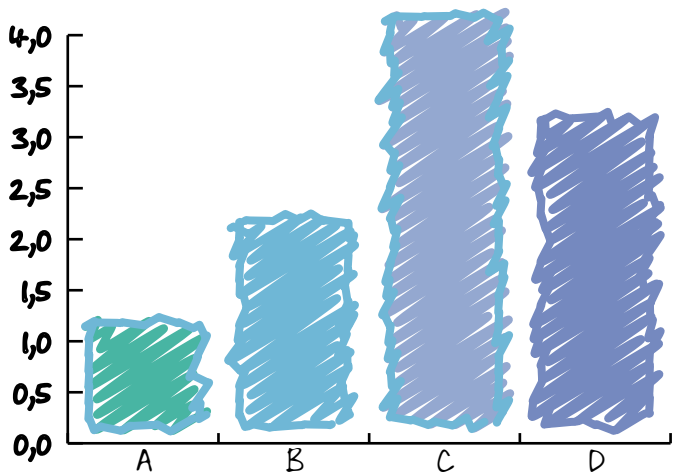
XXXIII Curso de Técnicas Estadísticas

Fase de Presente

Estadística con Excel



Este documento pretende ser una guía de estudio para las asignatura

Estadística con Excel

Tanto la detección de alguna errata como cualquier sugerencia que considere podría redundar en mejorar este documento en futuras entregas, deberían ser puestas en conocimiento del responsable de la asignatura.

Madrid, mayo de 2006

El responsable de la asignatura

ÍNDICE

1 Funciones estadísticas y relacionadas

1.1 Genéricas

- **CONTAR** Cuenta cuántos números hay en la lista de argumentos.
- **CONTARA** Cuenta cuántas celdas no vacías hay en la lista de argumentos.
- **NORMALIZACION**: Devuelve un valor normalizado.
- **PERMUTACIONES**: Devuelve el número de permutaciones para un número determinado de objetos.

1.2 Números aleatorios.

- **ALEATORIO**(): Devuelve un número aleatorio distribuido según una **U[0;1]**
- **ALEATORIO.ENTRE(a;b):** Devuelve un número aleatorio distribuido según una **U[a;b]**

1.3 Descriptivas de datos

- **COEFICIENTE.ASIMETRIA**: Devuelve el sesgo de una distribución
- **CUARTIL**: Devuelve el cuartil de un conjunto de datos
- **CURTOSIS**: Devuelve el coeficiente de curtosis de un conjunto de datos
- **DESVEST**: Calcula la (cuasi) desviación estándar de una muestra. Se pasan por alto los valores lógicos como VERDADERO y FALSO y el texto.
- **DESVESTA**: Calcula la (cuasi) desviación estándar de una muestra, incluidos números, texto y valores lógicos. Los argumentos que contengan VERDADERO se evaluarán como 1; los argumentos que contengan texto o FALSO se evaluarán como 0 (cero).
- **DESVESTP**: Calcula la desviación estándar de la población total. Se pasan por alto los valores lógicos como VERDADERO y FALSO y el texto.
- **DESVESTPA**: Calcula la desviación estándar de la población total, incluidos números, texto y valores lógicos. Los argumentos que contengan VERDADERO se evaluarán como 1; los argumentos que contengan texto o FALSO se evaluarán como 0 (cero).
- **DESVIA2**: Devuelve la suma de los cuadrados de las desviaciones.
- **DESVPROM**: Devuelve el promedio de las desviaciones absolutas de la media de los puntos de datos.
- **FRECUENCIA**: Devuelve una distribución de frecuencia como una matriz vertical.
- **INTERVALO.CONFIANZA**: Devuelve el radio del intervalo de confianza para la media de una población normal, supuesta conocida la varianza (usando la normal).
- **JERARQUIA** Devuelve la jerarquía de un número en una lista de números
- **K.ESIMO.MAYOR**: Devuelve el valor k-ésimo mayor de un conjunto de datos.
- **K.ESIMO.MENOR**: Devuelve el valor k-ésimo menor de un conjunto de datos.
- **MAX**: Devuelve el valor máximo de una lista de argumentos
- **MAXA**: Devuelve el valor máximo de una lista de argumentos, incluidos números, texto y valores lógicos.
- **MEDIA.ACOTADA**: Devuelve la media del interior de un conjunto de datos
- **MEDIA.ARMO**: Devuelve la media armónica.
- **MEDIA.GEOM**: Devuelve la media geométrica.
- **MEDIANA**: Devuelve la mediana de los números dados.
- **MIN**: Devuelve el valor mínimo de una lista de argumentos.

- **MINA**: Devuelve el valor mínimo de una lista de argumentos, incluidos números, texto y valores lógicos.
- **MODA**: Devuelve el valor más frecuente en un conjunto de datos.
- **PERCENTIL**: Devuelve el percentil k-ésimo de los valores de un rango.
- **PROBABILIDAD**: Devuelve la probabilidad de que los valores de un rango estén comprendidos entre dos límites.
- **PROMEDIO**: Devuelve el promedio de los argumentos.
- **PROMEDIOA**: Devuelve el promedio de los argumentos, incluidos números, texto y valores lógicos.
- **RANGO.PERCENTIL**: Devuelve el rango de un valor en un conjunto de datos como porcentaje del conjunto.
- **VAR**: Calcula la varianza de una muestra.
- **VARA**: Calcula la varianza de una muestra, incluidos números, texto y valores lógicos.
- **VARP**: Calcula la varianza de la población total.
- **VARPA**: Calcula la varianza de la población total, incluidos números, texto y valores lógicos.

1.4 Regresión y correlación lineal

- **COEF.DE.CORREL**: Devuelve el coeficiente de correlación de dos conjuntos de datos.
- **COVAR**: Devuelve la covarianza, el promedio de los productos de las desviaciones pareadas.
- **COEFICIENTE.R2**: Devuelve el cuadrado del coeficiente de correlación del momento del producto Pearson.
- **CRECIMIENTO**: Devuelve valores en una tendencia exponencial.
- **ERROR.TIPICO.XY**: Devuelve el error típico del valor de Y previsto para cada valor de X de la regresión.
- **ESTIMACION.LINEAL**: Devuelve los parámetros de una tendencia lineal
- **ESTIMACION.LOGARITMICA**: Devuelve los parámetros de una tendencia exponencial.
- **INTERSECCION.EJE**: Devuelve la intersección de la línea de regresión lineal.
- **PEARSON**: Devuelve el coeficiente de correlación del momento del producto Pearson.
- **PENDIENTE**: Devuelve la pendiente de la línea de regresión lineal
- **PRONOSTICO**: Devuelve un valor en una tendencia lineal.
- **TENDENCIA**: Devuelve los valores que resultan de una tendencia lineal.

1.5 P.Valores

- **PRUEBA.CHI.INV**: Devuelve el inverso de una probabilidad dada, de una sola cola, en una distribución chi cuadrado.
- **PRUEBA.CHI**: Devuelve la prueba de independencia.
- **PRUEBA.F**: Devuelve el resultado de una prueba F.
- **PRUEBA.FISHER.INV**: Devuelve el inverso de la transformación Fisher.
- **PRUEBA.T**: Devuelve la probabilidad asociada a una prueba t de Student.
- **PRUEBA.Z:** Devuelve el valor P de dos colas de una prueba Z.

1.6 Variables aleatorias

- **BINOM.CRIT**: Devuelve el menor valor menor cuya desviación binomial acumulativa es menor o igual que un valor de un criterio.
- **DIST.GAMMA.INV**: Devuelve el inverso de la función gamma acumulativa
- **DIST.GAMMA**: Devuelve la distribución gamma.
- **DISTR.BETA.INV**: Devuelve el inverso de la función de densidad de probabilidad beta acumulativa.
- **DISTR.BETA**: Devuelve la función de densidad de probabilidad beta acumulativa.
- **DISTR.BINOM**: Devuelve la probabilidad de distribución binomial de un término individual.
- **DISTR.CHI**: Devuelve la probabilidad de una sola cola de la distribución chi cuadrado.
- **DISTR.EXP**: Devuelve la distribución exponencial.
- **DISTR.F**: Devuelve la distribución de probabilidad F.
- **DISTR.HIPERGEOM**: Devuelve la distribución hipergeométrica.
- **DISTR.INV.F**: Devuelve el inverso de una distribución de probabilidad F.
- **DISTR.LOG.INV**: Devuelve el inverso de la distribución logarítmico-normal.
- **DISTR.LOG.NORM:** Devuelve la distribución logarítmico-normal acumulativa.
- **DISTR.NORM.ESTAND.INV**: Devuelve el inverso de la distribución normal acumulativa estándar.
- **DISTR.NORM.ESTAND**: Devuelve la distribución normal acumulativa estándar.
- **DISTR.NORM.INV**: Devuelve el inverso de la distribución normal acumulativa.
- **DISTR.NORM**: Devuelve la distribución normal acumulativa.
- **DISTR.T.INV**: Devuelve el inverso de la distribución t de Student.
- **DISTR.T**: Devuelve la distribución t de Student.
- **DISTR.WEIBULL**: Devuelve la distribución Weibull.
- **NEGBINOMDIST**: Devuelve la distribución binomial negativa.
- **POISSON**: Devuelve la distribución de Poisson.

1.7 Otras funciones de interés

ABS

Devuelve el valor absoluto de un número. El valor absoluto de un número es el número sin su signo.

ABS(número)

- **Número** es el número real cuyo valor absoluto desea obtener.

COINCIDIR

Devuelve la posición relativa de un elemento en una matriz que coincida con un valor especificado en un orden especificado. Utilice COINCIDIR en lugar de las funciones BUSCAR cuando necesite conocer la posición de un elemento en un rango en lugar del elemento en sí.

COINCIDIR(valor_buscado;matriz_buscada;tipo_de_coincidencia)

COCIENTE

Devuelve la parte entera de una división. Use esta función cuando desee descartar el residuo de una división. Si esta función no está disponible, ejecute el progra-

ma de instalación e instale las Herramientas para análisis. Para instalar este complemento, elija *Complementos* en el menú *Herramientas* y seleccione la casilla correspondiente.

COCIENTE(numerador; denominador)

- **Numerador** es el dividendo.
- **Denominador** es el divisor.

Observaciones

- Si uno de los argumentos no es un valor numérico, COCIENTE devuelve el valor de error #¡VALOR!
- COCIENTE(5; 2) es igual a 2
- COCIENTE(4,5; 3,1) es igual a 1
- COCIENTE(-10; 3) es igual a -3

CONTAR.SI

Cuenta las celdas, dentro del rango, que no están en blanco y que cumplen con el criterio especificado.

CONTAR.SI(rango; criterio)

- **Rango** es el rango dentro del cual desea contar el número de celdas que no están en blanco.
- **Criterio** es el criterio en forma de número, expresión o texto, que determina las celdas que se van a contar.

ENTERO

Devuelve un número hasta el entero inferior más próximo.

ENTERO(número)

- **Número** es el número real que desea redondear al entero inferior más próximo.

FACT

Devuelve el factorial de un número. El factorial de un número es igual a 1*2*3*...* número.

FACT(número)

- **Número** es el número no negativo cuyo factorial desea obtener. Si el argumento número no es un entero, se trunca.

NOD

Devuelve el valor de error #N/A, que significa "no hay ningún valor disponible". Utilice #N/A para marcar las celdas vacías. Si escribe #N/A en las celdas donde le falta información, puede evitar el problema de la inclusión no intencionada de celdas vacías en los cálculos. (Cuando una fórmula hace referencia a una celda que contiene #N/A, la fórmula devuelve el valor de error #N/A.)

NOD( )

- Debe incluir paréntesis vacíos con el nombre de la función. De lo contrario no se reconocerá como función.
- También puede escribir el valor #N/A directamente en la celda. La función NOD se proporciona por compatibilidad con otros programas para hojas de cálculo.

NUMERO.ROMANO

Convierte un número arábigo en número romano con formato de texto.

NUMERO.ROMANO(número; forma)

- **Número** es el número arábigo que desea convertir.
- **Forma** es un argumento que especifica la forma de número romano que desea. El estilo de número romano varía entre clásico y simplificado; cuanto más aumenta el valor del argumento forma, más conciso es el estilo devuelto. Vea los ejemplos siguientes.

PRODUCTO

Multiplica todos los números que figuran como argumentos y devuelve el producto.

PRODUCTO(número1;número2; ...)

- **Número1**; número2; ... son entre 1 y 30 números que desea multiplicar.
- Los argumentos que son números, valores lógicos o representaciones textuales de números se toman en cuenta; los argumentos que son valores de error o texto que no se puede convertir en números causan errores.
- Si un argumento es una matriz o una referencia, sólo se tomarán en cuenta los números de la matriz o de la referencia. Se pasan por alto las celdas vacías, valores lógicos, texto o valores de error en la matriz o en la referencia.

REDONDEA.PAR REDONDEA.IMPAR

Devuelve un número redondeado hasta el número entero par (impar) más próximo. Esta función puede usarse para procesar artículos que vienen en pares.

REDONDEA.PAR(número)

- **Número** es el valor que desea redondear.
- Si el argumento número es un valor no numérico, REDONDEA.PAR devuelve el valor de error #¡VALOR!
- Cuando un valor se ajusta alejándose de cero, se redondeará hacia arriba, independientemente del signo del número. Si el argumento número es un entero par, no se redondea.

REDONDEAR

Redondea un número al número de decimales especificado.

REDONDEAR(número;núm_de_decimales)

- **Número** es el número que desea redondear.
- **Núm_de_decimales** especifica el número de dígitos al que desea redondear el argumento número.
- Si el argumento núm_de_decimales es mayor que 0 (cero), número se redondeará al número de lugares decimales especificado.
- Si el argumento núm_de_decimales es 0, número se redondeará al entero más próximo.
- Si el argumento núm_de_decimales es menor que 0, número se redondeará hacia la izquierda del separador decimal.
- REDONDEAR(2,15; 1) es igual a 2,2
- REDONDEAR(2,149; 1) es igual a 2,1

- REDONDEAR(-1,475; 2) es igual a -1,48
- REDONDEAR(21,5; -1) es igual a 20

REDONDEAR.MAS

Redondea un número hacia arriba, en dirección contraria a cero.

REDONDEAR.MAS(número;núm_decimales)

- **Número** es cualquier número real que desee redondear.
- **Núm_decimales** es el número de decimales al cual desea redondear el número.
- La función REDONDEAR.MAS es similar a la función REDONDEAR, excepto que siempre redondea al número superior más próximo, alejándolo de cero.
- Si el argumento núm_decimales es mayor que 0 (cero), el número se redondea al valor superior (inferior para los números negativos) más próximo que contenga el número de lugares decimales especificado.
- Si el argumento núm_decimales es 0 o se omite, el número se redondea al entero superior (inferior si es negativo) más próximo.
- Si el argumento núm_decimales es menor que 0, el número se redondea al valor superior (inferior si es negativo) más próximo a partir de la izquierda de la coma decimal.
- REDONDEAR.MAS(3,2;0) es igual a 4
- REDONDEAR.MAS(76,9;0) es igual a 77

RESIDUO

Devuelve el residuo o resto de la división entre número y núm_divisor. El resultado tiene el mismo signo que núm_divisor.

RESIDUO(número;núm_divisor)

- **Número** es el número que desea dividir y cuyo residuo o resto desea obtener.
- **Núm_divisor** es el número por el cual desea dividir número. Si núm_divisor es 0, RESIDUO devuelve el valor de error #¡DIV/0!
- La función RESIDUO se puede expresar utilizando la función ENTERO:
- RESIDUO(n;d) = n - d*ENTERO(n/d)
- RESIDUO(3; 2) es igual a 1
- RESIDUO(-3; 2) es igual a 1
- RESIDUO(3; -2) es igual a -1
- RESIDUO(-3; -2) es igual a -1

SUBTOTALES

Devuelve un subtotal en una lista o base de datos. Generalmente es más fácil crear una lista con subtotales utilizando el comando Subtotales del menú Datos. Una vez creada la lista de subtotales, puede cambiarse modificando la fórmula SUBTOTALES.

SUBTOTALES(núm_función;ref1)

- **Núm_función** es un número de 1 a 11 que indica qué función debe ser utilizada para calcular los subtotales dentro de una lista.
- **Ref1** es el rango o referencia para el cual desea calcular los subtotales.

Núm_función	Función
1	PROMEDIO
2	CONTAR
3	CONTARA
4	MAX
5	MIN
6	PRODUCTO
7	DESVEST
8	DESVESTP
9	SUMA
10	VAR
11	VARP

- Si hay otros subtotales dentro de ref1 (o subtotales anidados), estos subtotales anidados se pasarán por alto para no repetir los cálculos.
- La función SUBTOTALES pasa por alto las filas ocultas. Esto es importante cuando sólo desea obtener el subtotal de los datos visibles que resulta de una lista filtrada.

SUMA

Suma todos los números de un rango.

SUMA(número1;número2; ...)

- **Número1; número2**; .. son entre 1 y 30 números cuya suma desea obtener.
- Se toman en cuenta números, valores lógicos y representaciones de números que escriba directamente en la lista de argumentos. Consulte los dos primeros ejemplos.
- Si un argumento es una matriz o una referencia, solamente se contarán los números de esa matriz o referencia. Se pasan por alto las celdas vacías, valores lógicos, texto o valores de error en esa matriz o referencia.
- Los argumentos que sean valores de error o texto que no se pueda traducir a números causarán errores.

SUMA.CUADRADOS

Devuelve la suma de los cuadrados de los argumentos.

SUMA.CUADRADOS(número1;número2; ...)

- **Número1; número2**; ... son de 1 a 30 argumentos para los cuales desea obtener la suma de sus cuadrados. También puede usar una sola matriz o una referencia a una matriz en lugar de argumentos separados con punto y coma.

SIGNO

Devuelve el signo de un número. Devuelve 1 si el argumento número es positivo, 0 si el argumento número es 0 y -1 si el argumento número es negativo.

SIGNO(número)

- **Número** es un número real cuyo signo desea saber.

SUMAR.SI

Suma las celdas en el rango que coinciden con el argumento criterio.

SUMAR.SI(rango;criterio;rango_suma)

- **Rango** es el rango de celdas que desea evaluar.
- **Criterio** es el criterio en forma de número, expresión o texto, que determina qué celdas se van a sumar.
- **Rango_suma** son las celdas que se van a sumar. Las celdas contenidas en rango_suma se suman sólo si las celdas correspondientes del rango coinciden con el criterio. Si rango_suma se omite, se suman las celdas contenidas en el argumento rango.

TRUNCAR

Trunca un número a un entero, suprimiendo la parte fraccionaria de dicho número.

TRUNCAR(número; núm_de_decimales)

- **Número** es el número que desea truncar.
- **Núm_de_decimales** es un número que especifica la precisión al truncar. El valor predeterminado del argumento núm_de_decimales es 0.
- TRUNCAR y ENTERO son similares, ya que ambos devuelven enteros. TRUNCAR suprime la parte fraccionaria del número. ENTERO redondea los números al entero menor más próximo, según el valor de la porción fraccionaria del número. ENTERO y TRUNCAR son diferentes solamente cuando se usan números negativos: TRUNCAR(-4,3) devuelve -4, pero ENTERO(-4,3) devuelve -5, ya que -5 es el número entero menor más cercano.
- TRUNCAR(8,9) es igual a 8
- TRUNCAR(-8,9) es igual a -8
- TRUNCAR(PI()) es igual a 3

2 Números aleatorios.

2.1 Procedimientos relacionados

Excel cuenta con dos procedimientos para obtener números aleatorios distribuidos según una forma determinada:

a) Utilizar las funciones ALEATORIO() y ALEATORIO.ENTRE(a;b)

b) Recurriendo al complemento de **Análisis de Datos.**

El segundo procedimiento se describe en el apartado 13.11 de este documento.

Por el primer procedimiento obtendremos números de una distribución Uniforme: ALEATORIO() según una $U_{[0;1]}$; ALEATORIO.ENTRE(a;b): según una $U_{[a;b]}$ discreta. Estas funciones son volátiles de manera que se recalcularan cada vez (si la opción de cálculo está puesta en automático).

- **ALEATORIO** Devuelve un número aleatorio mayor o igual que 0 y menor que 1, distribuido uniformemente. Cada vez que se calcula la hoja de cálculo, se devuelve un número aleatorio nuevo.

Su sintaxis es

ALEATORIO( )

- **ALEATORIO.ENTRE** Devuelve un número aleatorio entre los números que especifique. Devuelve un nuevo número aleatorio cada vez que se calcula la hoja de cálculo. Si esta función no está disponible, ejecute el programa de instalación e instale las Herramientas para análisis. Para instalar este complemento, elija **Complementos** en el menú **Herramientas** y active la casilla correspondiente.

Su sintaxis es

ALEATORIO.ENTRE(inferior; superior)

- **Inferior** es el menor número entero que la función ALEATORIO.ENTRE puede devolver.
- **Superior** es el mayor número entero que la función ALEATORIO.ENTRE puede devolver.

A pesar de contar únicamente con funciones para generar números distribuidos de forma uniforme podemos gracias a éstas, generar prácticamente cualquier distribución utilizando bien algoritmos descritos en la literatura, bien las funciones inversas cuando éstas están implementadas en Excel.

