	$4,492\ 838\ 10^{-8}$
T (K)	299,93	0,101 160	A	9	$8,617\ 343\ 10^{-5}$	$8,717\ 299\ 10^{-6}$
q (C)	$1,602\ 176\ 53\ 10^{-19}$	$1,4\ 10^{-26}$	B	$\infty$	$-1,613\ 180\ 10^{17}$	$-2.258\ 453\ 10^{-9}$

$V_T$ (V)=	0,025 845 997	$\nu_{\text{eff}}(V_T) =$	9,0005	$u_c(V_T) =$	$8,717\ 415\ 10^{-6}$
------------	---------------	---------------------------	--------	--------------	-----------------------

Como se puede comprobar, la incertidumbre combinada es casi igual a la incertidumbre debida a T, ya que la incertidumbre en las constantes físicas es bajísima.

Si deseamos un nivel de confianza ( $CL=1-\alpha$ ) de 95%, necesitaremos calcular la incertidumbre expandida ( $U=k \cdot u(V_T)$ ), y por tanto antes tendremos que calcular k (factor de cubrimiento) a partir de la distribución t-Student y el número de grados de libertad de  $V_T$ .

CL=	95%	k =	2,262 157	$U(V_T) =$	$1,972\ 10^{-5}$
-----	-----	-----	-----------	------------	------------------

Para calcular k, hemos utilizado el número efectivo de grados de libertad  $\nu_{\text{eff}}$ , truncándolo a entero. Aunque algunos prefieren interpolar; no tiene mucho sentido alcanzar precisiones extremas, cuando se ha calculado con fórmulas que usan distintas aproximaciones, y al final terminan reduciéndose las cifras significativas de la incertidumbre total y haciéndose un redondeo final.

Por tanto, tras la reducción de cifras significativas, y su redondeo asociado, el resultado final quedaría:

$V_T =$	$0,025\ 846 \pm 0,000\ 020$	V
---------	-----------------------------	---

**Ejemplo 2** Tenemos 42 datos de tensión e intensidad de un diodo (V,I). Deseamos calcular la tensión umbral  $V_\gamma$ , definida como el corte con el eje de abscisas de la recta de regresión  $y=a+bx$  (con  $y=I$ , y  $x=V$ ).

Tras el análisis de regresión lineal obtenemos los coeficientes  $a$  y  $b$ , las desviaciones típicas  $s(a)$  y  $s(b)$ , y la covarianza entre  $a$  y  $b$   $cov(a,b)$ . Esos datos se incluyen ya en la tabla de las incertidumbres.

Calculamos los coeficientes de sensibilidad ( $c_i$ ), derivando la función  $V_\gamma(a,b) = -a/b$ :

$$c_a = \frac{\partial V_\gamma}{\partial a} = -\frac{1}{b} \quad ; \quad c_b = \frac{\partial V_\gamma}{\partial b} = \frac{a}{b^2}$$

En la siguiente tabla agrupamos los valores numéricos de los valores de las variables ( $x_i$ ), las incertidumbres de las variables ( $u(x_i)$ ), el tipo de incertidumbre, los grados de libertad y los coeficientes de sensibilidad ( $c_i$ ).

$X_i$	$x_i$	$u(x_i)$	tipo	$v$	$c_i$	$c_i u(x_i)$
$a$ ( $\mu A$ )	-173,5	7,562	A	40	-3,938 56	-29,7834
$b$ ( $\mu A/mV$ )	0,2539	0,0102	A	40	-2691,37	-27,452
	$cov(a,b) =$	-0,002772	A	40	10600,13	-29,3836

$V_\gamma$ (mV)=	683,3399	$v_{eff} (V_\gamma) =$	73,788	$u_c(V_\gamma) =$	39,773 05
------------------	----------	------------------------	--------	-------------------	-----------

Si deseamos un nivel de confianza ( $CL=1-\alpha$ ) de 95%, necesitaremos calcular la incertidumbre expandida ( $U=k \cdot u(V_\gamma)$ ), y por tanto antes tendremos que calcular  $k$  (factor de cubrimiento) a partir de la distribución t-Student y el número de grados de libertad de  $V_\gamma$ .