La tabla siguiente representa este procedimiento para algunas de las funciones continuas más comunes:

Beta (α,β)	DISTR.BETA.INV(ALEATORIO();α,β, a , b).
χ^2_{GL}	PRUEBA.CHI.INV(ALEATORIO();GL)
Exponencial(β)	(1/β) * -LOG(ALEATORIO())
F(GL_1;GL_2)	DISTR.F.INV(ALEATORIO();GL_1;GL_2)
Gamma(α,β)	DISTR.GAMMA.INV(ALEATORIO();α;β)
LogNormal(μ,σ)	DISTR.LOG.INV(ALEATORIO();μ;σ)
Normal(μ,σ)	DISTR.NORM.INV(ALEATORIO();μ;σ)
	μ+σ*(RAIZ(-2*LOG(ALEATORIO()))*SENO(2*PI()*ALEATORIO()))
Triangular (a,b,c)	c + (a+ALEATORIO()*(b-a)-c)*MAX(ALEATORIO();ALEATORIO())
	c + (a+ALEATORIO()*(b-a)-c)*RAIZ(ALEATORIO())
T_{GL}	DISTR.T.INV(ALEATORIO();GL)*SIGNO(ALEATORIO()-0,5)
Pareto(α,β)	β*((1/(1-ALEATORIO()))^(1/α))
	β*(ALEATORIO()^(-1/α))

2.2 Dos funciones interesantes

- **INDICE** Devuelve el elemento del rango matriz que ocupa la posición dada por los índices de número de fila y de columna.

Sintaxis

INDICE(**matriz**; núm_fila; núm_columna)

- **Matriz**: es un rango de celdas o una matriz de constantes.
- Si matriz contiene sólo una fila o columna, el argumento núm_fila o núm_columna que corresponde es opcional.
- Si matriz tiene más de una fila y más de una columna y sólo utiliza núm_fila o núm_columna, INDICE devuelve una matriz con toda una fila o columna.

Observaciones

- Si se utilizan ambos argumentos núm_fila y núm_columna, INDICE devuelve el valor en la celda de intersección de los argumentos núm_fila y núm_columna.
- Si se define núm_fila o núm_columna como 0 (cero), INDICE devuelve la matriz de valores de toda la columna o fila, respectivamente. Para utilizar valores devueltos como una matriz, introduzca la función INDICE como una fórmula matricial en un rango horizontal de celdas para una fila y en un rango vertical de celdas para una columna. Para introducir una fórmula matricial, presione CTRL+MAYÚS+ENTRAR.
- Los argumentos núm_fila y núm_columna deben indicar una celda contenida en matriz; de lo contrario, INDICE devuelve el valor de error #¡REF!
- **JERARQUIA** Devuelve la "jerarquía" de un número dentro de una lista. La "jerarquía" de un número es su posición en la lista si ésta se considerara ordenada de menor a mayor

Sintaxis

JERARQUIA(número ; referencia ; orden)

- **número**: es el número cuya jerarquía desea conocer.
- **referencia**: es una matriz de o una referencia a una lista de números. Los valores no numéricos se pasan por alto.
- orden: es un número que especifica cómo clasificar el argumento número.

2.3 PROBLEMAS

2.3.1 Generar dos muestras de 100 valores cada una, comprendidos entre 0 y 1. Comprobar mediante un gráfico que se respetan los límites previstos.

2.3.2 Comprobar el efecto de la tecla "Calcular" (F9).

2.3.3 Simular el lanzamiento 100 veces de un dado equilibrado.

2.3.4 Simular el experimento "lanzar dos dados y calcular la suma de ambos".

2.3.5 Simular una distribución Uniforme no discreta $U_{[0,100]}$.

2.3.6 Generar una muestra (n=25) de una distribución $\mathbf{N}_{(10;1)}$ usando las dos fórmulas dadas en la tabla. Ordenar los valores obtenidos de menos a mayor.

2.3.7 Estimar mediante MonteCarlo la probabilidad de que al colocar 5 números distintos al azar, al menos dos de ellos sean consecutivos. (Utilizar la función JERARQUIA aplicada sobre un conjunto de números aleatorios para obtener un muestro sin reemplazamiento)

2.3.8 ¿Estimar la probabilidad de que al escribir **n** cartas y sus correspondientes **n** sobres y colocarlos al azar las unas en los otros, al menos uno de ellos contenga la carta correcta?

2.3.9 Un grupo de **2N** chicos y **2N** chicas se divide en dos grupos iguales. Hallar la probabilidad de que cada grupo tenga igual número de personas de cada sexo.

2.3.10 ¿Cuál de los tres sucesos siguientes es más probable?: SIXTO RÍOS Pág. 33 problema 27

a) obtener al menos un 6 al lanzar 6 dados.
b) al menos dos 6 al lanzar 12 dados.
c) al menos tres 6 al lanzar 18 dados.

Se trata de un problema elemental de probabilidad cuya solución analítica es evidente: la única forma de no sacar al menos un 6 (1/6 de probabilidad al aplicar el criterio de Laplace) es que ninguno de los lanzamiento lo sea, es decir, al lanzar n dados la probabilidad es: Pn = 1-(5/6)n. No obstante procederemos a la simulación de las tres alternativas para comprobar la validez del procedimiento de aproximación basado en el método de MonteCarlo.

2.3.11 Un jugador apuesta por uno de los dígitos 1,2,3,4,5 o 6. Se lanzan tres dados, si en uno, dos o tres de los dados sale el número apostado el jugador recibe dos, tres o cuatro veces su apuesta; si no sale su número, pierde lo apostado. SIXTO RÍOS Pág. 61 problema 21.

a) Simular 100 veces el experimento.
b) ¿Cuál es la esperanza de pérdida si apuesta n euros?.

2.3.12 Dos personas deciden verse para lo cual se citan en un determinado lugar ofreciéndose cada una llegar entre las 6 y las 6:50 y no esperar a la otra más de 10 minutos, estando como mucho hasta las 7. ¿Cuál es la probabilidad de que lleguen a encontrarse?.

2.3.13 Una compañía aérea vende sus billetes a 15 euros. Cada pasajero suponen un coste de 3 euros. El avión tiene 100 plazas. La probabilidad de que se presente un pasajero con reserva previa es P. Los billetes reservados y no atendidos (*overbooking*) se compensan con 30 euros. Optimizar el número máximo posible de reservas.

2.3.14 Sixto Ríos (1983) refiere el siguiente problema "Un sultán tiene el propósito de establecer un política de control de la natalidad que incremente la proporción de mujeres de la población. Para ello adopta promulga el siguiente edicto: Tan pronto como una mujer tenga su primer hijo le estará prohibido

tener más descendencia". Suponiendo que la probabilidad de que nazca un niño es igual a la de que nazca una niña, comentar la eficacia del edicto del sultán.

La manera que proponemos de abordar el problema es la siguiente: supondremos un número N = 20 de mujeres a las que hacemos parir un número suficiente de hijos (digamos que también 20) cuyo género se adapte a las probabilidades de nacimiento de cada uno. Tendremos que generar para cada mujer, un experimento de Bernouilli de probabilidad p, cosa que en Excel es extraordinariamente sencillo ya que basta con asignar a la celda en cuestión la fórmula siguiente:

IF(ALEATORIO()<=p;Éxito;Fracaso)

Donde p es la probabilidad del suceso que hemos llamado (arbitrariamente éxito); así, si como es nuestro caso, tenemos que:

	Éxito	Fracaso
Probabilidad	p	1-p
Suceso	Mujer(M)	Hombre (H)

bastará que, suponiendo que la celda Ref contiene el valor de p escribamos la fórmula

IF(ALEATORIO()<=Ref;"M";"H")

hecho esto tendremos asociada a cada madre un progenie aleatoria distribuida con arreglo al valor de p; por ejemplo:

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
H	H	M	M	M	H	M	H	M	H	M	M	M	M	M	M	M	M	M	H
H	M	M	H	M	H	H	H	H	M	H	M	M	M	H	H	H	M	M	H
M	M	M	M	M	H	M	H	H	H	H	M	M	M	M	M	M	H	M	H
H	M	H	M	M	H	H	H	H	H	H	M	M	H	H	M	H	M	M	M
H	H	H	M	H	M	M	H	M	M	M	H	M	H	H	H	H	M	H	M
M	M	H	H	H	H	H	M	M	M	M	H	M	H	M	H	H	M	M	M
H	M	M	M	H	H	M	M	M	M	M	M	M	M	M	M	H	M	H	H
M	M	M	H	M	M	M	M	M	H	M	M	M	M	M	H	H	M	H	H
M	H	M	M	M	H	H	M	H	H	H	M	M	H	H	M	M	M	M	M

Ahora sólo queda aplicar el edicto del sultán contando únicamente los hijos tenidos hasta que hubiera aparecido la primera "H", para ello utilizamos la función de Excel COINCIDIR, cuya sintaxis es:

COINCIDIR(valor_buscado;matriz_buscada;tipo_de_coincidencia)

- *Valor_buscado es el valor que se usa para encontrar el valor deseado en la tabla.*
- *Matriz_buscada es un rango múltiple de celdas que contienen posibles valores a buscar*
- *Tipo_de_coincidencia es el número -1, 0 ó 1 y especifica cómo hace coincidir*

Puesto que buscamos las "H" la fórmula será:

=COINCIDIR("H";ICol:Fcol;0)

siendo Icol:Fcol las direcciones en las que buscar, es decir la progenie de cada mujer sin considerar el edicto.

El número obtenido por la aplicación de esta fórmula será el ordinal del primer varón de la progenie. Bastará entonces con sumar estos números (restándole una unidad a cada uno de ellos) para obtener el número de Hijas, siendo el número de Hijos igual al de madres consideradas. La estimación de las proporciones finales de unos y otros en la población, nos llevará a concluir que el edicto del sultán, no sólo es vejatorio para sus súbditos, sino que además es absolutamente inútil por cuanto no cumple el objetivo con el que fue promulgado.

3 Distribución de frecuencias.

3.1 Procedimientos relacionados

Excel cuenta con dos procedimientos para obtener la distribución de frecuencias de una variable:

a) Utilizar la función FRECUENCIA.

b) Recurrir al complemento de **Análisis de Datos (HISTOGRAMA).**

El segundo procedimiento se describe en el apartado 13.9 de este documento.

La función de Excel para la obtención de las distribución de frecuencias es:

- **FRECUENCIA**: Devuelve una distribución de frecuencia como una matriz vertical

Su sintaxis es

FRECUENCIA(**datos**; **grupos**)

- **Datos**: es una matriz de un conjunto e valores o una referencia a un conjunto de valores cuyas frecuencias desea contar. Si datos no contiene ningún valor, FRECUENCIA devuelve una matriz de ceros.
- **Grupos**: es una matriz de intervalos o una referencia a intervalos dentro de los cuales desea agrupar los valores del argumento datos. Si grupos no contiene ningún valor, FRECUENCIA devuelve el número de elementos contenido en datos.

Observaciones

- FRECUENCIA se introduce como una **fórmula matricial** después de seleccionar un rango de celdas adyacentes en las que se desea que aparezca el resultado de la distribución.
- El número de elementos de la matriz devuelta supera en una unidad el número de elementos de grupos. El elemento adicional de la matriz devuelta devuelve la suma de todos los valores superiores al mayor intervalo.
- La función FRECUENCIA pasa por alto celdas en blanco y texto.

Una observación se cuenta como perteneciente al intervalo cuya marca de clase es C_i si se verifica que: $C_{i-1} < x_i \leq C_i$

1 $x_i \leq 1$

2 $1 < x_i \leq 2$

3 $2 < x_i \leq 3$

$3 < x_i$

En general:

$C_{i-1} < x_i \leq C_i$

3.2 PROBLEMAS

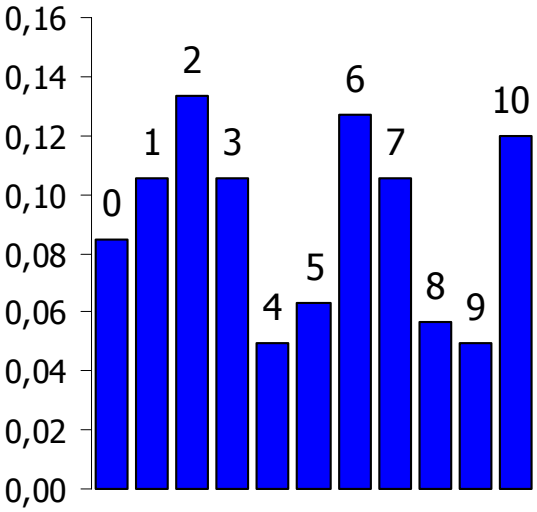
3.2.1 Generar 100 valores (discretos) comprendidos entre 0 y 10.

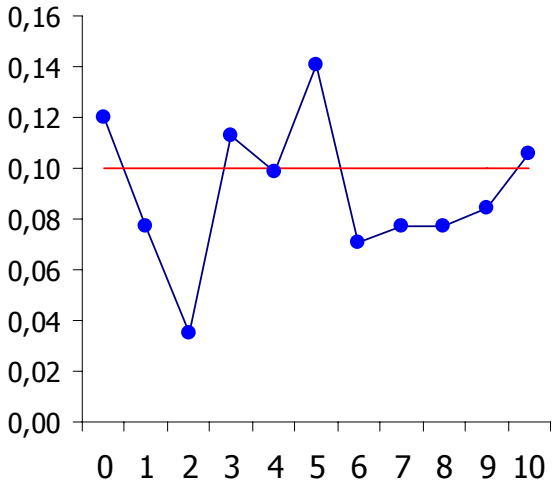
a) Construir la tabla de distribución de frecuencias de dichos valores.

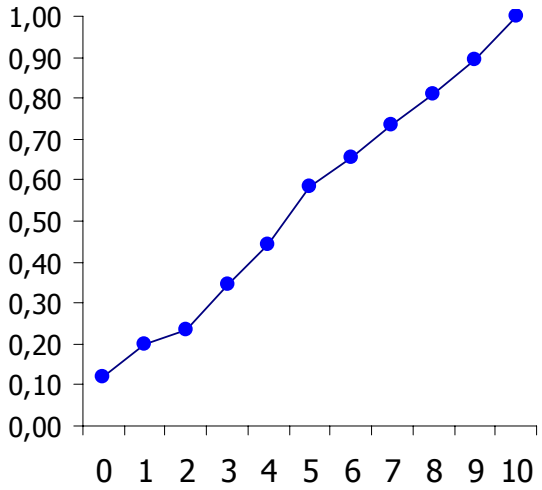
b) Representar gráficamente los valores mediante un diagrama de barras y polígonos de frecuencias acumulados y no acumulados.

Distribución de frecuencias

Ci	ni	fi	Ni	Fi
0	12	0,084507	12	0,084507
1	15	0,105634	27	0,190141
2	19	0,133803	46	0,323944
3	15	0,105634	61	0,429577
4	7	0,049296	68	0,478873
5	9	0,063380	77	0,542254
6	18	0,126761	95	0,669014
7	15	0,105634	110	0,774648
8	8	0,056338	118	0,830986
9	7	0,049296	125	0,880282
10	17	0,119718	142	1,000000
	142	1		







3.2.2 Copiar la siguiente fórmula descrita anteriormente:

=500+200*(RAIZ(-2*LOG(ALEATORIO()))*SENO(2*PI()*ALEATORIO()))

$$\sqrt{-2\cdot \mathrm{Ln}(U)}\cdot \mathrm{sen}(2\pi U) \quad U \approx \mathbf{U}_{[0;1]}$$

y utilizarla para generar 1000 valores de una variable aleatoria **N**(500;200). Una vez obtenidos los valores,

a) Construir la tabla de su distribución de frecuencias,
b) Realizar un histograma de los valores,
c) Aproximar su función de densidad mediante un polígono de frecuencias.

		Ci		ni	fi	Ni	Fi
A	*0*	**25**	*50*	2	0,0004	2	0,0004
	50	**75**	*100*	6	0,0012	8	0,0016
	100	**125**	*150*	14	0,0029	22	0,0045
	150	**175**	*200*	44	0,0091	66	0,0136
	200	**225**	*250*	79	0,0163	145	0,0298
	250	**275**	*300*	182	0,0375	327	0,0673
	300	**325**	*350*	299	0,0615	626	0,1288
	350	**375**	*400*	459	0,0945	1085	0,2233
	400	**425**	*450*	597	0,1229	1682	0,3462
	450	**475**	*500*	745	0,1533	2427	0,4995
	500	**525**	*550*	731	0,1504	3158	0,6499
	550	**575**	*600*	597	0,1229	3755	0,7728
	600	**625**	*650*	485	0,0998	4240	0,8726
	650	**675**	*700*	308	0,0634	4548	0,9360
	700	**725**	*750*	144	0,0296	4692	0,9656
	750	**775**	*800*	108	0,0222	4800	0,9879
	800	**825**	*850*	39	0,0080	4839	0,9959
	850	**875**	*900*	15	0,0031	4854	0,9990
	900	**925**	*950*	4	0,0008	4858	0,9998
B	*950*	**975**	*1000*	1	0,0002	4859	1,0000
				4859	1		
				4859			

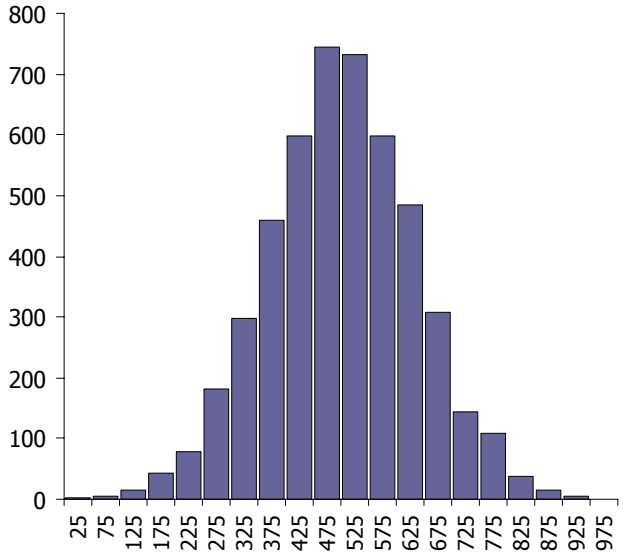
A	MIN/MAX		
B	Redondeo a enteros		
C	Redondeo a múltiplos de 10		
	A	**B**	**C**
Min	12,6	12,0	0,0
Max	984,8	985,0	990,0

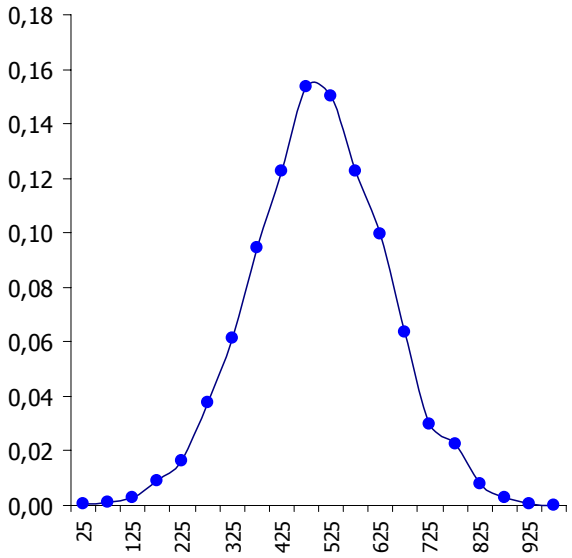
Rango

A	MIN(B:B)
B	REDONDEAR.MENOS(MIN(B:B);0)
C	REDOND.MULT(MAX(0;MIN(B:B)-10);10)

Intervalos

70	REDONDEAR.MAS(RAIZ(CONTAR(B:B));0)
70	ENTERO(RAIZ(CONTAR(B:B)))+1





4 Medidas de tendencia central, variación y forma.

4.1 Procedimientos relacionados

Excel cuenta con dos procedimientos para obtener la descripción mediante estadísticos de una muestra o una población:

a) Utilizar las funciones relacionadas expuestas a continuación.

b) Recurrir al complemento de **Análisis de Datos** (Estadística Descriptiva). Descrito en el apartado 13.5 de este documento.

4.2 Funciones para el cálculo de la tendencia central.

Media.

- **PROMEDIO**: Devuelve la media aritmética de los argumentos.
- **PROMEDIOA**: Devuelve la media incluidos texto y valores lógicos.
- **MEDIA.ACOTADA**: Devuelve la media recortada de un conjunto de datos

  MEDIA.ACOTADA(matriz ; porcentaje)

  Matriz es la matriz o el rango de valores que desea acotar y de los cuales se calculará la media. **Porcentaje** es el número fraccionario de puntos de datos que se excluyen del cálculo. Por ejemplo, si porcentaje = 0,2, se eliminarán cuatro puntos de un conjunto de datos de 20 puntos (20 x 0,2), dos de la parte superior y dos de la parte inferior.

- **MEDIA.ARMO**: Devuelve la media armónica.
- **MEDIA.GEOM**: Devuelve la media geométrica.

Mediana.

- **MEDIANA**: Devuelve la mediana de los números dados.

Moda.

- **MODA**: Devuelve el valor más frecuente en un conjunto de datos.

4.3 Funciones para el cálculo de la variación.

Rango medio.

- Usar MAX y MIN:

  RM = PROMEDIO(MAX(Datos)+MIN(Datos))

Cuartiles.

- **CUARTIL**:

  CUARTIL(matriz ; cuartil)

  Matriz, es la matriz o rango de celdas de valores numéricos cuyo cuartil desea obtener. Cuartil, indica el valor que se devolverá, el código es (0 = mínimo; 1 = primer cuartil; 2 = mediana; 3 = tercer cuartil;4 = máximo).

- **PERCENTIL**: Devuelve el percentil k-ésimo de los valores de un rango

  PERCENTIL(matriz ; k)

  Matriz es la matriz o rango de datos que define la posición relativa. **K** es el valor de percentil en el intervalo de 0 a 1, inclusive.

- **RANGO.PERCENTIL**: Devuelve el % del los valores que son menores que cifra dentro de matriz

  RANGO.PERCENTIL(matriz;x;cifra_significativa)

Matriz es la matriz o rango de datos con valores numéricos que define la posición relativa. **X**, es el valor cuyo rango percentil desea conocer. **Cifra_significativa** es un valor opcional que identifica el número de cifras significativas para el valor de porcentaje devuelto. Si se omite este argumento, RANGO.PERCENTIL utiliza tres dígitos.

Rango intercuartílico.

- Usar cualquiera de las dos alternativas siguientes:

  PERCENTIL (RI = PERCENTIL(datos;0,75)-PERCENTIL(datos;0,25))

  CUARTIL (RI = CUARTIL(datos;3)- CUARTIL(datos;1)).

Varianza y desviación típica.

- **VAR**(A): Calcula la (cuasi)varianza de una muestra.
- **VARP**(A): Calcula la varianza de la población.
- **DESVEST**(A): Calcula la (cuasi) desviación estándar de una muestra.
- **DESVESTP**(A): Calcula la desviación estándar de la población total.

Coeficiente de variación.

- Usar PROMEDIO y DESVEST

4.4 Funciones para el cálculo de la forma.

Simetría

- **COEFICIENTE.ASIMETRIA**: Devuelve el sesgo de una distribución

Curtosis

- **CURTOSIS**: Devuelve el coeficiente de curtosis de un conjunto de datos

4.5 PROBLEMAS

4.5.1 Para los datos siguientes

{7,4,9,7,3,12}

Calcular todos los estadísticos descritos anteriormente

Datos		
7	Media	7,00
4	Mediana	7,00
9	Moda	7,00
7	Rango Medio	4,5
3	Eje medio	6,625
12	Rango	9
	Rango intercualtílico	3,75
	Varianza	10,80
	Desviación	3,29
	Coeficiente de variación	0,3550

4.5.2 Para los datos anteriores, comprobar los resultados de las siguientes funciones:

a) PROMEDIO;
b) MEDIA.ARMO;
c) MEDIA.GEOM;
d) VAR;
e) VARP;
f) COEFICIENTE.ASIMETRIA;
g) CURTOSIS

con los obtenidos al hacer los cálculos directamente sobre la hoja.

4.5.3 Para los datos anteriores calcular los momentos de tercer y cuarto orden centrados en torno a la media.

$$m_3 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^3 \quad ; \quad m_4 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^4$$

4.5.4 Para el siguiente conjunto de datos

1, 4, 3, 8, 9, 10, 10, 7, 3, 1, 8, 7, 5, 5, 8, 10, 1

calcular la MEDA, definida como:

$$MEDA_X = \text{mediana}\left\{|x_1 - med_x|; |x_2 - med_x|; \cdots; |x_N - med_x|\right\}$$

1	4	3	8	9	10	10	7	3	1	8	7	5	5	8	10	1	**7**

6	3	4	1	2	3	3	0	4	6	1	0	2	2	1	3	6	**3**

4.5.5 ¿Cuál es la media geométrica del siguiente conjunto de valores?

-1, 3, 9

4.5.6 Suponga el siguiente conjunto de datos

13, 15, 14, 17, 13, 16, 15, 16, 16

a) Suponga que por error, el último dato se introdujo como 61 en vez de 16. Compare la descripción de los dos conjunto de datos.

A	B
13	13
15	15
14	14
17	17
13	13
16	16
15	15
16	16
16	61

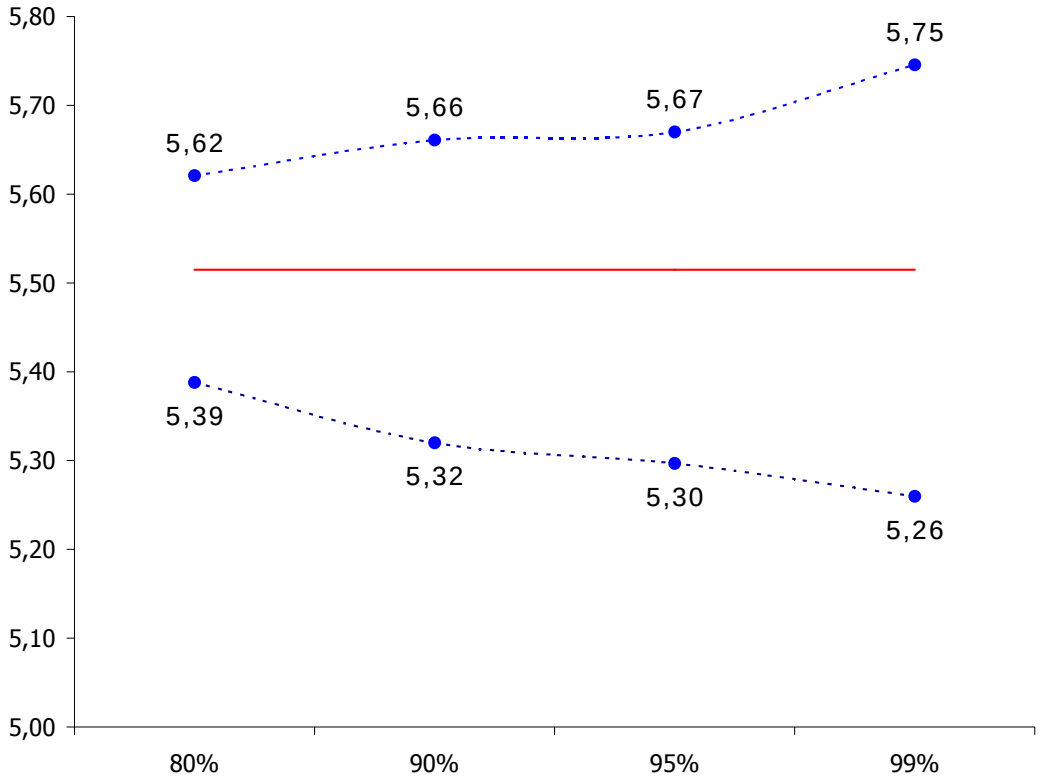
	A	B
Media	15	20
Error típico	0,471	5,145
Mediana	15	15
Moda	16	13
Desviación estándar	1,414	15,435
Varianza de la muestra	2	238,25
Curtosis	-1,089	8,800
Coeficiente de asimetría	-0,341	2,955
Rango	4	48
Mínimo	13	13
Máximo	17	61
Suma	135	180
Cuenta	9	9
Mayor (1)	17	61
Menor(1)	13	13
Nivel de confianza(95,0%)	1,09	11,86

4.5.7 Para los datos del ejercicio 3.45 (Pág. 143)

5,65	5,34	5,57	5,62	5,47	5,32	5,77	5,50	5,61	5,63
5,44	5,54	5,40	5,56	5,40	5,67	5,57	5,32	5,45	5,50
5,42	5,45	5,53	5,46	5,47	5,29	5,42	5,50	5,44	5,57
5,40	5,52	5,54	5,44	5,61	5,49	5,58	5,53	5,25	5,67
5,53	5,41	5,55	5,51	5,53	5,55	5,58	5,58	5,56	5,36

a) Calcular un intervalo centrado en torno a la mediana que contenga el {80%;90%;95%;99%} de los datos.

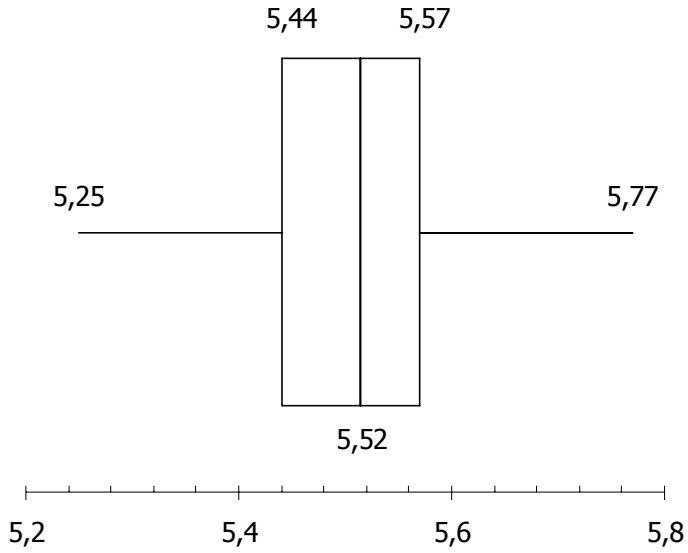
b) Hacer un gráfico que incluya la mediana.



4.5.8 Con los datos anteriores

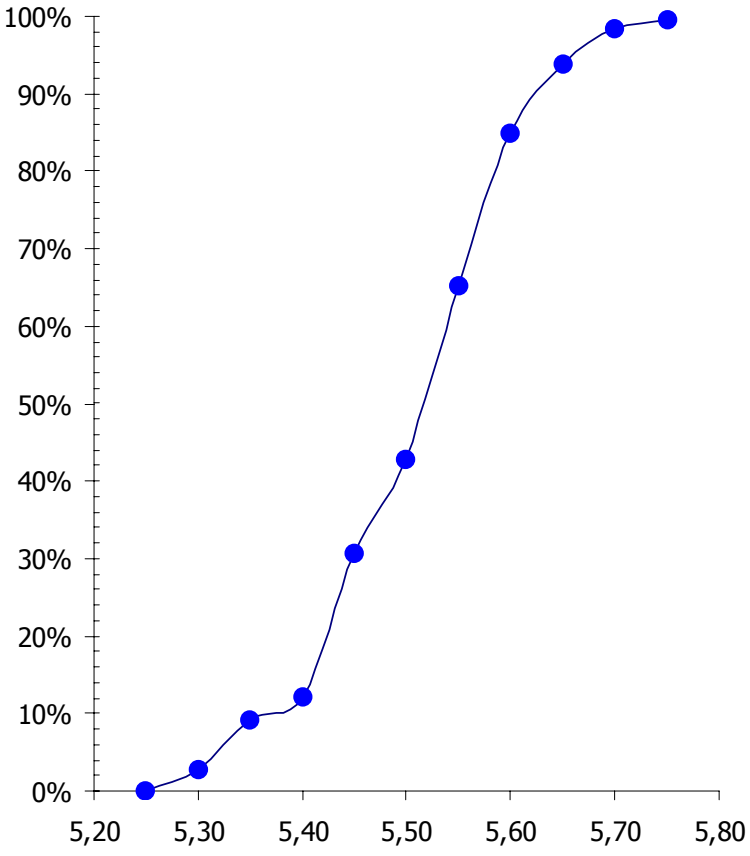
a) Construya un diagrama de caja. Para ello deberá crear la siguiente estructura de datos y representarla gráficamente.

Mínimo	2
Q1	2
Q1	3
Mediana	3
Mediana	1
Mediana	3
Q3	3
Q3	2
Máximo	2
Q3	2
Q3	1
Q1	1
Q1	2



b) Construya una tabla y un gráfico para, conociendo un valor concreto (comprendido entre 5,25 y 5,75), se pueda deducir qué porcentaje de bolsas tendrán un peso inferior o superior.

Valor	%
5,25	0,000
5,30	0,027
5,35	0,091
5,40	0,122
5,45	0,306
5,50	0,428
5,55	0,653
5,60	0,850
5,65	0,938
5,70	0,985
5,75	0,995



4.5.9 Utilizando la fórmula siguiente

DISTR.LOG.INV(ALEATORIO();μ;σ)

genere una muestra aleatoria de 500 valores de una distribución LogNormal de media μ = 10 y desviación σ = 3.

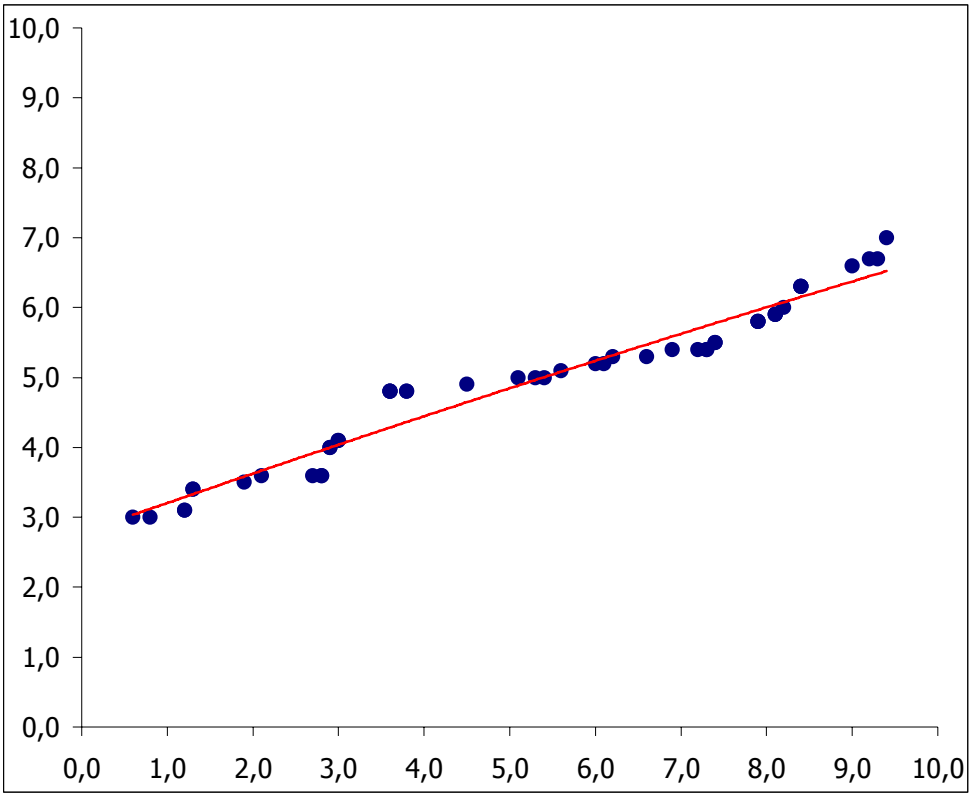
> *Se conviene en considerar valores **atípicos** aquellos que son mayores que Q_3 + 1,5·RI o menores que Q_1 - 1,5·RI; y en considerar valores **extremos** aquellos que son mayores que Q_3 + 3·RI o menores que Q_1 - 3·RI.*

Para los datos recién generados

a) representar el histograma;
b) obtener una aproximación a la función de densidad;
c) un diagrama de caja;
d) detectar si hay datos atípicos y/o extremos.

4.5.10 Para los datos anteriores compare gráficamente la mediana con la media recortada (MEDIA.ACOTADA) al α% $\alpha \in$(0;5;19;15;20}.

4.5.11 Se dispone de las notas en 2 asignaturas (A y B) de un mismo grupo de alumnos. Se quiere corregir las notas de B de manera que la nueva nota B' sea la que corresponda, por estar en la misma posición de orden, que la del grupo B (La nota más alta de B se convertirá en la que sea más alta de A, la segunda de B en la segunda de A ...). Suponer que A ≈ **U**[0;10] y que B ≈ **U**[3;5]. Hacer un gráfico de la transformación.



4.5.12 Repetir el problema anterior suponiendo que A y B están referidos a dos cursos con distinto número de alumnos.

4.5.13 La tabla siguiente muestra la distribución de frecuencias absolutas de una variable **X**. Con esta información calcular

a) la media de X;
b) su varianza.

X_i	n_i
1	35
2	5
3	21
4	32
5	47
6	24
7	12
8	32
9	7
10	30
11	33
12	23

4.5.14 Utilizar el módulo de Análisis de datos para generar una muestra aleatoria de una distribución binomial **B**(n=40;p=0,18).

a) Obtener la distribución de frecuencias absolutas.
b) Calcular media, varianza, desviación, coeficiente de asimetría (CAs) y coeficiente de apuntamiento (CAp) los datos no agrupados
c) Lo mismo utilizando los datos agrupados.

$$\bar{x} = \sum_{i=1}^{i=k} (c_i \cdot f_i) \quad ; \quad S_x = \sqrt{\sum_{i=1}^{i=k} (c_i - \bar{x})^2 \cdot f_i}$$

$$CA_s = \frac{\sum_{i=1}^{i=k} (c_i - \bar{x})^3 \cdot f_i}{S_x^3} \quad ; \quad CA_P = \frac{\sum_{i=1}^{i=k} (c_i - \bar{x})^4 \cdot f_i}{S_x^4}$$

4.5.15 Comprobar las siguientes propiedades de la media aritmética:

a) La suma de las desviaciones de los valores de la variable respecto de su media aritmética siempre es cero.

$$\sum (x_i - \bar{x}) = 0$$

b) La media de las desviaciones cuadráticas de los valores de la variable respecto de un constante k cualquiera es mínima cuando k es la media de x.

$$\min \left\{ \sum (x_i - k)^2 \right\} \Leftrightarrow \quad (k = \bar{x})$$

c) Se verifica que:

$$V = (ax + b) \quad \Rightarrow \quad \bar{V} = a\bar{x} + b$$

4.5.16 Comprobar que se verifica que

$$H_x \le G_x \le \bar{x}$$

siendo H_x y G_x las media armónica y geométrica respectivamente

4.5.17 Comprobar la desigualdad de Tchebychev.

Para cualquier conjunto de datos (de una población o una muestra) y cualquier constante k mayor que 1, el porcentaje de los datos que debe caer dentro de k-veces la desviación típica a cualquier lado de la media es, como mínimo:

$$P_r\left\{x \in (\mu \mp k\sigma)\right\} \geq \left(1 - \frac{1}{k^2}\right)$$

POBLACIÓN				
Media	12			
Desviación	1			
k	1,35			
MUESTRA				
Media	12,033			
Desviación	1,225			
Minimo	9,490	4	8,0%	
Inferior	10,379	41	82,0%	45,1%
Superior	13,687			
Maximo	15,229	5	10,0%	
		50	100,0%	

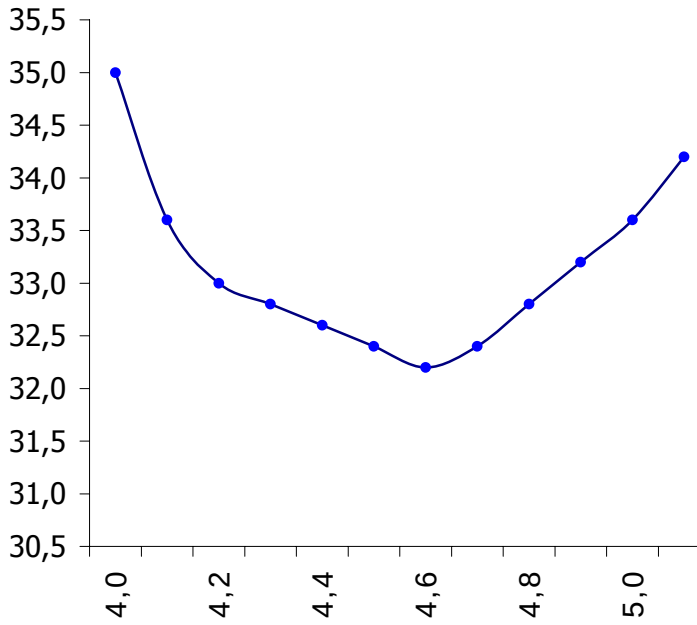
4.5.18 Sea el siguiente conjunto de datos:

6,6	3,7	5,9	4,0	3,6	3,1	3,2	6,1	3,7	5,2	5,8	5,0	5,7	4,1	4,2
3,1	4,7	4,2	4,1	4,1	6,5	7,0	6,0	6,9	4,6	4,1	6,6	4,6	3,0	6,4

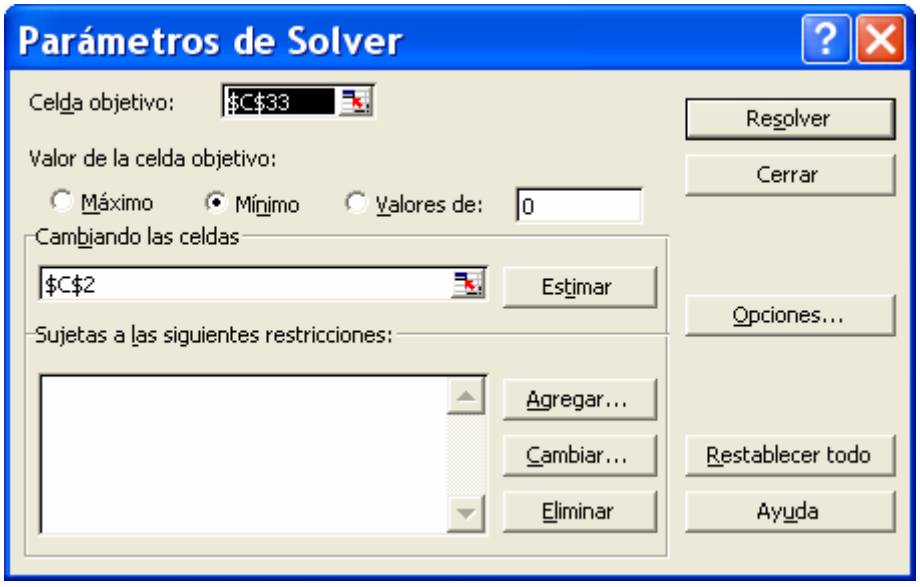
a) Calcular qué valor **k**, en torno a su media, hace mínima la expresión