$CL=$	95%	$k =$	1,992 543	$U(V_\gamma) =$	79,2495
-------	-----	-------	-----------	-----------------	---------

Por tanto, tras la reducción de cifras significativas, y su redondeo asociado, el resultado final quedaría:

$V_\gamma =$	680 $\pm$ 80	mV
--------------	--------------	----

## 2.3. Apéndices

### 2.3.1. Deducción de la fórmula RSS y de la suma algebraica

Si la relación funcional entre la magnitud  $Y$ , y las magnitudes  $X_i$  es:

$$Y = f(X_1, X_2, \dots, X_N)$$

y tenemos las medidas de las magnitudes  $X_i$  y sus incertidumbres:

$$X_i = x_i \pm u(x_i) = x_i \pm \Delta x_i$$

El valor de la medida tomado para la magnitud  $Y$  será:

$$y = f(x_1, x_2, \dots, x_N)$$

Si desarrollamos en serie de Taylor, eliminando los términos de segundo orden y superiores:

$$\begin{aligned} y + \Delta y &= f(x_1 + \Delta x_1, x_2 + \Delta x_2, \dots, x_N + \Delta x_N) = \\ &= f(x_1, x_2, \dots, x_N) + \left[ \frac{\partial f}{\partial x_1} \Delta x_1 + \frac{\partial f}{\partial x_2} \Delta x_2 + \dots + \frac{\partial f}{\partial x_N} \Delta x_N \right] \end{aligned}$$

donde se han evaluado todas las derivadas parciales en los valores medidos  $x_1, x_2, \dots, x_N$ .

Despejamos  $\Delta y$ :

$$\Delta y = \sum_{i=1}^N \frac{\partial f}{\partial x_i} \Delta x_i$$

Si en la anterior fórmula se toman valores absolutos en las derivadas parciales, tenemos la ley de propagación de errores según la suma algebraica:

$$\Delta y = \sum_{i=1}^N \left| \frac{\partial f}{\partial x_i} \right| \Delta x_i$$

donde el valor absoluto se toma para tener en cuenta el peor caso posible, es decir que todos los errores  $\Delta x_i$  se sumaran en un solo sentido. Más adelante se discutirá su aplicabilidad, e inconvenientes respecto a la fórmula RSS para la propagación de errores.

Seguimos con la deducción de la fórmula RSS. Calculamos ahora la desviación típica de los valores medidos ( $y$ ), y que llamaremos  $\sigma = \sigma(y) = u_c(y)$  en las siguientes fórmulas; también llamaremos  $\sigma_i = \sigma(x_i) = u(x_i)$  a las desviaciones típicas de las  $x_i$ . Si se quiere resaltar el hecho de que no son desviaciones típicas poblacionales, sino muestrales, se sustituiría  $\sigma$  por  $s$ . Aunque esta deducción se hace para variables aleatorias, el uso de la fórmula RSS se utiliza para todo tipo de errores: sistemáticos y aleatorios.

$$\begin{aligned}
\sigma^2 &= \frac{1}{n} \sum_{k=1}^n (y_k - \bar{y})^2 = \frac{1}{n} \sum_{k=1}^n \Delta y_k^2 = \frac{1}{n} \sum_{k=1}^n \left( \sum_{i=1}^N \frac{\partial f}{\partial x_i} \Delta x_{ik} \right)^2 = \\
&= \frac{1}{n} \sum_{k=1}^n \left[ \sum_{i=1}^N \left( \frac{\partial f}{\partial x_i} \right)^2 \Delta x_{ik}^2 + 2 \sum_{i=1}^{N-1} \sum_{j=i+1}^N \frac{\partial f}{\partial x_i} \frac{\partial f}{\partial x_j} \Delta x_{ik} \Delta x_{jk} \right] = \\
&= \sum_{i=1}^N \left( \frac{\partial f}{\partial x_i} \right)^2 \frac{1}{n} \sum_{k=1}^n \Delta x_{ik}^2 + 2 \sum_{i=1}^{N-1} \sum_{j=i+1}^N \frac{\partial f}{\partial x_i} \frac{\partial f}{\partial x_j} \frac{1}{n} \sum_{k=1}^n \Delta x_{ik} \Delta x_{jk} = \\
&= \sigma^2 = \sum_{i=1}^N \left( \frac{\partial f}{\partial x_i} \right)^2 \sigma_i^2 + 2 \sum_{i=1}^{N-1} \sum_{j=i+1}^N \frac{\partial f}{\partial x_i} \frac{\partial f}{\partial x_j} \sigma_{ij}
\end{aligned}$$

Con lo que queda demostrada la fórmula RSS de propagación de incertidumbres. El índice k recorre las n medidas de las magnitudes  $x_i$ . A la medida k de la magnitud  $x_i$  le llamamos  $x_{ik}$ . Se supone que todas las magnitudes  $x_i$  tienen exactamente el mismo número de grados de libertad n.