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

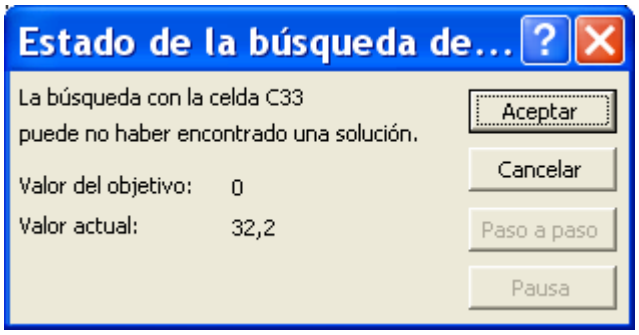
Datos	4,6	4,0	4,1	4,2	4,3	4,4	4,5	4,6	4,7	4,8	4,9	5,0	5,1
6,6	2,0	2,6	2,5	2,4	2,3	2,2	2,1	2,0	1,9	1,8	1,7	1,6	1,5
3,7	0,9	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1,0	1,1	1,2	1,3	1,4
5,9	1,3	1,9	1,8	1,7	1,6	1,5	1,4	1,3	1,2	1,1	1,0	0,9	0,8
4,0	0,6	0,0	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1,0	1,1
3,6	1,0	0,4	0,5	0,6	0,7	0,8	0,9	1,0	1,1	1,2	1,3	1,4	1,5
3,1	1,5	0,9	1,0	1,1	1,2	1,3	1,4	1,5	1,6	1,7	1,8	1,9	2,0
3,2	1,4	0,8	0,9	1,0	1,1	1,2	1,3	1,4	1,5	1,6	1,7	1,8	1,9
6,1	1,5	2,1	2,0	1,9	1,8	1,7	1,6	1,5	1,4	1,3	1,2	1,1	1,0
3,7	0,9	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1,0	1,1	1,2	1,3	1,4
5,2	0,6	1,2	1,1	1,0	0,9	0,8	0,7	0,6	0,5	0,4	0,3	0,2	0,1
5,8	1,2	1,8	1,7	1,6	1,5	1,4	1,3	1,2	1,1	1,0	0,9	0,8	0,7
5,0	0,4	1,0	0,9	0,8	0,7	0,6	0,5	0,4	0,3	0,2	0,1	0,0	0,1
5,7	1,1	1,7	1,6	1,5	1,4	1,3	1,2	1,1	1,0	0,9	0,8	0,7	0,6
4,1	0,5	0,1	0,0	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1,0
4,2	0,4	0,2	0,1	0,0	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9
3,1	1,5	0,9	1,0	1,1	1,2	1,3	1,4	1,5	1,6	1,7	1,8	1,9	2,0
4,7	0,1	0,7	0,6	0,5	0,4	0,3	0,2	0,1	0,0	0,1	0,2	0,3	0,4
4,2	0,4	0,2	0,1	0,0	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9
4,1	0,5	0,1	0,0	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1,0
4,1	0,5	0,1	0,0	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1,0
6,5	1,9	2,5	2,4	2,3	2,2	2,1	2,0	1,9	1,8	1,7	1,6	1,5	1,4
7,0	2,4	3,0	2,9	2,8	2,7	2,6	2,5	2,4	2,3	2,2	2,1	2,0	1,9
6,0	1,4	2,0	1,9	1,8	1,7	1,6	1,5	1,4	1,3	1,2	1,1	1,0	0,9
6,9	2,3	2,9	2,8	2,7	2,6	2,5	2,4	2,3	2,2	2,1	2,0	1,9	1,8
4,6	0,0	0,6	0,5	0,4	0,3	0,2	0,1	0,0	0,1	0,2	0,3	0,4	0,5
4,1	0,5	0,1	0,0	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1,0
6,6	2,0	2,6	2,5	2,4	2,3	2,2	2,1	2,0	1,9	1,8	1,7	1,6	1,5
4,6	0,0	0,6	0,5	0,4	0,3	0,2	0,1	0,0	0,1	0,2	0,3	0,4	0,5
3,0	1,6	1,0	1,1	1,2	1,3	1,4	1,5	1,6	1,7	1,8	1,9	2,0	2,1
6,4	1,8	2,4	2,3	2,2	2,1	2,0	1,9	1,8	1,7	1,6	1,5	1,4	1,3
	32,2	**35,0**	**33,6**	**33,0**	**32,8**	**32,6**	**32,4**	**32,2**	**32,4**	**32,8**	**33,2**	**33,6**	**34,2**

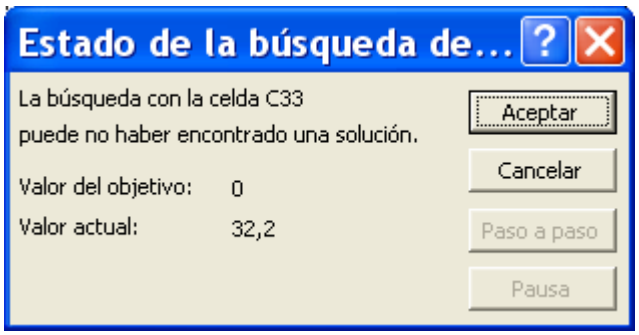


b) Hallar **k** utilizando **SOLVER**



c) Hallar **k** utilizando **BUSCAR OBJETIVO**





d) Comparar k con la mediana de los datos.

4.5.19 Comprobar que:

$$V = (ax + b) \quad \Rightarrow \quad \sigma_V = |a|\,\sigma_x$$

5 Medidas de asociación lineal

5.1 Procedimientos relacionados

Excel cuenta con dos procedimientos para obtener medidas de la relación lineal entre variables:

a) Utilizar las funciones relacionadas expuestas a continuación.
b) Recurrir al complemento de Análisis de Datos en donde encontraremos varios procedimientos asociados:
   - **Covarianza** (descrito en el punto 13.4)
   - **Regresión** (descrito en el punto 13.13)

5.2 Funciones para el cálculo del grado de asociación lineal.

Covarianza.

- **COVAR** Devuelve la **covarianza**, o promedio de los productos de las desviaciones para cada pareja de puntos de datos.

COVAR(matriz1;matriz2)

- **Matriz1** es el primer rango de celdas de números enteros.
- **Matriz2** es el segundo rango de celdas de números enteros.

- Los argumentos deben ser números o nombres, matrices o referencias que contengan números.
- Si el argumento matricial o de referencia contiene texto, valores lógicos o celdas vacías, estos valores se pasan por alto; sin embargo, se incluirán las celdas con el valor cero.
- Si los argumentos matriz1 y matriz2 tienen números distintos de puntos de datos, COVAR devuelve el valor de error #N/A.
- Si los argumentos matriz1 o matriz2 están vacíos, COVAR devuelve el valor de error #¡DIV/0! .

Coeficiente de correlación.

- **COEF.DE.CORREL** Devuelve el coeficiente de correlación entre dos rangos de celdas definidos por los argumentos matriz1 y matriz2. Use el coeficiente de correlación para determinar la relación entre dos propiedades. Por ejemplo, para examinar la relación entre la temperatura promedio de una localidad y el uso de aire acondicionado.

COEF.DE.CORREL(matriz1;matriz2)

- **Matriz1** es un rango de celdas de valores.
- **Matriz2** es un segundo rango de celdas de valores.

- Los argumentos deben ser números, o bien nombres, matrices o referencias que contienen números.
- Si el argumento matricial o de referencia contiene texto, valores lógicos o celdas vacías, estos valores se pasan por alto; sin embargo, se incluirán las celdas con el valor cero.
- Si los argumentos matriz1 y matriz2 tienen un número diferente de puntos de datos, COEF.DE.CORREL devuelve el valor de error #N/A.
- Si el argumento matriz1 o matriz2 está vacío, o si s (la desviación estándar de los valores) es igual a cero, COEF.DE.CORREL devuelve el valor de error #¡DIV/0!

5.3 PROBLEMAS

5.3.1 Para el siguiente conjunto de datos:

X	48	11	17	49	8	25	37	14	39	12	21	33	45	29	42
Y	55	19	22	61	8	38	40	24	49	18	33	36	46	30	44

a) Calcular la covarianza y el coeficiente de correlación de los datos.
b) Comprobar el resultado de las funciones con cálculos "a mano".
c) Realizar el correspondiente diagrama de dispersión.

	X	Y	$\frac{1}{n}\sum^{n}(x_i-\bar{x})(y_i-\bar{y})$	$(x_i-\bar{x})^2$	$(y_i-\bar{y})^2$
1	48	55	389,244	373,78	405,35
2	11	19	280,311	312,11	251,75
3	17	22	150,111	136,11	165,55
4	49	61	531,378	413,44	682,95
5	8	8	555,244	427,11	721,82
6	25	38	-11,489	13,44	9,82
7	37	40	42,778	69,44	26,35
8	14	24	159,378	215,11	118,08
9	39	49	146,044	106,78	199,75
10	12	18	281,111	277,78	284,48
11	21	33	14,311	58,78	3,48
12	33	36	4,911	18,78	1,28
13	45	46	181,844	266,78	123,95
14	29	30	-1,622	0,11	23,68
15	42	44	121,778	177,78	83,42
	28,67	34,87	2845,333	13,826	14,380

Covarianza	189,689	COVAR(B4:B18;C4:C18)
Covarianza	189,689	D19/15
Correlación	0,9541	COEF.DE.CORREL(B4:B18;C4:C18)
Correlación	0,9541	F22/(DESVESTP(C4:C18)*DESVESTP(B4:B18))
Correlación	0,9541	E22/(E19*F19)

5.3.2 El coeficiente de correlación por rangos de Spearman está definido de la forma siguiente:

$$\rho = 1 - \frac{6\sum_{i=1}^{i=N} d_i^2}{N^3 - N}$$

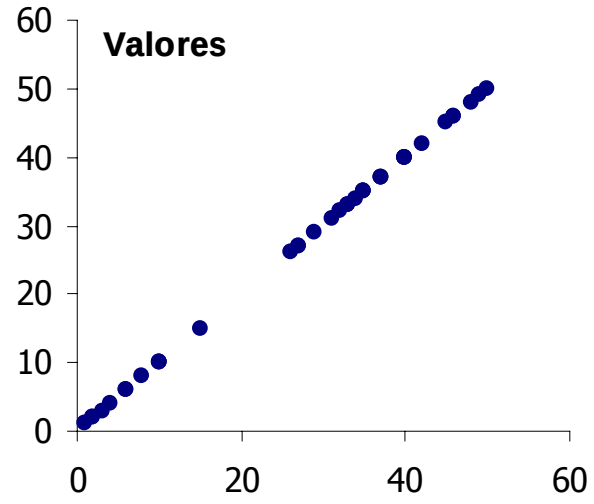
siendo $d_i = x_i - y_i$, con x_i ; y_i los rangos de las observaciones en ambas variables. Generar dos variables aleatorias, **X** e **Y,** de la forma siguiente:

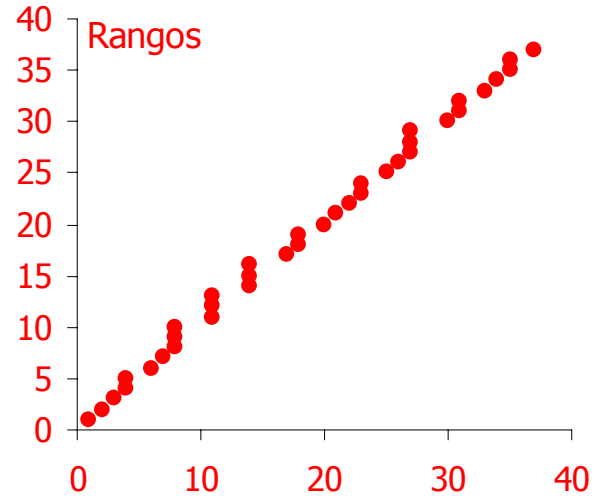
$$\begin{cases} \mathbf{X} \approx \mathbf{U}_{[0;50]} \\ \mathbf{Y} \Rightarrow y_i = \lambda x_i + (1-\lambda)\cdot \mathbf{U}_{[0;1]} \end{cases}$$

comparar los valores de ρ con los de r^2 para $0 \le \lambda \le 1$

X	Y			
33	33	18	19	1
10	10	27	27	0
32	32	20	20	0
35	35	14	15	1
37	37	11	11	0
8	8	30	30	0
35	35	14	14	0
2	2	35	36	1
6	6	31	31	0
10	10	27	29	4
35	35	14	16	4
37	37	11	13	4
4	4	33	33	0
45	45	6	6	0
40	40	8	8	0
42	42	7	7	0
26	26	25	25	0
46	46	4	4	0
40	40	8	9	1
3	3	34	34	0
6	6	31	32	1
15	15	26	26	0
48	48	3	3	0
27	27	23	23	0
37	37	11	12	1
46	46	4	5	1
1	1	37	37	0
50	50	1	1	0
33	33	18	18	0
49	49	2	2	0
31	31	21	21	0
29	29	22	22	0
40	40	8	10	4
10	10	27	28	1
2	2	35	35	0
34	34	17	17	0
27	27	23	24	1

Peso 0,999
Correlación 1,000 Spearman 0,997



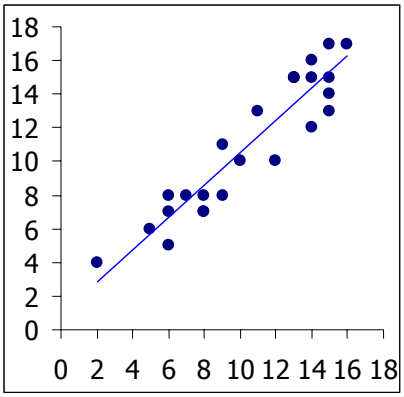


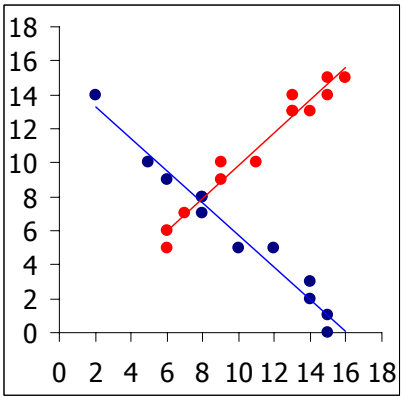
5.3.3 Suponga que tiene 3 variables A, B y C cuyos coeficientes de correlación con otra variable X son los dados en la tabla siguiente

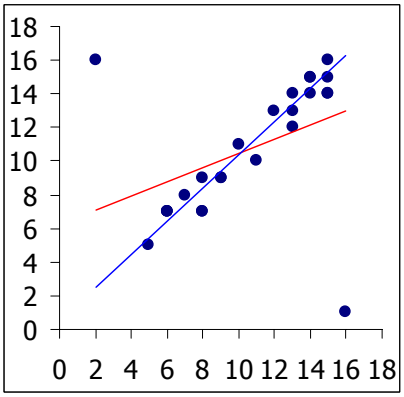
	X
A	**0,932**
B	**0,007**
C	**0,415**

a) Interprete el grado de relación de A, B y C con X antes y después de realizar el diagrama de dispersión

X	A	B1	B2	B	C
15	**13**	0		**0**	**15**
15	**17**	1		**1**	**14**
12	**10**	5		**5**	**13**
14	**16**	3		**3**	**15**
8	**7**	8		**8**	**7**
14	**12**	2		**2**	**14**
8	**7**	8		**8**	**9**
8	**8**	7		**7**	**7**
5	**6**	10		**10**	**5**
10	**10**	5		**5**	**11**
6	**5**	9		**9**	**7**
2	**4**	14		**14**	**16**
6	**8**		6	**6**	**7**
14	**15**		13	**13**	**15**
9	**11**		9	**9**	**9**
13	**15**		14	**14**	**13**
15	**14**		15	**15**	**14**
15	**15**		14	**14**	**16**
16	**17**		15	**15**	**1**
9	**8**		10	**10**	**9**
7	**8**		7	**7**	**8**
11	**13**		10	**10**	**10**
13	**15**		13	**13**	**12**
6	**7**		5	**5**	**7**
13	**15**		13	**13**	**14**







	X
A	0,932
B	0,007
C	0,415

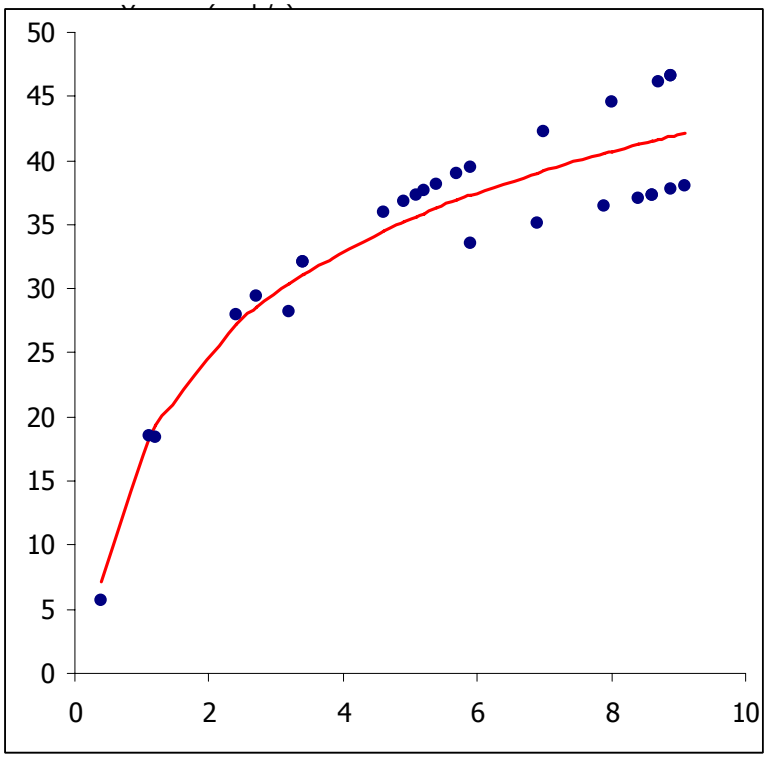
A X+ALEATORIO.ENTRE(-3;3)
B1 (16-X)+ALEATORIO.ENTRE(-1;1)
B2 X+ALEATORIO.ENTRE(-3;3)
C X+ALEATORIO.ENTRE(-1;1)

5.3.4 Sobre los datos de la hoja 3 ajustar utilizando **SOLVER** los siguientes modelos:

a. $\hat{y}_i = e^{\left(a-\frac{b}{x_i}\right)} + \varepsilon_i$

b. $\hat{y}_i = a \cdot Ln(x_i) + b + \varepsilon_i$

x	y		
a	3,441	11,214	
b	0,711	17,375	
0,4	5,676	7,10007	2,02656
1,1	18,552	18,44414	0,01173
1,2	18,457	19,41989	0,92645
2,4	28,008	27,19282	0,66398
2,7	29,384	28,51364	0,75803
3,2	28,192	30,41888	4,96032
3,4	32,120	31,09873	1,04398
3,4	32,120	31,09873	1,04398
4,6	35,940	34,48850	2,10720
4,9	36,794	35,19698	2,55160
5,1	37,348	35,64560	2,89951
5,2	37,621	35,86336	3,09029
5,4	38,159	36,28658	3,50776
5,7	38,950	36,89288	4,23061
5,9	39,466	37,27961	4,78232
5,9	33,566	37,27961	13,78747
6,9	35,054	39,03537	15,85220
7,0	42,195	39,19673	8,99182
7,9	36,424	40,55309	17,05338
8,0	44,556	40,69415	14,91098
8,4	37,077	41,24128	17,34350
8,6	37,333	41,50515	17,40492
8,6	37,333	41,50515	17,40492
8,7	46,161	41,63479	20,48245
8,9	46,613	41,88967	22,31378
8,9	37,713	41,88967	17,44110
8,9	46,613	41,88967	22,31378
9,1	37,964	42,13888	17,42921
			257,334



5.3.5 Sobre los datos de la hoja 2 hacer lo siguiente:

- Diagrama de dispersión + Tendencia lineal + Ecuación.
- Utilizar las funciones **INTERSECCION.EJE** y **PENDIENTE** para calcular la recta según un modelo lineal.
- Calcular, con los valores anteriores la predicción para X = {10,..,15}, comparar los resultados con los de la función **TENDENCIA** y los de la función **PRONOSTICO**.
- Calcular el coeficiente de correlación comparar con el resultados de la función **PEARSON**.

6 Variables aleatorias discretas.

6.1 Binomial.

Supongamos que un experimento aleatorio tiene las siguientes características:

a) En cada prueba del experimento sólo son posibles dos resultados: el suceso A (éxito) y su contrario $\bar{A}$ (fracaso).

b) El resultado obtenido en cada prueba es independiente de los resultados obtenidos anteriormente.

c) La probabilidad del suceso A es constante, la representamos por p, y no varía de una prueba a otra. La probabilidad de es $\bar{A}$ 1- p y la representamos por q.

d) El experimento consta de un número n de pruebas.

Todo experimento que tenga estas características diremos que sigue el modelo de la distribución Binomial. A la variable X que expresa el número de éxitos obtenidos en cada prueba del experimento, la llamaremos variable aleatoria binomial. La variable binomial es una variable aleatoria discreta, sólo puede tomar los valores 0, 1, 2, 3, 4, ..., n suponiendo que se han realizado n pruebas. Como hay que considerar todas las maneras posibles de obtener k-éxitos y (n-k) fracasos debemos calcular éstas por combinaciones.

Una v.a. Binomial representa el número de éxitos que ocurren en n repeticiones independientes de un ensayo de Bernouilli cuya probabilidad de éxito es p. Así de distribuyen con arreglo a esta distribución, el número de piezas defectuosas en un lote de tamaño n (moderado) cuando cada pieza tiene una probabilidad p de ser defectuosa; el tamaño de un conjunto si éste es aleatorio y no demasiado grande; el número de artículos demandados de un almacén, el número de encuestados que están a favor de determinada cuestión, etc.

La notación habitual es X~B(n,p).

La función de densidad es:

$$p(x) = \binom{n}{X} p^X (1-p)^{1-X}$$

La función de distribución es:

$$F(x) = \sum_{i=0}^{X} \binom{n}{i} p^X (1-p)^{1-X}$$

La media y varianza son (respectivamente):

$$np \quad ; \quad np(1-p)$$

Propiedades.

Si $(X_1,X_2,..X_m) \sim B(n_i,p)$ entonces $(X_1+X_2+..+X_m) \sim B(n_1+n_2+..n_m,p)$; si $X \sim B(n,p)$ entonces la variable $(n-X) \sim B(n,1-p)$. La distribución es simétrica sólo si p=1/2

Generación.

Puesto que Excel cuenta con una función para la inversa de la función de distribución, la generación de variables aleatorias puede hacerse, bien a través del módulo de Análisis de datos, bien directamente por inversión utilizando la fórmula siguiente:

=BINOM.CRIT(n;p;ALEATORIO())

6.2 Poisson

Una v.a. de Poisson es en realidad una v.a. Binomial llevada al límite, es decir cuando $n\to\infty$ (aunque basta con que sea suficientemente grande) y $p\to 0$ (aunque basta con que sea muy pequeño).

En general un suceso "raro" puede ser perfectamente modelizado por un v.a. de Poisson, ejemplos típicos son el número de remaches defectuosos en un avión (porque un avión puede llegar a tener varios millones de ellos y al ser un mecanismo tan simple es realmente difícil que sea defectuoso); el número de erratas en un libro (que contiene un gran número de palabras que difícilmente están mal escritas) el número de llegadas a un servicio si la distribución entre los tiempos es exponencial, el número de accidentes laborales en un mes en una gran empresa, el número de personas que entran en un supermercado en un minuto, etc.

La notación habitual es X~Poisson(λ). El único parámetro debe ser positivo $\lambda>0$.

La función de densidad es:

$$p(x) = \frac{e^{-\lambda}\lambda^{X}}{x!}$$

La función de distribución es:

$$F(x) = e^{-\lambda}\sum_{i=0}^{i=X}\frac{\lambda i}{i!}$$

La media y varianza coinciden en el único parámetro λ.

Propiedades.

Si $(X_1,X_2,..X_m)$ ~Poisson(λ_i) entonces $(X_1+X_2+..+X_m)$~Poisson($\lambda_1+\lambda_2+..\lambda_m$); si X~B(n,p) entonces la variable (n-X)~B(n,1-p).

Generación.

Excel cuenta con una función para la distribución y densidad de Poisson, cuenta también con la posibilidad de obtener muestras aleatorias así distribuidas (Herramientas + Análisis de Datos + Generación de números aleatorios). En cualquier caso es posible obtener números que se distribuyan según una Poisson aleatorios utilizando la fórmula siguiente:

BINOM.CRIT(λ/0,001;0,001;ALEATORIO())

Caracterización.

El parámetro pueden ser estimado fácilmente de la forma siguiente:

$$\hat{\lambda} = \overline{\mathbf{x}}_{(\mathbf{n})}$$

6.3 Uniforme (Discreta)

Esta v.a. es el equivalente discreto de la de mismo nombre dentro de las distribuciones continuas. Se utiliza cuando un conjunto de posibles resultados es igualmente probable, la cantidad de caras con un determinado número al lanzar un dado, la probabilidad de cada número en un sorteo legal, etc.

La notación habitual es X~UD(a,b). El único parámetro debe ser positivo □>0.

La función de densidad es:

$$\mathbf{p(x)} = \frac{1}{\mathbf{a}-\mathbf{b}+1}$$

La función de distribución es:

$$\mathbf{F(x)} = \frac{\mathbf{X - a + 1}}{\mathbf{a - b + 1}}$$

La media y varianza son:

$$\frac{\mathbf{a + b}}{2} \quad ; \quad \frac{(\mathbf{a - b + 1})^2 - 1}{12}$$

Excel cuenta con una función directa para generar muestras aleatorias así distribuidas

ALEATORIO.ENTRE(a;b)

Caracterización.

Los parámetros pueden ser estimados fácilmente de la forma siguiente:

$$\hat{\mathbf{a}} =, \min\left\{\mathbf{X_{(n)}}\right\} \quad ; \quad \hat{\mathbf{b}} =, \max\left\{\mathbf{X_{(n)}}\right\}$$

6.4 Geométrica

Una v.a. Geométrica representa el número de fracasos que ocurren hasta obtener el primer éxito en la realización de ensayos de Bernouilli con probabilidad p de éxito. Así, el número de artículos examinados de un lote hasta que aparece el primer defectuoso, el número de candidatos a entrevistar cuando se quiere encontrar una persona idónea para un puesto de trabajo, el número de melones que un cliente exigente manosea antes de conseguir aquél que satisface sus criterios, etc.

La notación habitual es X~Geom(p) o, a veces, G(p).

La función de densidad es:

$$\mathbf{p(x)} = \mathbf{p}\,(1 - \mathbf{p})^{\mathbf{X}}$$

La función de distribución es:

$$\mathbf{F(x)} = 1 - (1 - \mathbf{p})^{\mathbf{x}+1}$$

La media y varianza son respectivamente.

$$\frac{(1 - \mathbf{p})}{\mathbf{p}} \quad ; \quad \frac{(1 - \mathbf{p})}{\mathbf{p}^2}$$

Propiedades.

La primera propiedad es evidente: se trata de una particularización de la binomial negativa, es decir, se verifica que BN(1,p) ≡ Geom(p). Si $(X_1,X_2,..X_m)$~G(p) entonces $(X_1+X_2+..+X_m)$~BN(m,p).

Es el equivalente discreto de la Exponencial en el sentido de que es la única distribución discreta que "no guarda memoria" ya que el número de fallos ocurridos hasta un instante dado no modifica la probabilidad de que el próximo intento sea un éxito.

Generación.

Excel no cuenta con una función para la distribución y densidad de la distribución Geométrica, sin embargo es fácil generar muestras aleatorias por inversión de la función de Distribución utilizando la fórmula siguiente

REDONDEAR.MENOS(LN(ALEATORIO())/LN(1-p);0)

Caracterización.

Se verifica que:

$$\hat{p} = \frac{1}{\overline{X}_{(n)} + 1}$$

6.5 Binomial Negativa

Una v.a. Binomial negativa representa el número de fracasos que ocurren hasta obtener el n-ésimo éxito en la realización de ensayos de Bernouilli con probabilidad p de éxito. Así, el número de artículos examinados de un lote hasta que aparece el n-ésimo defectuoso, el número de candidatos a entrevistar cuando se quiere formar un equipo de n personas idóneas para un puesto de trabajo, etc.

La notación habitual es X~NegBin(n,p) o, a veces, BN(n,p).

La función de densidad es:

$$p_{(x)} = \binom{n+X-1}{X} p^X (1-p)^X$$

La función de distribución es:

$$F_{(x)} = \sum_{i=0}^{i=X} \binom{n+i-1}{i} p^n (1-p)^i$$

La media y varianza son respectivamente.

$$\frac{n(1-p)}{p} \quad ; \quad \frac{n(1-p)}{p^2}$$

Propiedades.

Si $(X_1,X_2,..X_m)\sim BN(n_i)$ entonces $(X_1+X_2+..+X_m)\sim BN(n_1+ n_2+.. n_m)$. También es conocida como distribución de Pascal o distribución de Polya. Se verifica que $BN(1,p) \equiv Geom(p)$.

Generación.

Excel cuenta con una función para la distribución y densidad de la Binomial Negativa aunque no con la inversa de la distribución. No cuenta tampoco con la posibilidad de obtener muestras aleatorias a partir del módulo de Análisis de Datos + Generación de números aleatorios.

En cualquier caso es posible obtener números que se distribuyan según una esta distribución utilizando la fórmula siguiente:

BINOM.CRIT(DISTR.GAMMA.INV(U;n;(1-p)/p)/ε;ε;U)

siendo ε un número suficientemente pequeño (obtendremos buenos resultados con ε= 0,0001) y U la Uniforme (0;1), es decir U = ALEATORIO().

6.6 Distribución Hipergeométrica

Una v.a. Hipergeométrica representa el número de éxitos que ocurrirán cuando de una población en la que hay N éxitos y M fracasos se extrae una muestra, sin repetición, de tamaño n. Es importante notar que el muestreo se hace sin repetición, es decir sin devolver los objetos al seno de la población antes de cada ensayo, porque esta característica es la única que diferencia esta distribución de la distribución binomial.

Se distribuyen según una Hipergeométrica magnitudes tales como el numero de hombres (o de mujeres) que incluye una selección al azar de un grupo en el que ambos géneros están presentes, el numero de temas estudiados por un opositor que ha decidido estudiar sólo unos cuantos del temario de su oposición cuando el examen consta de varios temas, etc.

La notación habitual es X~HiperGeom(n,N,M) o también X~H(n,N,M). Todos los parámetros deben ser lógicamente positivos y representan n el tamaño de la muestra

extraída; N número de éxitos que contiene la población, M el número total de elementos de la población.

La función de densidad es:

$$p_{(x)} = \frac{\binom{M}{X}\binom{N-M}{n-X}}{\binom{N}{n}}$$

La función de distribución es:

$$F_{(x)} = \frac{1}{\binom{N}{n}} \sum_{i=0}^{i=X} \binom{M}{i}\binom{N-M}{n-i}$$

La media y varianza son:

$$\frac{nM}{N} \quad ; \quad \left(\frac{N-n}{N-1}\right)\left(\frac{nM}{N}\right)\left(1-\frac{M}{N}\right)$$

Propiedades.

Es evidente que ha de verificarse que: $\mathbf{Max}(0, \mathbf{n} - \mathbf{N} + \mathbf{M}) \leq \mathbf{X} \leq \mathbf{Min}(\mathbf{M}, \mathbf{n})$

Generación.

Excel cuenta con una función para la distribución y densidad, no cuenta sin embargo, con la posibilidad de obtener muestras aleatorias

6.7 Funciones Excel relacionadas

BINOMIAL

Recordamos que la función de cuantía de la distribución B(n,p) es:

$$p(x) = \binom{n}{X} p^X (1-p)^{1-X}$$

mientras que la función de distribución es

$$F(x) = \sum_{i=0}^{X} \binom{n}{i} p^X (1-p)^{1-X}$$

La función de Excel que nos da ambas es:

DISTR.BINOM(k ; n ; p ; acumulado)

- **k** es el valor sobre el que hallaremos la probabilidad;
- **n** y **p** los parámetros que definen la distribución;
- **acumulado** es un valor lógico que determina la forma de la función. Si el argumento acumulado es VERDADERO, DISTR.BINOM devuelve la función de distribución; si es FALSO, devuelve la función de masa de probabilidad.

Una segunda función de Excel relacionada con la binomial es:

BINOM.CRIT(n ; p ; alfa)

Función que devuelve el menor valor cuya distribución binomial acumulativa es menor o igual que un valor (alfa) de criterio.

- **n** y **p** los parámetros que definen la distribución;
- **alfa** el criterio (0<alfa<1).

BINOMIAL NEGATIVA

La función de cuantía es de NegBin(n,p) es:

$$p_{(x)} = \binom{n+X-1}{X} p^X (1-p)^X$$

La función de Excel para la cuantía es:

NEGBINOMDIST(núm_fracasos;núm_éxitos;prob_éxito)

- Núm_fracasos: es el número de fracasos.
- Núm_éxitos: es el número límite de éxitos.
- Prob_éxito: es la probabilidad de obtener un éxito.

Observaciones

- Los argumentos núm_fracasos y núm_éxitos se truncan a enteros.
- Si uno de los argumentos no es numérico, NEGBINOMDIST devuelve el valor de error #¡VALOR!
- Si el argumento prob_éxito < 0 o si probabilidad > 1, NEGBINOMDIST devuelve el valor de error #¡NUM!.
- Si los argumentos (núm_fracasos + núm_éxitos - 1) ≤ 0, la función NEGBINOMDIST devuelve el valor de error #¡NUM!.

POISSON

Recordamos que la función de cuantía de la distribución Poisson(λ) es:

$$p_{(x)} = \frac{e^{-\lambda}\lambda^X}{x!}$$

La función de distribución es:

$$F_{(x)} = e^{-\lambda}\sum_{i=0}^{i=X}\frac{\lambda i}{i!}$$

La función de Excel que nos da ambas es:

POISSON(x ; media ; acumulado)

- **x** el valor que toma la variable;
- **media**, el parámetro λ;
- **acumulado** es un valor lógico que determina la forma de la función. Si el argumento acumulado es VERDADERO, devuelve la función de distribución; si es FALSO, devuelve la función de masa de probabilidad.

HIPERGEOMÉTRICA

La función de cuantía de la HiperGeom(n,N,M) es:

$$p_{(x)} = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$$

La función de Excel a utilizar es:

DISTR.HIPERGEOM(x; n; M; N)

- **x** es el número de éxitos en la muestra.
- **n** es el tamaño de la muestra.
- **M** es el número de éxitos en la población.
- **N** es el tamaño de la población.

Observaciones

- Todos los argumentos se truncan a enteros.
- Si uno de los argumentos no es numérico, DISTR.HIPERGEOM devuelve el valor de error #¡VALOR!
- Si el argumento x < 0 o si x es mayor que el menor de los números entre el argumento n o N, DISTR.HIPERGEOM devuelve el valor de error #¡NUM!
- Si el argumento x es menor que el mayor número entre 0 o (n - N + M), DISTR.HIPERGEOM devuelve el valor de error #¡NUM!
- Si el argumento n, M, N < 0 o si n, M > N, DISTR.HIPERGEOM devuelve el valor de error #¡NUM!

6.8 PROBLEMAS

6.8.1 Representar la función de masa de las siguientes distribuciones:

a) **B**(n=10;p=0,4).
b) **Geom**(p=0,65).

6.8.2 Una máquina fabrica una determinada pieza y se sabe que produce un 7 por 1000 de piezas defectuosas.

c) Hallar la probabilidad de que al examinar 50 piezas sólo haya una defectuosa.
d) Generar una lista del nº de piezas defectuosas y su probabilidad asociada.

6.8.3 La probabilidad de éxito de una determinada vacuna es 0,72. Calcular la probabilidad de que, una vez administrada a 15 pacientes:

a) Ninguno sufra la enfermedad
b) Todos sufran la enfermedad
c) Dos de ellos contraigan la enfermedad

6.8.4 La probabilidad de que el carburador de un coche salga de fábrica defectuoso es del 4 por 100. Hallar :

a) El número de carburadores defectuosos esperados en un lote de mil
b) La varianza y la desviación típica.

6.8.5 Un profesor ha sometido a sus estudiantes a un examen de 18 preguntas, cada una de las cuales tenia cuatro posibles respuestas, de las que únicamente una era la correcta. ¿A partir de qué puntuación obtenida por los alumnos es razonable (95% de confianza) suponer que las respuestas no han sido escogidas al azar?.

6.8.6 Un fabricante vende bolsas de semillas de maíz de calidad extra que germinan en un 98% de los casos. Las vende en bolsas de 500 granos y garantiza la germinación de un 96% de las semillas como mínimo. ¿Cuál es la probabilidad de que no cumpla la garantía?.

6.8.7 Generar una muestra de 100 valores de una distribución de Poisson de parámetro arbitrario.

a) Estimar el parámetro de la distribución.
b) Representar la función de masa observada y esperada.

6.8.8 Generar una muestra (n = 100) de una $\mathbf{U}_{[1;7]}$

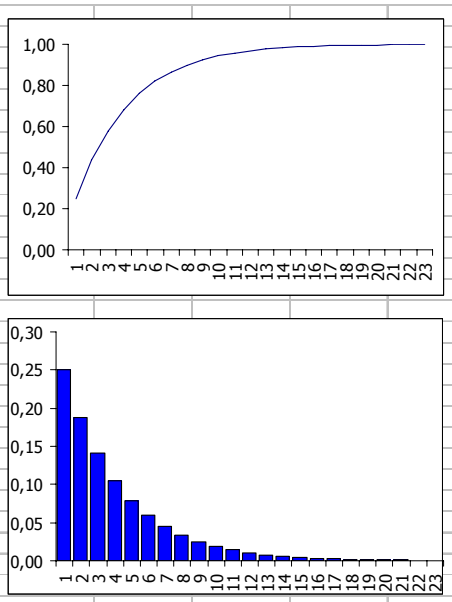
a) Estimar su media y varianza.
b) Construir la distribución de frecuencias.
c) Comparar las frecuencias esperadas con las observadas.
d) Comparar la media y varianza esperadas con las observadas.

6.8.9 Igual que el anterior con una **Geom**(p=0,25).

6.8.10 Suponga que la probabilidad de encontrar una bujía defectuosa es del 25%.

a) ¿Cuál es la probabilidad de que sea necesario examinar 12 bujías antes de encontrar una defectuosa?.
b) Genere una tabla para todas las posibilidades.
c) Realice un histograma del valor de la variable aleatoria y de su función de densidad acumulada.
d) ¿Cuál será el número máximo de bujías que será necesario examinar (95%)?.

	Éxito		
	0,250	**0,025**	**Acum**
1	0,2500	0,0250	0,2500
2	0,1875	0,0244	0,4375
3	0,1406	0,0238	0,5781
4	0,1055	0,0232	0,6836
5	0,0791	0,0226	0,7627
6	0,0593	0,0220	0,8220
7	0,0445	0,0215	0,8665
8	0,0334	0,0209	0,8999
9	0,0250	0,0204	0,9249
10	0,0188	0,0199	0,9437
11	0,0141	0,0194	0,9578
12	0,0106	0,0189	0,9683
13	0,0079	0,0184	0,9762
14	0,0059	0,0180	0,9822
15	0,0045	0,0175	0,9866
16	0,0033	0,0171	0,9900
17	0,0025	0,0167	0,9925
18	0,0019	0,0163	0,9944
19	0,0014	0,0158	0,9958
20	0,0011	0,0155	0,9968
21	0,0008	0,0151	0,9976
22	0,0006	0,0147	0,9982
23	0,0004	0,0143	0,9987

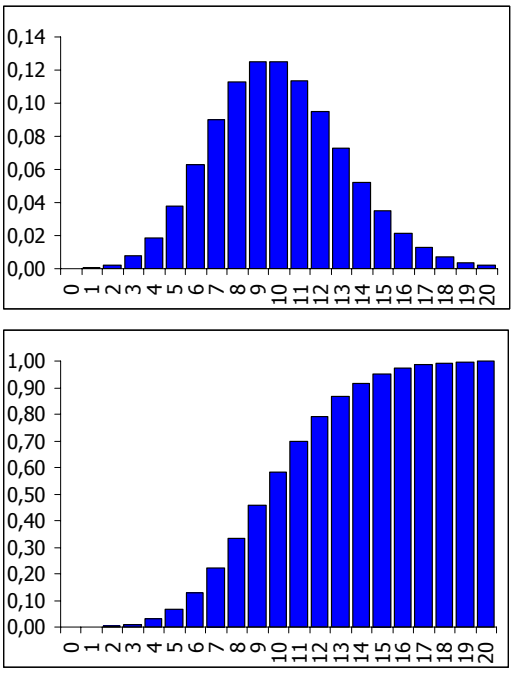


6.8.11 La variable X se distribuye con arreglo a una Poisson (λ=10).

a) Generar una tabla para los primeros valores de f(x) y F(x).
b) Gráficos de ambas funciones
c) Calcular la media de los 20 primeros valores

	10	
0	0,000045	0,000045
1	0,000454	0,000499
2	0,002270	0,002769
3	0,007567	0,010336
4	0,018917	0,029253
5	0,037833	0,067086
6	0,063055	0,130141
7	0,090079	0,220221
8	0,112599	0,332820
9	0,125110	0,457930
10	0,125110	0,583040
11	0,113736	0,696776
12	0,094780	0,791556
13	0,072908	0,864464
14	0,052077	0,916542
15	0,034718	0,951260
16	0,021699	0,972958
17	0,012764	0,985722
18	0,007091	0,992813
19	0,003732	0,996546
20	0,001866	0,998412

Media 9,96546

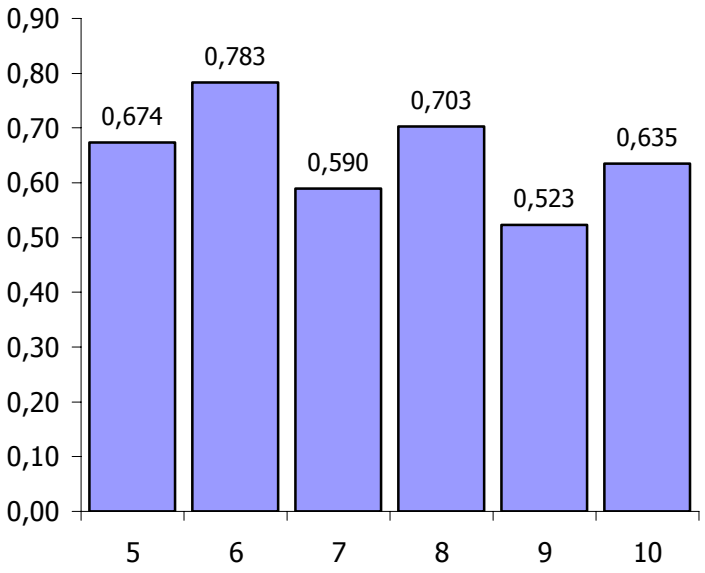


6.8.12 Supóngase que de un grupo de 50 diputados de una determinada cámara legislativa, 30 están a favor de una determinada modificación a cierta ley.

a) Se selecciona un grupo al azar de 5 diputados, ¿cuál es la probabilidad de que en dicho grupo haya mayoría a favor de la modificación de la ley?.

b) ¿Cuál será el tamaño óptimo del grupo de diputados para que la probabilidad de modificación sea máxima, sabiendo que el reglamento de la cámara lo restringe a "una cifra comprendida entre 5 y 10 diputados"?.