### 2.3.2. Aplicabilidad comparada de las fórmulas RSS y de la suma algebraica

Las incertidumbres se pueden combinar según la fórmula de la suma algebraica (S.A.) (izquierda) o bien con la fórmula de la raíz cuadrada de la suma de los cuadrados de las incertidumbres (RSS) (derecha).

$$\Delta y = \sum_{i=1}^N \left| \frac{\partial f}{\partial x_i} \right| \Delta x_i \qquad \sigma^2 = \sum_{i=1}^N \left( \frac{\partial f}{\partial x_i} \right)^2 \sigma_i^2 + 2 \sum_{i=1}^{N-1} \sum_{j=i+1}^N \frac{\partial f}{\partial x_i} \frac{\partial f}{\partial x_j} \sigma_{ij}$$

Si utilizamos la notación estándar para incertidumbres, quedan:

$$u_c(y) = \sum_{i=1}^N \left| \frac{\partial f}{\partial x_i} \right| u(x_i) \qquad u_c^2(y) = \sum_{i=1}^N \left( \frac{\partial f}{\partial x_i} \right)^2 u^2(x_i) + 2 \sum_{i=1}^{N-1} \sum_{j=i+1}^N \frac{\partial f}{\partial x_i} \frac{\partial f}{\partial x_j} u(x_i, x_j)$$

No existe unanimidad entre los usuarios en cómo se deben combinar las incertidumbres, y eso es así porque entre otras cosas, no existe una forma absolutamente correcta de aplicación universal. Desde principios de los noventa, los organismos internacionales se han ido decantando por usar siempre la fórmula RSS. Inicialmente, en Física se trataban las incertidumbres (los errores) con la fórmula S.A., lo cual tenía ventajas e inconvenientes.

Ventajas de la fórmula S.A.:

- Es rápida y simple de aplicar.
- Sobreestimaba los errores, lo que no es malo, pero mucha gente solía entender que proporcionaban niveles de confianza del 100%. En realidad daba una falsa sensación de seguridad, puesto que no llegaba a niveles de confianza del 100%, ya que en la serie de Taylor se eliminaban términos de segundo orden y superiores.

### Inconvenientes de la fórmula S.A.:

- Al sobreestimar las incertidumbres, resultan valores muy altos, que corresponderían a una visión extremadamente pesimista del experimento. Si dicha fórmula se aplica varias veces y a bastantes componentes de incertidumbre, resulta una incertidumbre combinada que correspondería a una situación altamente improbable, en la que todas las incertidumbres estuvieran jugando correlacionadamente en el mismo sentido de aumentar la incertidumbre total.
- En un proceso normal de medida, suelen tomarse medidas repetidas para dar la media como el mejor valor estimado de la medida (se recurre a la estadística). En ese proceso, el propio experimentador suele eliminar los datos experimentales que presentan valores muy desviados de la media (se vuelve a recurrir a la intuición estadística del experimentador). Es decir, estamos utilizando la Estadística constantemente, pero no los utilizamos en la combinación de incertidumbres; es poco coherente.
- Como se ha dicho antes, no se consiguen niveles de confianza del 100%, sino menores. Como no se usan métodos estadísticos, no podemos dar un nivel de confianza para la medida experimental, y por supuesto tampoco podríamos crear intervalos de confianza.
- Las incertidumbres procedentes de errores aleatorios, se procesan estadísticamente, pero no se combinan de forma coherente con la fórmula S.A.
- No se puede usar cuando hay variables correlacionadas (a pesar de que algunos dicen lo contrario, como supuesta ventaja de la fórmula S.A sobre la fórmula RSS).

En las dos fórmulas siempre se tiene un error por truncamiento de la serie de Taylor (de los términos de segundo orden y superiores), ese error es importante cuando la incertidumbre es muy grande. Algunos autores dicen que en casos de incertidumbres grandes debe utilizarse sólo la fórmula RSS, cuando la fórmula RSS siempre da errores menores que los de la fórmula S.A.