N	30					
M	50					
	0,67405	0,78328	0,58965	0,70312	0,52301	0,63503
n	5	6	7	8	9	10
0	0,00732	0,00244	0,00078	0,00023	0,00007	0,00002
1	0,06860	0,02927	0,01164	0,00433	0,00151	0,00049
2	0,23405	0,13263	0,06752	0,03140	0,01346	0,00533
3	0,36408	0,29126	0,19693	0,11724	0,06281	0,03064
4	0,25869	0,32767	0,31278	0,24731	0,16959	0,10341
5	0,06726	0,17936	0,27107	0,30260	0,27558	0,21509
6		0,03737	0,11889	0,21014	0,27017	0,28006
7			0,02038	0,07584	0,15439	0,22593
8				0,01090	0,04672	0,10826
9					0,00571	0,02786
10						0,00292

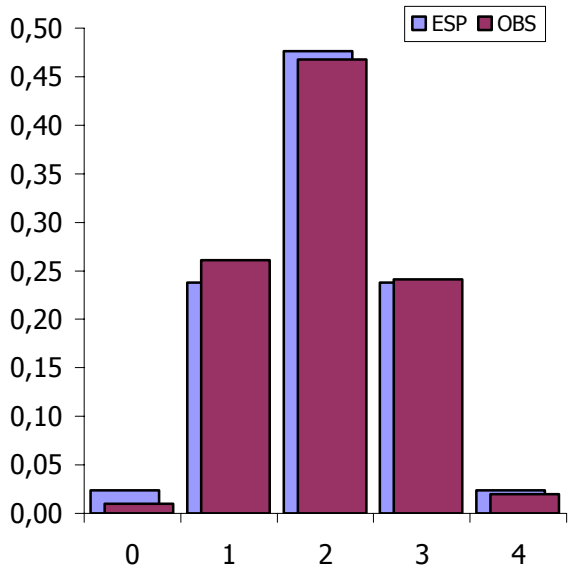


6.8.13 Simular el experimento aleatorio correspondiente a una hipergeométrica de parámetros n=5, M=10; N=4

a) Obtener las probabilidades de la v.a. observadas en la simulación y compararlas con las teóricas esperadas

NOTA Utilizar las funciones **JERARQUIA** e **INDICE** para la simulación

	M	10				OBS	ESP			
	N	4		0	2	0,01	0,02			
	n	5		1	53	0,26	0,24			
				2	95	0,47	0,48			
				3	49	0,24	0,24			
				4	4	0,02	0,02			
					203					
	1	**1**	**1**	**1**	**0**	**0**	**0**	**0**	**0**	**0**
	1	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***
4	1	1	1	1	0	0	0	0	0	0
2	1	0	0	0	1	0	1	1	0	0
2	1	1	0	0	0	1	0	0	0	1
3	1	1	0	0	1	0	0	0	1	0
2	1	1	0	0	0	0	1	1	0	0
2	1	0	0	1	0	1	0	1	0	0



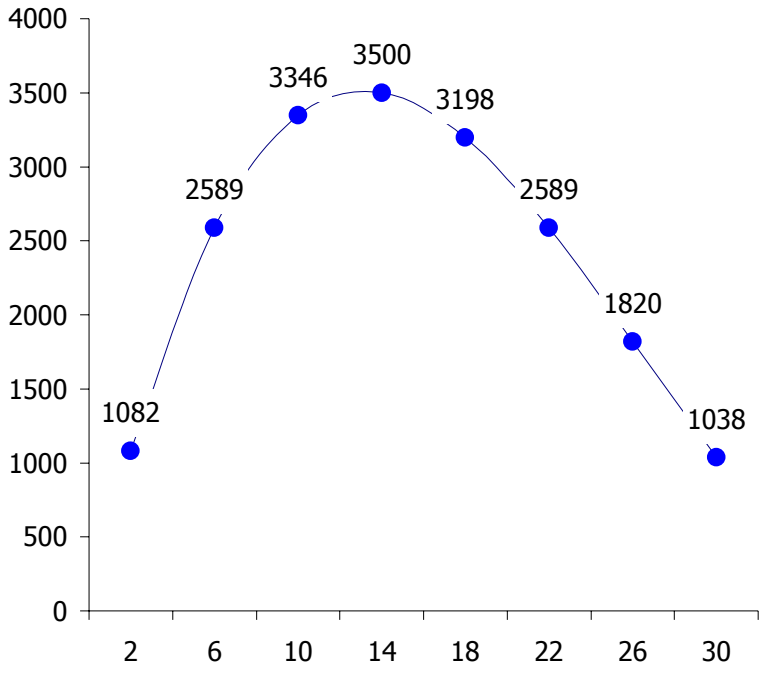
6.8.14 Considérese un empresario que compra motores a una compañía que los fabrica. El empresario recibe lote de 40 motores, su plan de aceptación de lote consiste en lo siguiente:

Seleccionar 8 motores del lote y someterlos a prueba. Si ninguno presenta defectos aceptar el lote, en caso contrario rechazarlo.

a) ¿Cuál es la probabilidad de aceptar un lote en el que 2 motores están defectuosos?.
b) Suponga que el empresario tiene una función de beneficio que es de la forma:

$$C_n = 600 \cdot n \cdot P$$

c) Siendo n el tamaño del lote que inspecciona y P la probabilidad de aceptar un lote que contiene 2 defectuosos. ¿Cuál es el tamaño óptimo (n*) del lote que debe inspeccionar?



	1082	2589	3346	3500	3198	2589	1820	1038
Pr	90	72	56	42	30	20	12	6
Ct	12	36	60	84	108	132	156	180
n	2	6	10	14	18	22	26	30
0	**0,9013**	**0,7192**	**0,5577**	**0,4167**	**0,2962**	**0,1962**	**0,1167**	**0,0577**
1	0,0974	0,2615	0,3846	0,4667	0,5077	0,5077	0,4667	0,3846
2	0,0013	0,0192	0,0577	0,1167	0,1962	0,2962	0,4167	0,5577
	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000

6.8.15 Supóngase que para personas de determinada edad, la probabilidad de que mueran por una enfermedad transmisibles es 0,001. ¿Cuántas personas de este grupo pueden exponerse a la enfermedad de manera que la probabilidad de que no más de una persona muera sea por lo menos del 95%?.

No se puede resolver de forma analítica ya que la ecuación

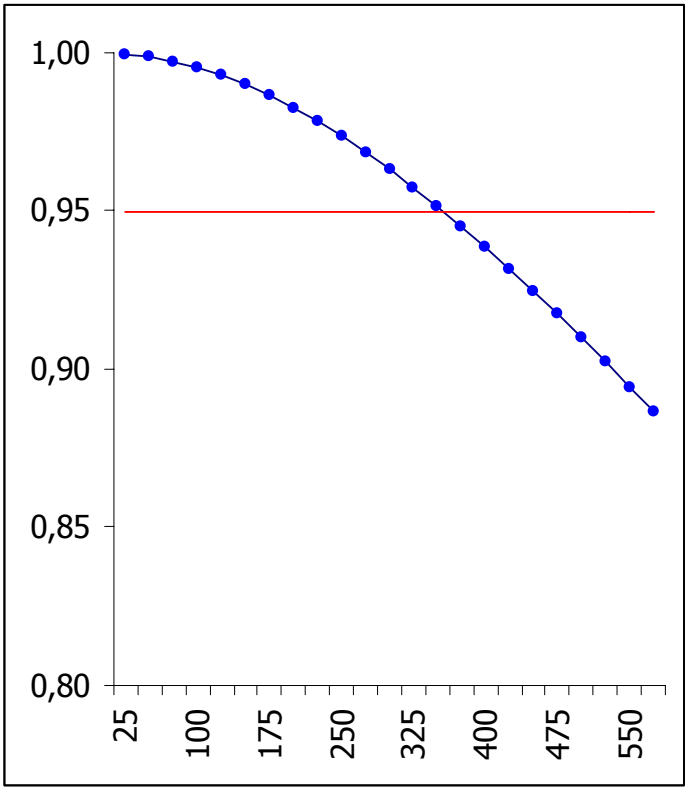
$$\binom{n}{0}(0.001^{0})(0.999^{n})+\binom{n}{1}(0.001^{0})(0.999^{n})=0.95$$

no se resuelve de manera explicita

Primer método: búsqueda bruta

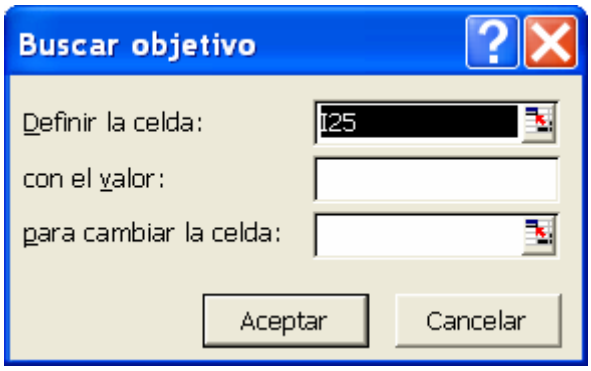
p 0,001

n	DISTR.BINOM(1;n;p;1)
25	1,000
50	0,999
75	0,997
100	0,995
125	0,993
150	0,990
175	0,986
200	0,983
225	0,978
250	0,974
275	0,969
300	0,963
325	0,957
350	0,951
375	0,945
400	0,939
425	0,932
450	0,925
475	0,917
500	0,910
525	0,902
550	0,894
575	0,886



Segundo método: utilizando "Buscar objetivo"

- Definir la celda: Introducir la celda que depende de otra y que se pretende que alcance el valor v
- con el valor: introducir el valor v
- para cambiar la celda: Introducir la celda que contiene el valor que se quiere encontrar

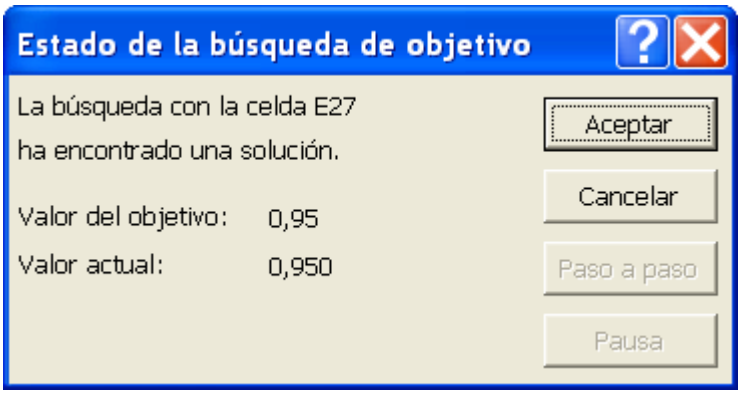


1	1,000	0,95

DISTR.BINOM(1;D27;D1;1)

356	0,950	0,95

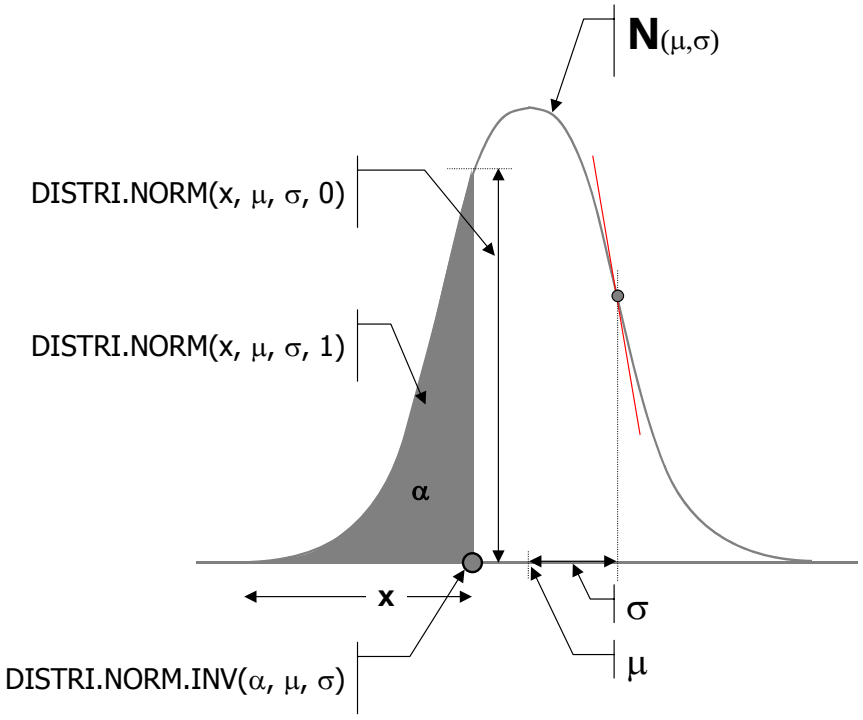
DISTR.BINOM(1;D27;D1;1)



7 Variables aleatorias continuas.

7.1 Funciones relacionadas con la Normal

Existen en total cuatro funciones relacionadas con la distribución normal, dos de ellas referidas a la normal estándar **N**(0;1) y las otras dos a una normal genérica **N**(π;σ). Cada una de ellas tiene además una función para la densidad y otra para la inversa.



1. **DISTR.NORM**: Devuelve la distribución normal acumulativa

   DISTR.NORM(**x** ; **media** ; **desv_estándar** ; **acum**)

   - **X**: es el valor cuya distribución desea obtener.
   - **Media**: es la media aritmética de la distribución.
   - **Desv_estándar**: es la desviación estándar de la distribución.
   - **Acum**: es un valor lógico que determina si la función devuelve la densidad (Acum = 0) o la función de Distribución (Acum = 1).
   - Si los argumentos media o desv_estándar no son numéricos, DISTR.NORM devuelve el valor de error #¡VALOR!
   - Si el argumento desv_estándar ≤ 0, la función DISTR.NORM devuelve el valor de error #¡NUM!

2. **DISTR.NORM.ESTAND:** Devuelve la distribución normal acumulativa estándar.

   DISTR.NORM.ESTAND(z)

   - **Z:** es el valor para el cual desea obtener la distribución.

3. **DISTR.NORM.ESTAND.INV** Devuelve el inverso de la distribución normal acumulativa estándar.

DISTR.NORM.ESTAND.INV(**probabilidad**)

- **Probabilidad**: es una probabilidad correspondiente a la distribución normal.
- Si el argumento probabilidad no es numérico, DISTR.NORM.ESTAND.INV devuelve el valor de error #¡VALOR!
- Si probabilidad < 0 o si probabilidad > 1, DISTR.NORM.ESTAND devuelve el valor de error #¡NUM!
- La función DISTR.NORM.ESTAND.INV se calcula utilizando una técnica iterativa. Dado un valor de probabilidad, DISTR.NORM.ESTAND.INV itera hasta que el resultado tenga una exactitud de ± 3x10^-7. Si no converge después de 100 iteraciones, la función devuelve el valor de error #N/A.

4. **DISTR.NORM.INV:** Devuelve el inverso de la distribución normal acumulativa

DISTR.NORM.INV(**probabilidad** ; **media** ; **desv_estándar**)

- Probabilidad: es l probabilidad correspondiente a la distribución normal.
- Media: es la media aritmética de la distribución.
- Desv_estándar: es la desviación estándar de la distribución.
- Si uno de los argumentos no es numérico, DISTR.NORM.INV devuelve el valor de error #¡VALOR!
- Si probabilidad < 0 o si probabilidad > 1, DISTR.NORM.INV devuelve el valor de error #¡NUM!
- Si desv_estándar ≤ 0, DISTR.NORM.INV devuelve el valor de error #¡NUM!

7.2 Funciones relacionadas con otras distribuciones

- **DIST.GAMMA.INV** Devuelve el inverso de la función gamma acumulativa
- **DIST.GAMMA** Devuelve la distribución gamma
- **DISTR.BETA.INV** Devuelve el inverso de la función de densidad de probabilidad beta acumulativa
- **DISTR.BETA** Devuelve la función de densidad de probabilidad beta acumulativa
- **DISTR.CHI** Devuelve la probabilidad de una sola cola de la distribución chi cuadrado
- **DISTR.EXP** Devuelve la distribución exponencial
- **DISTR.F** Devuelve la distribución de probabilidad F
- **DISTR.INV.F** Devuelve el inverso de una distribución de probabilidad F
- **DISTR.LOG.INV** Devuelve el inverso de la distribución logarítmico-normal
- **DISTR.LOG.NORM** Devuelve la distribución logarítmico-normal acumulativa
- **DISTR.T.INV** Devuelve el inverso de la distribución t de Student
- **DISTR.T** Devuelve la distribución t de Student
- **DISTR.WEIBULL** Devuelve la distribución Weibull

7.3 Beta

Usos.

Debido a su gran flexibilidad se utiliza en situaciones en las que la ausencia de datos concretos no impide, sin embargo, tener una idea del comportamiento "global" de la variable aleatoria. Si suponemos conocidos, o razonablemente supuestos, valores tales como el máximo, mínimo, media o moda y el tipo de simetría (o asimetría), entonces es posible encontrar una distribución Beta que se adapte a dichas suposiciones.

También se utiliza para simular la proporción (o el número total) de productos defectuosos en un lote de fabricación, la duración de un proceso (en PERT/CPM), o la mediana de una muestra aleatoria.

Notación y parámetros.

La notación habitual es $X \sim B_e(\alpha,\beta)$ o bien $X \sim Beta(\alpha,\beta)$, los dos parámetros son de *forma* ($\alpha, \beta > 0$). En Excel la notación es diferente y se basa en el hecho de que la distribución puede ser fácilmente reescalada a un intervalo (a,b) ya que si $X \sim B_e(\alpha,\beta) \rightarrow 0 \le X \le 1$ al hacer $X' = a+(b-a)X$ tendríamos $X' \sim B_e(\alpha,\beta)$ pero ahora con $a \le X' \le b$. Así, la notación en Excel es $X \sim B_e(\alpha,\beta,a,b)$; en este caso los parámetros a y b son de *escala* en la distribución.

Densidad y Distribución.

$$f(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B_{(\alpha,\beta)}}$$

siendo $B(\alpha,\beta)$ la función Beta:

$$B_{(\alpha,\beta)} \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx$$

para la Beta de cuatro parámetros, usada en Excel, tendremos:

$$f(x) = \frac{1}{B_{(\alpha,\beta)}} \frac{(x-a)^{\alpha-1}(b-x)^{\beta-1}}{(b-a)^{\alpha+\beta-1}}$$

F(x) no tiene, en general, forma cerrada.

Estadísticos.

La media y varianza son (respectivamente):

$$\frac{\alpha}{\alpha+\beta} \quad ; \quad \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

el sesgo, la curtosis y el coeficiente de variación son (respectivamente):

$$\frac{2(\beta-\alpha)}{(\alpha+\beta+2)}\sqrt{\frac{\alpha+\beta+1}{\alpha\beta}} \quad ; \quad \frac{3(\alpha+\beta+1)[\alpha\beta(\alpha+\beta-6)+2(\alpha+\beta)^2]}{\alpha\beta(\alpha+\beta+2)(\alpha+\beta+3)} \quad ; \quad \sqrt{\frac{\beta}{\alpha(\alpha+\beta+1)}}$$

Generación.

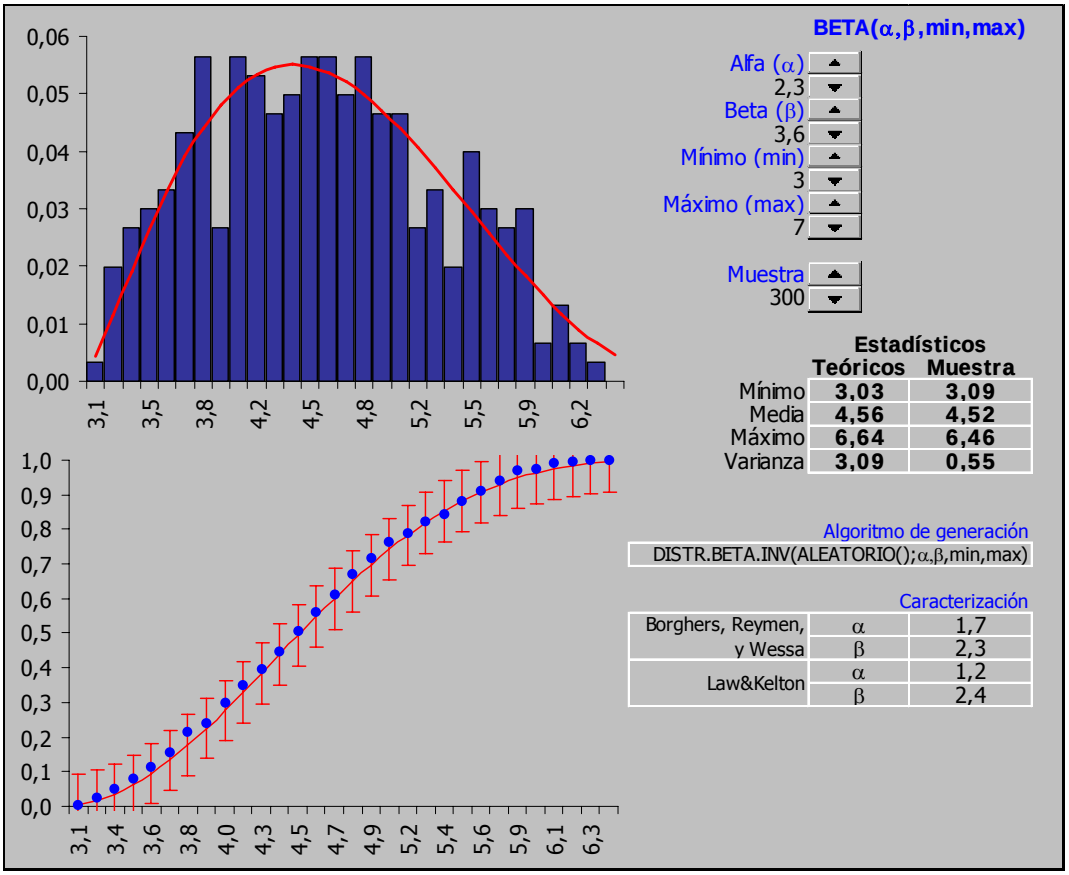
Puesto que Excel cuenta con una función para la inversa de la función de distribución, la generación de variables aleatorias puede hacerse directamente por inversión utilizando la fórmula siguiente:

DISTR.BETA.INV(ALEATORIO();α, β , a , b).

Caracterización.

Los parámetros pueden ser estimados de la forma siguiente [W1]:

$$\hat{\alpha} = \bar{x}\left[\left[\frac{\bar{x}(1-\bar{x})}{s^2}\right]-1\right] \quad ;\hat{\beta} = (1-\bar{x})\left[\left[\frac{\bar{x}(1-\bar{x})}{s^2}\right]-1\right]$$



7.4 Chi cuadrado (χ^2)

Usos.

Es sabido que la suma de n variables normales estándar al cuadrado sigue una distribución χ^2 de n grados de libertad, sin embargo, este hecho no convierte a la distribución χ^2 en candidata para la modelización de ninguna magnitud, excepto si ésta fuera precisamente la suma anterior. Su uso en Simulación, o MonteCarlo, está más relacionada con el test de bondad del ajuste que lleva su nombre.

Notación y parámetros.

La notación habitual es $X \sim \chi^2(v)$, siendo **v** el parámetro conocido como grados de libertad (v>0).

Propiedades.

La distribución χ^2 es un caso particular de la distribución Gamma, $\chi^2_n \equiv$ Gamma(0,2,n)

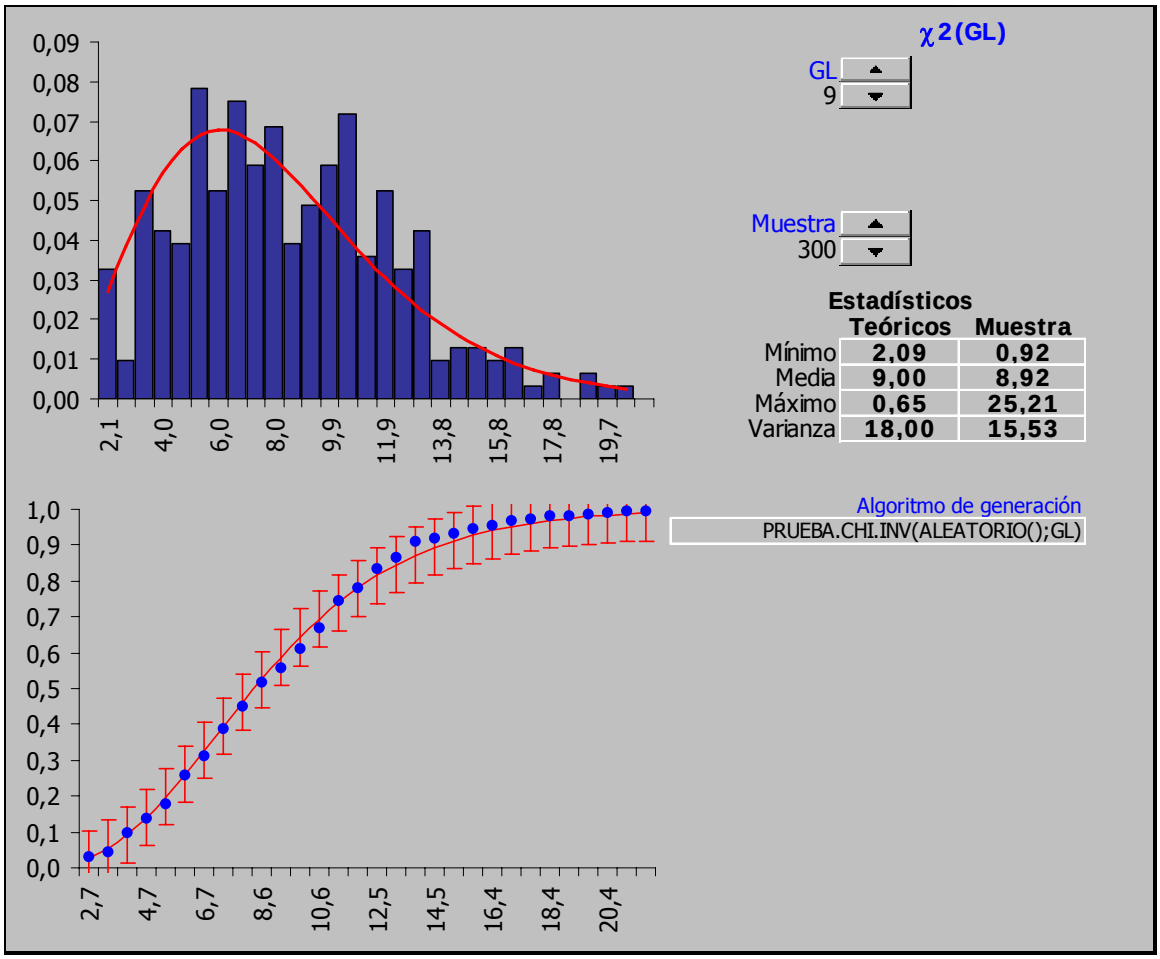
Generación.

La generación es inmediata usando la función de librería de Excel:

PRUEBA.CHI.INV(ALEATORIO();GL)

Hoja de cálculo.

El fichero Chi2.xls contiene una hoja que posibilita la descripción gráfica y la generación, su aspecto es el siguiente:



7.5 Exponencial

La distribución exponencial es una de las más utilizadas en simulación, sus valores son siempre positivos lo que la liga fundamentalmente con la modelización de "tiempos", pero lo que la convierte en sumamente importante es el hecho de que se trata de la única distribución continua cuya tasa de fallo es constante, o dicho de otra forma, no tiene memoria. Esto supone que la magnitud simulada, el tiempo necesario para que se complete una tarea, el tiempo hasta el fallo de un dispositivo mecánico, el tiempo entre llegadas de los clientes a una cola, es independiente del instante del tiempo en que nos encontremos y por tanto del tiempo transcurrido hasta ese momento.

Esta propiedad (conocida en la literatura anglosajona como "memoryless property") es harto frecuente, determinados dispositivos electrónicos, por ejemplo, no sufren desgaste y por lo tanto prácticamente no envejecen por lo que su probabilidad de fallo no aumenta a lo largo de su vida útil. Por otra parte, si el número de sucesos ocurridos en un intervalo de tiempo sigue una distribución de Poisson, lo cual es harto frecuente, entonces el tiempo entre dos de estos sucesos se distribuye de forma exponencial.

Notación y parámetros.

La notación habitual es **X~Exp(β)**, β es parámetro de *escala* (β>0).

Densidad y Distribución.

La función de densidad es:

$$f(x) = \frac{1}{\beta} e^{-\left(\frac{X}{\beta}\right)}$$

,la función de distribución es:

$$F(x) = 1 - e^{-\left(\frac{X}{\beta}\right)}$$

Estadísticos.

La media es β, la varianza $\beta 2$; el sesgo 2, la curtosis 9 y el coeficiente de variación 1.

Propiedades.

Es un caso particular de la distribución Gamma verificándose que Gamma(α,1) $\equiv$ Exp(α); también es un caso particular de la Weibull Weibull(α,1) $\equiv$ Exp(α); la suma de exponenciales independientes de parámetro β es una distribución Erlang(k;β)

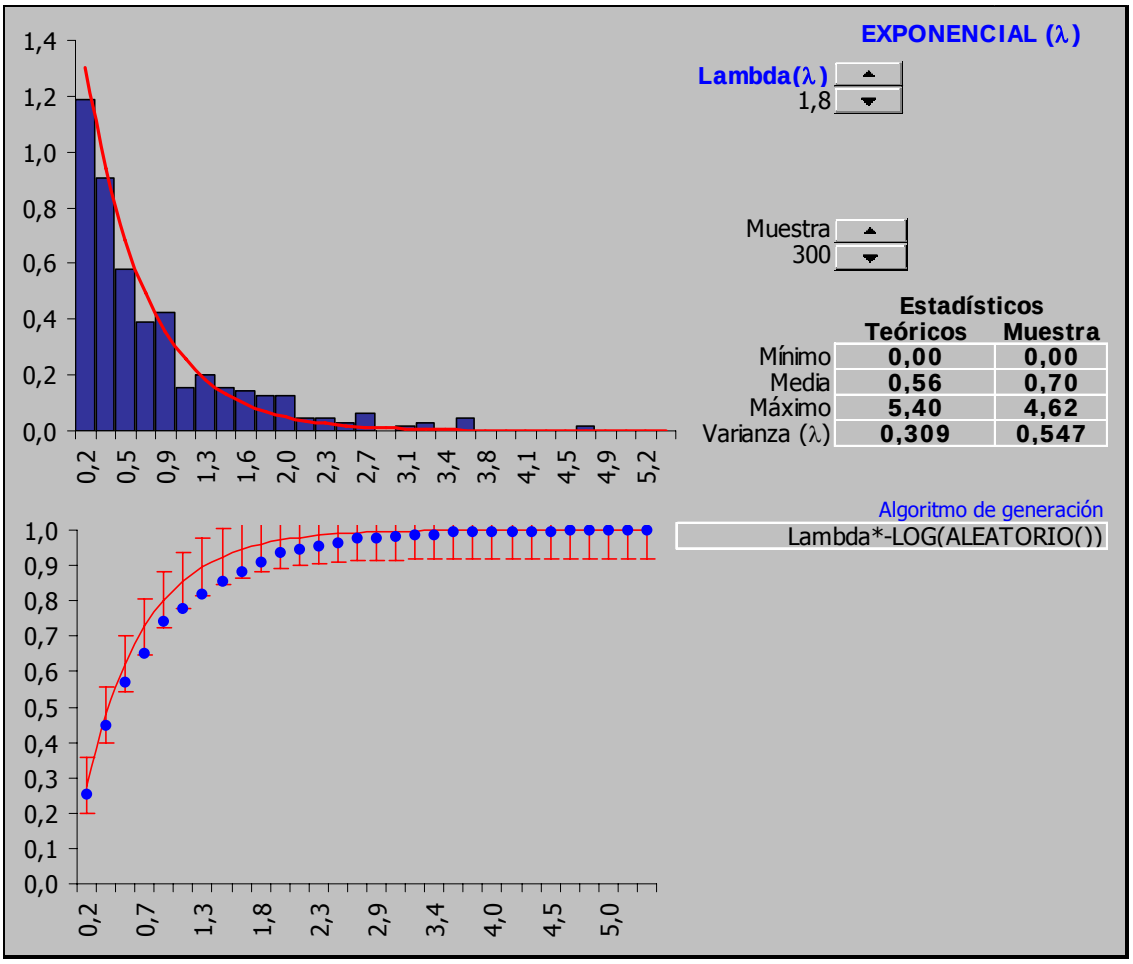
Generación.

Excel no cuenta con una función para la inversa de la función de distribución, sin embargo, la generación de variables aleatorias puede hacerse utilizando la fórmula siguiente:

(1/β) * -LOG(ALEATORIO())

Hoja de cálculo.

El fichero Exponencial.xls es una plantilla para la generación y análisis de esta distribución en Excel. Nótese que en la hoja se ha utilizado una notación ligeramente distinta (cambiando tasa por media) de manera que λ=1/β.



7.6 F (de Snedecor)

Usos.

Esta distribución tiene un papel fundamental en determinados contrastes de hipótesis (pruebas sobre las varianzas y ANOVA), fuera de estas aplicaciones no suele usarse para modelizar magnitud alguna.

Notación y parámetros.

La notación habitual es $X \sim F(gl_1, gl_2)$, ambos parámetros, conocidos como grados de libertad del numerador y g.l. del denominador son de *forma* ($gl_1; gl_2 > 0$).

Densidad y Distribución.

La función de densidad es:

$$f(x) = \frac{\left(\frac{gl_1}{gl_2}\right)^{\frac{gl_1}{2}} X^{\frac{gl_1}{2}-1}}{B\left[\frac{gl_1}{2};\frac{gl_2}{2}\right]\left[1+X\left(\frac{gl_1}{gl_2}\right)\right]^{\frac{gl_1+gl_2}{2}}}$$

mientras que la función de distribución no tiene forma cerrada.

Estadísticos.

La media y varianza son (respectivamente):

$$\frac{gl_1}{gl_1-2} \quad ; \quad \frac{2gl_1^{\,2}(gl_1+gl_2-2)}{gl_2(gl_1-4)(gl_1-2)^2}$$

Propiedades.

Nótese que la media de la distribución no depende de gl_1; al aumentar los grados de libertad de la distribución, ésta se aproxima cada vez más a la distribución Normal; se verifica que: $F(gl_1,gl_2) \equiv 1/\ F(gl_2,gl_1)$

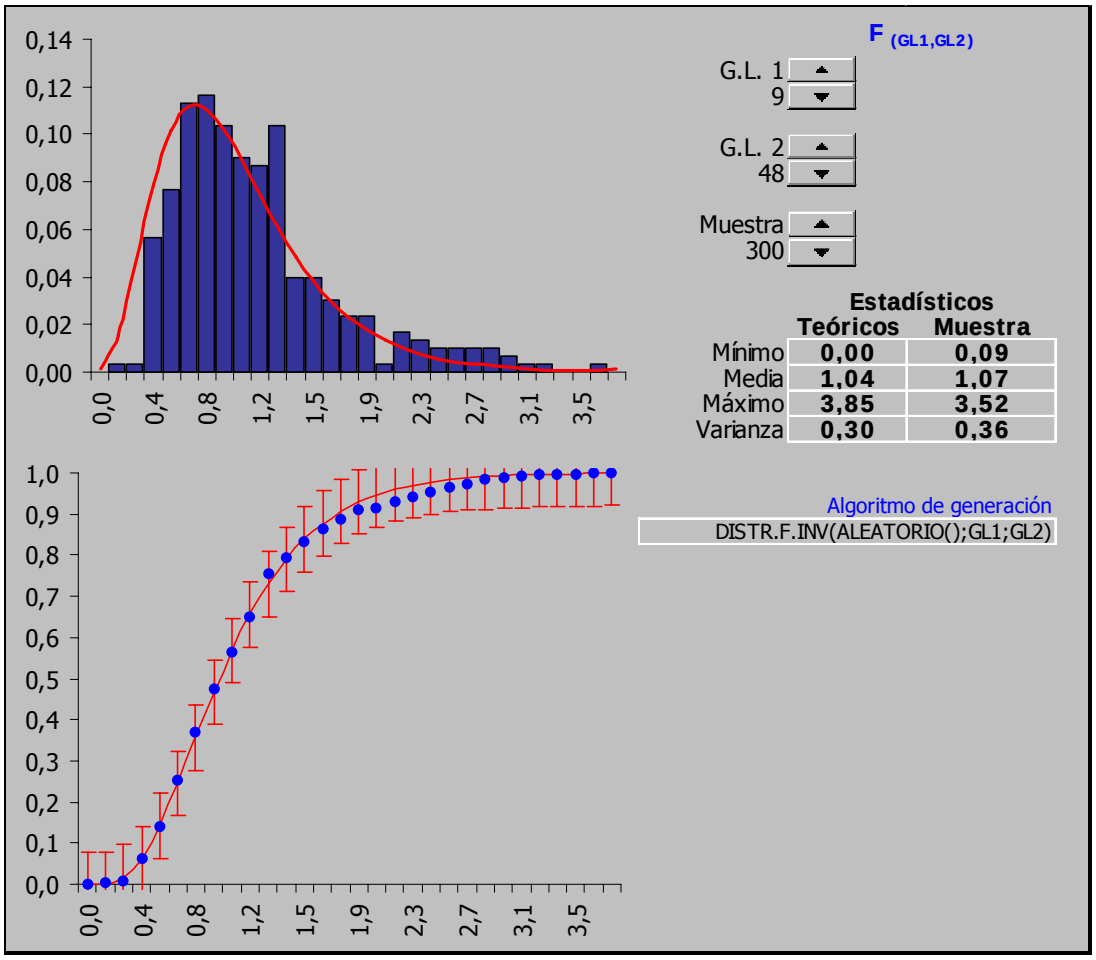
Generación.

Excel cuenta con una función para la inversa de la función de distribución, la generación de variables aleatorias puede hacerse utilizando la fórmula siguiente:

DISTR.F.INV(ALEATORIO();GL_1;GL_2)

Hoja de cálculo.

El fichero FSnedecor.xls es una plantilla para la generación y análisis de esta distribución en Excel. Su aspecto es el siguiente:



7.7 Gamma

La distribución Gamma es la generalización de algunas de las distribuciones más usadas en la modelización de fenómenos para su simulación: la exponencial, y la Erlang no son sino casos particulares (junto con la χ^2) de la distribución Gamma. Su empleo en Simulación/MonteCarlo está relacionado con los fenómenos de espera, el hecho de que sea siempre positiva la liga a magnitudes como el tiempo para realizar una tarea o el tiempo hasta el fallo de un dispositivo, entre otras posibles aplicaciones.

Estas aplicaciones se derivan del hecho de que puede considerarse como la probabilidad de que ocurran α sucesos en un periodo $(1/\beta)$ de tiempo (por ejemplo

que fallen los k subsistemas de un dispositivo que harán que éste finalmente deje de funcionar; que se lleven a cabo las k subtareas que componen un tarea principal con lo que ésta puede considerarse terminada, etc.)

Notación y parámetros.

La notación habitual es X~Gamma(α,β), α (α>0) es un parámetro de *forma* y β (β>0) de *escala*.

Densidad y Distribución.

La función de densidad es:

$$f(x) = \frac{\beta^{-\alpha} X^{\alpha-1} e^{\left(\frac{X}{\beta}\right)}}{\Gamma_\alpha}$$

,la función de distribución es:

$$F(x) = 1 - e^{-\left(\frac{X}{\beta}\right)} \left[\sum_{j=0}^{j=\alpha-1} \frac{\left(X/\beta\right)^j}{j!} \right]$$

Estadísticos.

La media y varianza son (respectivamente):

$$\alpha\beta \quad ; \quad \alpha\beta^2$$

el sesgo, la curtosis y el coeficiente de variación son (respectivamente):

$$2\sqrt{\frac{1}{\beta}} \quad ; \quad 3 + \frac{6}{\beta} \quad ; \quad \sqrt{\frac{1}{\beta}}$$

Propiedades.

Gamma(1,β) $\equiv$ Exp(β) ; si k es un entero positivo a la distribución Gamma(k,β) se la conoce como k-Erlang; a la distribución Gamma(v/2,2) se la conoce como $\chi^2_{v.}$

Si {X_1,X_2,..X_n} se distribuyen como Gamma(α_1,β), Gamma(α_2,β),.. entonces la suma X_1+X_2+.. se distribuye según Gamma($\alpha_1+\alpha_2+$...,β).