Vistos los inconvenientes, pronto se hizo necesario tratar al menos, la combinación de incertidumbres de errores aleatorios con la fórmula RSS. Durante muchos años se combinaban las incertidumbres procedentes de errores aleatorios con la fórmula RSS, y se dejaba la fórmula S.A para las incertidumbres que provenían de errores sistemáticos. Así se solucionaron algunos problemas, pero no todos. De hecho había muchos resultados experimentales en los que se hacía el tratamiento totalmente separado de los errores aleatorios y de los errores sistemáticos. Se daban entonces dos incertidumbres según proviniesen de errores aleatorios o sistemáticos; otros en cambio combinaban esas dos incertidumbres con la fórmula:

$$u_c^2(y) = \sqrt{u_{\text{SISTEM}}^2 + u_{\text{ALEAT}}^2}$$

Aún así, no se solucionaban los problemas, ya que si  $u_{\text{SISTEM}}$  no tenía un nivel de confianza definido, pero sí  $u_{\text{ALEAT}}$ , entonces ¿cuál sería el nivel de confianza final? Naturalmente, no se pueden construir los intervalos de confianza. Por estas razones, a principios de los noventa, se decidió utilizar siempre la fórmula RSS, para combinar incertidumbres, y establecer unos métodos universales para calcular los niveles de confianza, incluso cuando las incertidumbres proceden de errores sistemáticos. En estos apuntes se siguen esas normas de los años noventa.



### 2.3.3. Sobre la covarianza

A veces, resulta que tenemos una magnitud (Y) que depende de magnitudes que están correlacionadas (A y B) simplemente porque comparten alguna misma magnitud ( $X_i$ ) en la dependencia funcional.

Por ejemplo:

$$Y = f(A, B) \quad \text{con:} \quad A = f_b(X_1, X_2, \dots, X_N) \quad \text{y} \quad B = f_a(X_1, X_2, \dots, X_N)$$

En esos casos lo mejor es expresar la dependencia funcional de la magnitud Y de tal forma que dependa directamente de las variables ( $X_i$ ); de esta forma, además se ahorra en cálculos, aproximaciones y en posibilidades de error. Si no hubiera más remedio, entonces se utiliza que la covarianza de las medidas a y b (de las magnitudes A y B) se puede calcular (si las  $X_i$  son independientes entre sí) con la siguiente fórmula:

$$u(a, b) = \sum_{i=1}^N \left( \frac{\partial f_a}{\partial x_i} \right) \left( \frac{\partial f_b}{\partial x_i} \right) u^2(x_i) = \sum_{i=1}^N c_{ai} c_{bi} u^2(x_i)$$

Podemos tener casos en los que el grado de correlación no se puede calcular de ninguna manera, y nos es útil tener al menos una cota superior del valor que podría alcanzar la incertidumbre combinada cuando las variables están correlacionadas. En esos casos, las medidas relacionadas (p.ej. a y b) se combinan de acuerdo a la fórmula: donde se ha supuesto que la medida y sólo depende de las medidas

$$u^2(y) = c_a^2 u^2(a) + c_b^2 u^2(b) + 2c_a c_b u(a, b) \leq (|c_a u(a)| + |c_b u(b)|)^2$$

a y b. En la práctica si la covarianza  $u(a, b)$  es desconocida, usaríamos el último término de la ecuación como cota superior de la incertidumbre.

## 3. Presentación de resultados numéricos

En la presentación de los resultados, los números nunca se dejan con todas las cifras significativas que da el cálculo matemático, sino que se reducen hasta dejar sólo un número limitado de cifras significativas, pero que continúen describiendo perfectamente el número.

El primer paso consiste en reducir el número de cifras significativas en las incertidumbres, para ello será necesario conocer las reglas generales para el redondeo de números (estas reglas las revisaremos inmediatamente). El segundo paso sería reducir el número de cifras significativas del valor numérico de la magnitud medida, en el que de nuevo utilizaremos las reglas para el redondeo de números.

### 3.1. Reglas para el redondeo de números

Redondear un número consiste en reducir el número de cifras significativas. Primero se decide cuantas cifras significativas se van a dejar en el número redondeado, o bien cuantas cifras de las menos significativas se van a eliminar.

1- Si el dígito más significativo de los que se van a eliminar es menor de 5, el dígito precedente no se cambia. Si fuera mayor que 5, el dígito precedente se aumentaría en una unidad.

*Ejemplos:* 1234,56 se redondea a 3 dígitos como 1230  
45,6789 se redondea a 2 dígitos como 46

Si el dígito más significativo de los que se van a eliminar es 5 y en el resto de los dígitos hay al menos uno mayor que cero, el dígito precedente se incrementa en uno.

*Ejemplos:* 34,15001 se redondea hasta la décima como 34,2  
34,15999 se redondea hasta la décima como 34,2

3- Si el dígito más significativo de los que se van a eliminar es 5 y el resto de los dígitos a eliminar son cero, el dígito precedente se incrementa en uno si es impar, y se deja igual si es un número par. (Siempre resultaría par la cifra final del número redondeado.)

*Ejemplos:* 34,15000 se redondea hasta la décima como 34,2  
34,25000 se redondea hasta la décima como 34,2

4- Si todos los dígitos a eliminar son cero, no se altera ninguna cifra.

*Ejemplo:* 34,15000 se redondea hasta la centésima como 34,15

La regla del punto 3 es arbitraria en el sentido de que podría haberse escogido para favorecer los números acabados en cifra impar; lo importante de esa regla, es asegurar que en el proceso de redondeo se compensen los errores.

En campos donde se precisa la mayor seguridad posible, se admite que siempre se redondee en el sentido de maximizar la seguridad.

### 3.2. Reducción del número de cifras significativas en las incertidumbres

Utilizando las reglas para el redondeo de números.

Método 1: A las incertidumbres sólo se les deja una cifra significativa, y el redondeo se produce con las reglas ya mostradas anteriormente.

Método 2: A las incertidumbres sólo se les dejan dos cifras significativas si la primera cifra es un 1. En los demás casos sólo se deja una cifra significativa. El redondeo se produce con las reglas ya mostradas anteriormente.

Método 3: A las incertidumbres sólo se les dejan dos cifras significativas si la primera cifra es un 1, ó también si es un 2 y la segunda cifra es menor de 5. En los demás casos sólo se deja una cifra significativa. El redondeo se produce con las reglas ya mostradas anteriormente.

Método 4: Si los tres dígitos más significativos de la incertidumbre están entre 100 y 354, se dejan dos cifras significativas. Si esos tres dígitos están entre 355 y 949, se deja sólo una cifra significativa. Finalmente, si esos tres dígitos están entre 950 y 999 se redondea a 1000 y se dejan dos cifras significativas.

Método 5: A las incertidumbres sólo se les dejan dos cifras significativas.

Como se ha observado, en cualquier método, se redondean las cifras según las reglas antes dadas. Algo que debe tenerse en cuenta (en todos los métodos) es que nunca se permitirá un redondeo hacia abajo que implique un cambio en la incertidumbre mayor de un 5%, en ese caso el redondeo se haría hacia arriba. Esto es algo muy importante, que suele olvidarse. Si no se tiene cuidado, podemos encontrarnos con incertidumbres que al redondearse, pierden hasta un 33% de su valor, por lo que el nivel de confianza de la medida se reduce drásticamente. Es por esto, que para empezar a tratar con redondeo de incertidumbres es mejor usar el método 5, que no requiere esa comprobación del 5% (siempre se cumple que es menor del 5%).

Cada rama, cada grupo personas, suele utilizar, por costumbre, un determinado método de reducción de cifras significativas. A veces convendrá cambiar de método según en el campo donde se presenten los resultados. Lo que debe evitarse totalmente, es dar las incertidumbres con más de dos cifras significativas.

<i>Ejemplos:</i>	Método de redondeo:	(1)	(3)	(4)	(5)
	0,10021 se redondea a	0,1	0,10	0,10	0,10
	101,012 se redondea a	100	100	100	100
	121,012 se redondea a	100*	120	120	120
	0,02431 se redondea a	0,02*	0,024	0,024	0,024
	0,02561 se redondea a	0,03	0,03	0,026	0,026
	35,4012 se redondea a	40	40	35	35
	35,5012 se redondea a	40	40	40	36
	0,03552 se redondea a	0,04	0,04	0,04	0,36
	94,1012 se redondea a	90	90	90	94
	0,94912 se redondea a	0,9*	0,9*	0,9*	0,95
	9500,01 se redondea a	10000	10000	10000	9500
	0,95011 se redondea a	1	1	1,0	0,95

Los números marcados con asterisco, tienen un redondeo hacia abajo mayor de un 5%, no deberían darse así, sino que deberían redondearse hacia arriba.

### 3.3. Reducción del número de cifras significativas en valor de la magnitud estimado

Utilizando las reglas para el redondeo de números, se dejan sólo las cifras significativas de peso mayor o igual que las cifras significativas que quedaron en la incertidumbre tras el redondeo.

<i>Ejemplos:</i>	0,12345 con incertidumbre 0,0012 se redondea a 0,1234
	112,543 con incertidumbre 1,1 se redondea a 112,5
	456,781 con incertidumbre 30 se redondea a 460
	500,001 con incertidumbre 100 se redondea a 500
	95,423 con incertidumbre 11 se redondea a 95

(Ejemplos hechos según el método 3 de reducción de cifras significativas, ya que si fuese por el método 4 o el 5, la incertidumbre 30 tendría 2 cifras significativas, y el resultado final sería 457 y no 460)