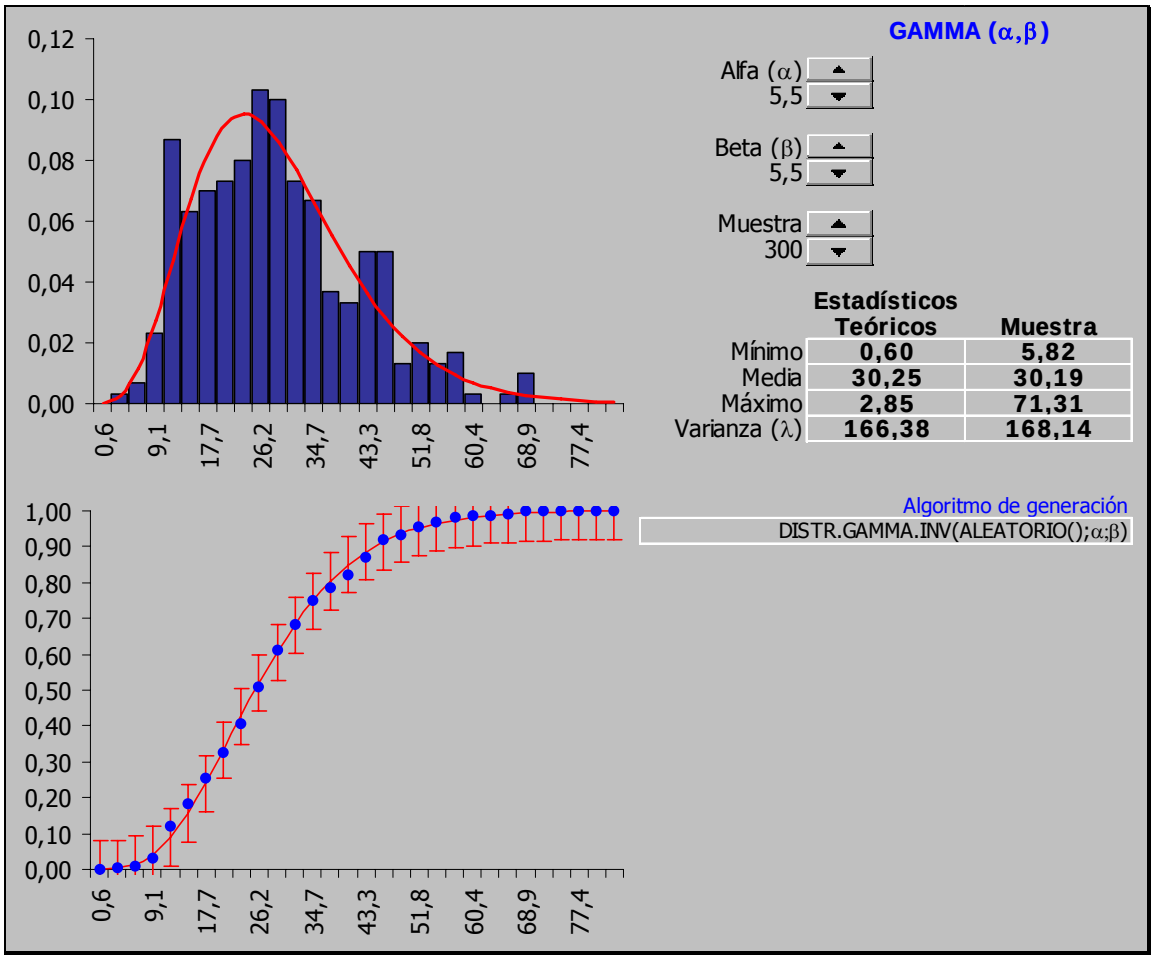
Generación.

Excel cuenta con una función para la inversa de la función de distribución, la generación de variables aleatorias puede hacerse utilizando la fórmula siguiente:

DISTR.GAMMA.INV(ALEATORIO();α;β)

Hoja de cálculo.

El fichero Gamma.xls es una plantilla para la generación y análisis de esta distribución en Excel. Su aspecto es el siguiente:



7.8 LogNormal

De la misma manera que la suma de un número (suficiente) de variables aleatorias positivas se distribuye de forma normal, el producto de un número (suficiente) de variables aleatorias positivas se distribuye de forma log-normal.

Puesto que la distribución es siempre positiva, se emplea también para modelizar tiempos: tiempo hasta el fallo de un dispositivo; tiempo para llevar a cabo una tarea.

Notación y parámetros.

La notación habitual es $X \sim LN(\mu,\sigma^2)$; μ es el parámetro de *escala* y σ el de *forma* ($\sigma>0$).

Densidad y Distribución.

La función de densidad es:

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

la función de distribución no tiene forma cerrada.

Estadísticos.

La media y la varianza son, respectivamente:

$$e^{\left(\frac{\mu+\sigma^2}{2}\right)} \quad ; \quad e^{2\mu+\sigma^2}\left(e^{\sigma^2}-1\right)$$

el sesgo, la curtosis y el coeficiente de variación son (respectivamente):

$$\left(e^{\sigma^2}+2\right)\sqrt{e^{\sigma^2}-1} \quad ; \quad e^{4\sigma^2}+2e^{3\sigma^2}+3e^{2\sigma^2}-3 \quad ; \quad \sqrt{e^{\sigma^2}-1}$$

Propiedades.

También conocida como distribución Cobb-Douglas. Siempre es sesgada hacia la derecha y nunca toma valores negativos.

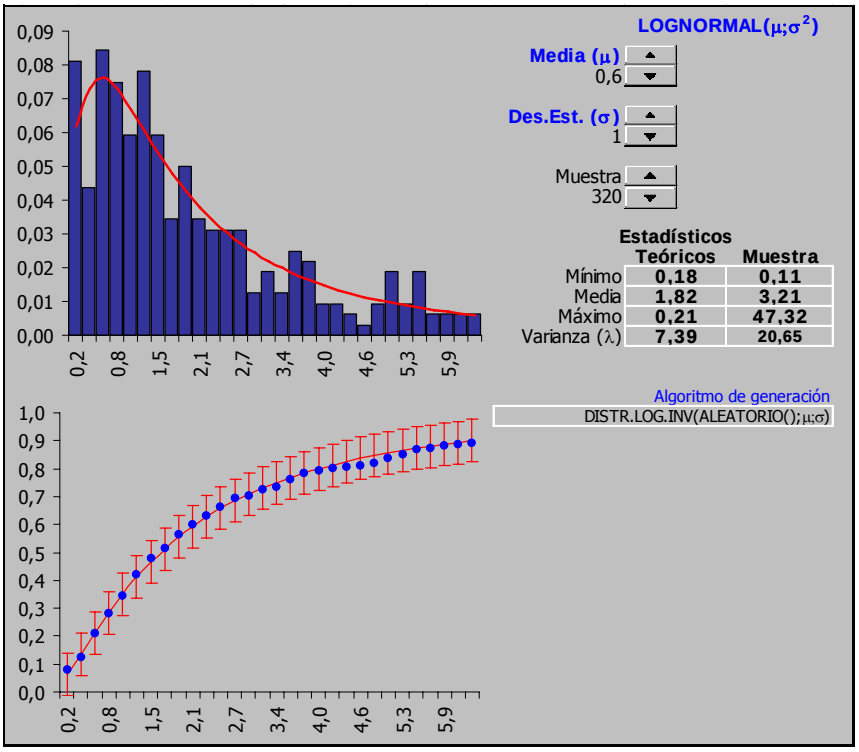
Generación.

Puesto que Excel cuenta con la función de distribución inversa entre sus funciones estadísticas, la generación es extraordinariamente sencilla, basta emplear la fórmula siguiente:

DISTR.LOG.INV(ALEATORIO();μ;σ)

Hoja de cálculo.

El fichero LogNorm.xls contiene una hoja que posibilita la descripción gráfica y la generación de v.a. log-normales. Su aspecto es el siguiente:



7.9 Normal

En virtud del Teorema Central de Límite cualquier magnitud que sea suma de otras magnitudes, seas éstas como sean, se distribuirá de forma normal.

Notación y parámetros.

La notación habitual es $X \sim N(\mu,\sigma)$, siendo μ el parámetro de *posición* y σ el parámetro de escala ($\sigma>0$).

Densidad

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

Estadísticos.

La media es μ, la varianza σ^2, el sesgo 0, la curtosis 3 y el coeficiente de variación σ/μ.

Propiedades.

La distribución es simétrica, centrada en μ y con puntos de inflexión en $\mu\pm\sigma$; la suma de n variables $N(\mu,\sigma^2)$ es $N(n\mu,n\sigma^2)$; un gran número de distribuciones están relacionadas con la Normal: t, F, χ^2, LogNormal, Cauchy.

Generación.

Excel cuenta con la función inversa de la distribución:

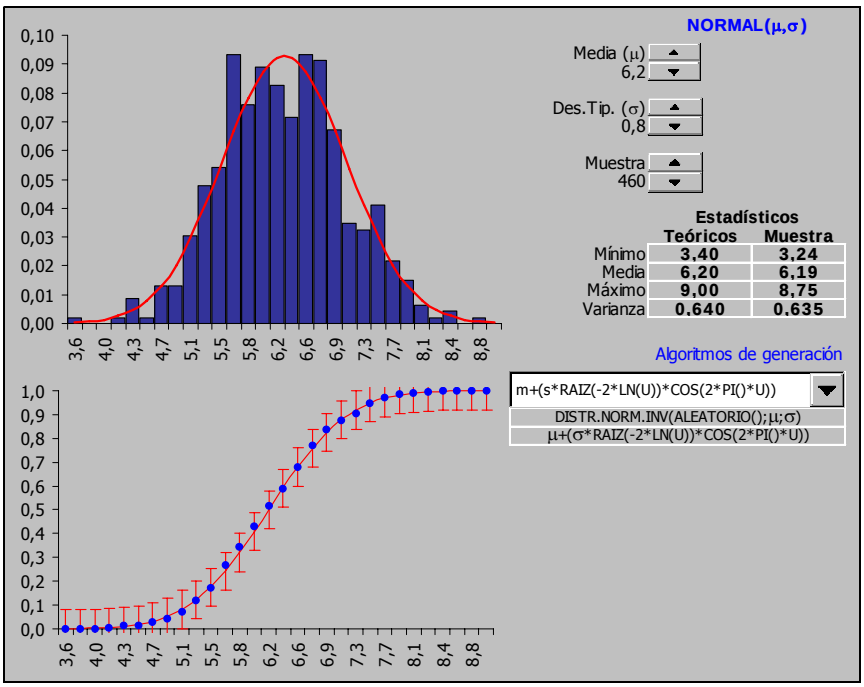
DISTR.NORM.INV(ALEATORIO();μ;σ)

también en la literatura aparecen descritos diversos métodos para generar Normales uno de los más efectivos es el conocido como Box-Muller:

μ+σ*RAIZ(GL*(ALEATORIO()^(-2/GL)-1))*COS(2*PI()*ALEATORIO())

Hoja de cálculo.

El fichero Normal.xls contiene una hoja que posibilita la descripción gráfica y la generación, por los dos métodos expuestos, de v.a. Normales.



7.10 t de Student

Esta distribución tiene un papel fundamental en determinados contrastes de hipótesis (pruebas sobre igualdad de medias), fuera de esta aplicación podría usarse para modelizar la desviación de la media de una muestra respecto de la media de la población de la que ésta procede.

Notación y parámetros.

La notación habitual es X~*t*(GL) siendo GL el único parámetro de *forma* (GL>0).

Densidad y Distribución.

La función de densidad es:

$$f(x) = \frac{\Gamma\left(\frac{GL+1}{2}\right)}{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{GL}{2}\right)} GL^{-1/2}\left[1+\frac{X^2}{GL}\right]^{\frac{-GL+1}{2}}$$

Estadísticos.

La media (para GL>1) y la varianza (para GL>2) son, respectivamente:

$$0 \quad ; \quad \frac{Gl}{(GL-2)}$$

Propiedades.

Para GL>30 la distribución es prácticamente una Normal; se verifica que t(1) ≡ $\text{Cauchy}_{(0,1)}$

Generación.

Excel cuenta con la función inversa de la distribución si bien sólo para valores positivos de X de manera que es necesaria una pequeña modificación:

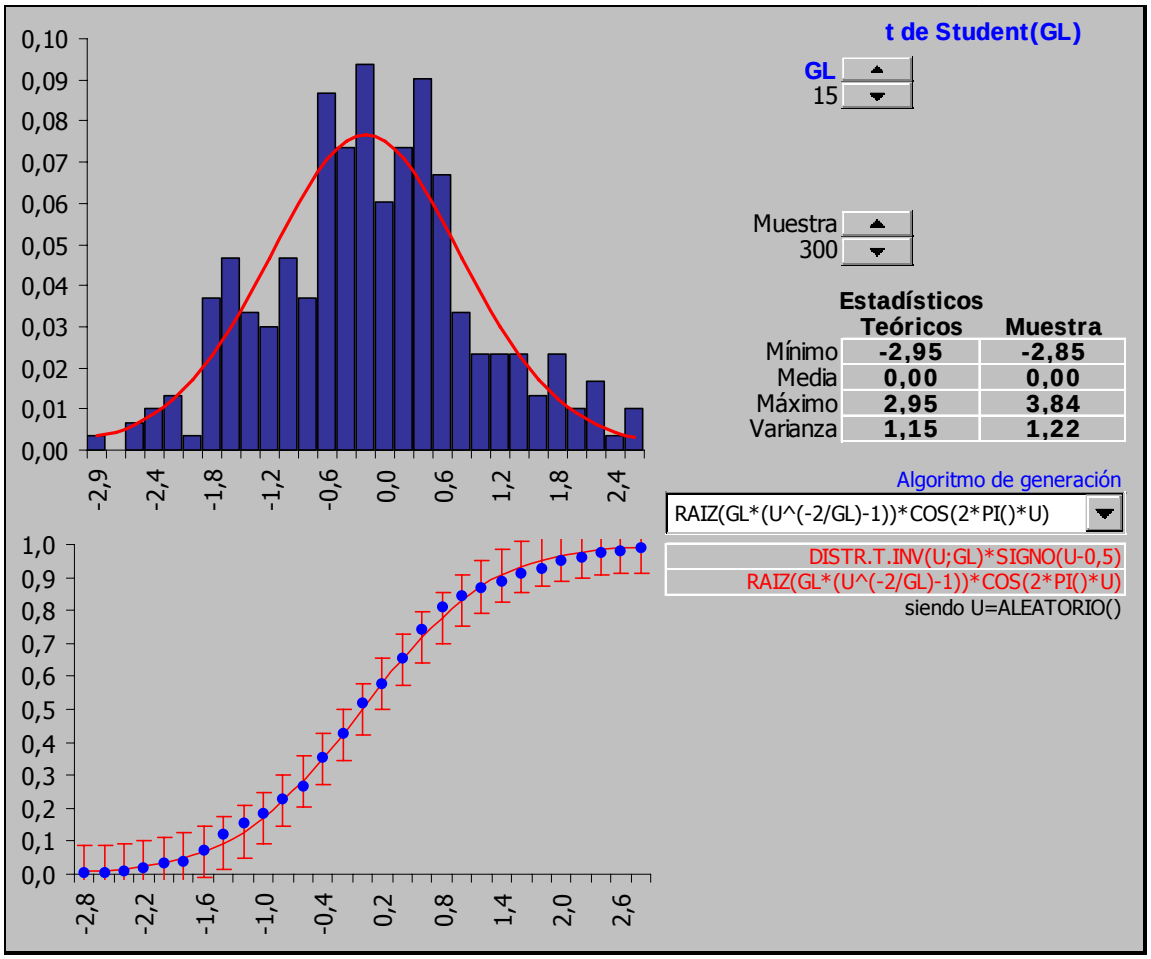
DISTR.T.INV(ALEATORIO();GL)*SIGNO(ALEATORIO()-0,5)

también en la literatura aparecen descritos diversos métodos para generar v.a. distribuidas según una ***t*** de Student, uno de los más efectivos es el que utiliza la fórmula siguiente:

RAIZ(GL*(ALEATORIO()^(-2/GL)-1))*COS(2*PI()*ALEATORIO())

Hoja de cálculo.

El fichero Student.xls contiene una hoja que posibilita la descripción gráfica y la generación, por los dos métodos expuestos, de v.a. de Pareto. Su aspecto es el siguiente:



7.11 Pareto

La distribución de Pareto aparece asociada a multitud de magnitudes naturales. Es profusamente empleada para modelizar aspectos tales como: la distribución de la renta de los individuos (cuando ésta supera un cierto umbral β); las reclamaciones de seguros; la distribución de recursos naturales en zonas geográficas; el tamaño de las ciudades; el numero de empleados de las empresas; las fluctuaciones de los precios en los mercados de valores, entre otras. En algunos textos la encontramos exclusivamente asociada a la distribución de los ingresos de los individuos: *"la probabilidad de que la renta de un individuo supere una cierta cantidad A es una variable aleatoria de Pareto(α=A,)"*.

En general, es una distribución a tener en cuenta para modelizar una magnitud (positiva) cuando en ésta se cumpla que un pequeño porcentaje de valores aparece un gran número de veces y es posible un elevado número de valores extremos aunque muy poco probables.

Notación y parámetros.

La notación habitual es X~Par(α,β), ambos parámetros son de *escala* (α,β>0), además β indica el valor mínimo posible de la variable ($\beta \le X < \infty$).

Densidad y Distribución.

La función de densidad es:

$$f(x) = \frac{\alpha \beta^{\alpha}}{X^{\alpha+1}}$$

,la función de distribución es:

$$F(x) = 1 - \left(\frac{\beta}{X}\right)^{\alpha}$$

Estadísticos.

La media (para α>1) y la varianza (para α>2) son, respectivamente:

$$\frac{\alpha\beta}{\alpha-1} \quad ; \quad \frac{\alpha\beta^2}{(\alpha-1)^2(\alpha-2)}$$

el sesgo y la curtosis son (respectivamente):

$$\frac{2(\alpha+1)}{(\alpha-3)}\sqrt{\frac{\alpha-2}{\alpha}} \quad ; \quad \frac{3(3\alpha^2+\alpha+2)(\alpha-2)}{\alpha(\alpha-3)(\alpha-4)}$$

Propiedades.

La distribución siempre es sesgada hacia la derecha y nunca toma valores negativos, nótese que los momentos de orden k sólo existen si α>k.

Generación.

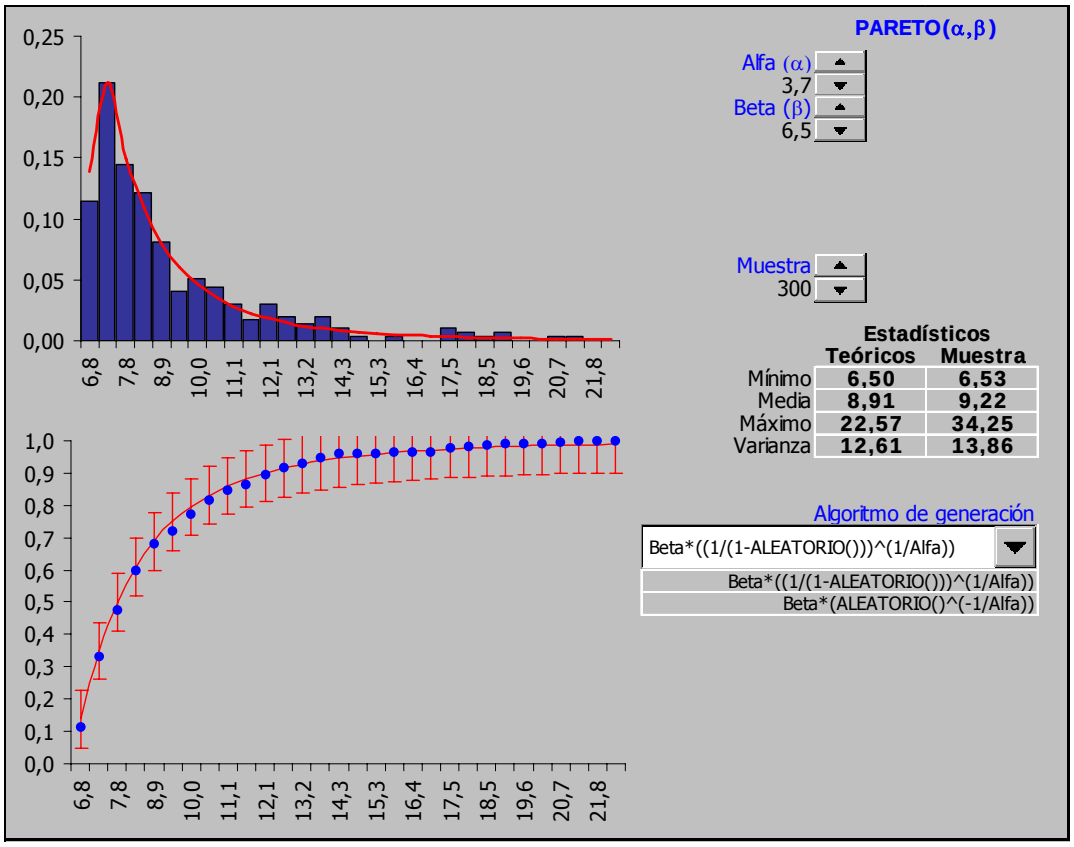
En la literatura aparecen descritos diversos métodos para generar v.a. de Pareto. En Excel es posible obtener v.a. a través de cualquiera de las fórmulas siguientes:

β*((1/(1-ALEATORIO()))^(1/α))

β*(ALEATORIO()^(-1/α))

Hoja de cálculo.

El fichero Pareto.xls contiene una hoja que posibilita la descripción gráfica y la generación, por los dos métodos expuestos, de v.a. de Pareto. Su aspecto es el siguiente:



7.12 Triangular

Su uso es como aproximación a la modelización de una magnitud aleatoria de la que no se cuenta con datos y únicamente puede aventurarse un mínimo y máximo absolutos y un valor modal.

Notación y parámetros.

La notación habitual es X~Tri(a,b,c), el parámetro a es de *posición* mientras que b es de *forma* y c es parámetro de *escala:* ($a \le b \le c$) y ($a \le X \le c$).

Densidad y Distribución.

La función de densidad es:

$$f(x) = \begin{cases} \dfrac{2(X-a)}{(b-a)(c-a)} & a \le X \le c \\ \dfrac{(b-X)}{(b-a)(b-c)} & c < X \le b \end{cases}$$

,la función de distribución es:

$$F(x) = \begin{cases} \dfrac{(X-a)^2}{(b-a)(c-a)} & a \le X \le c \\ 1 - \dfrac{(b-X)}{(b-a)(b-c)} & c < X \le b \end{cases}$$

Estadísticos.

La media y varianza son (respectivamente):

$$\frac{a+b+c}{3} \quad ; \quad \frac{a^2+b^2+c^2-ab-ac-bc}{18}$$

Propiedades.

Si a=c la distribución se convierte en una Triangular izquierda; si c=b la distribución se convierte en una triangular derecha.

Generación.

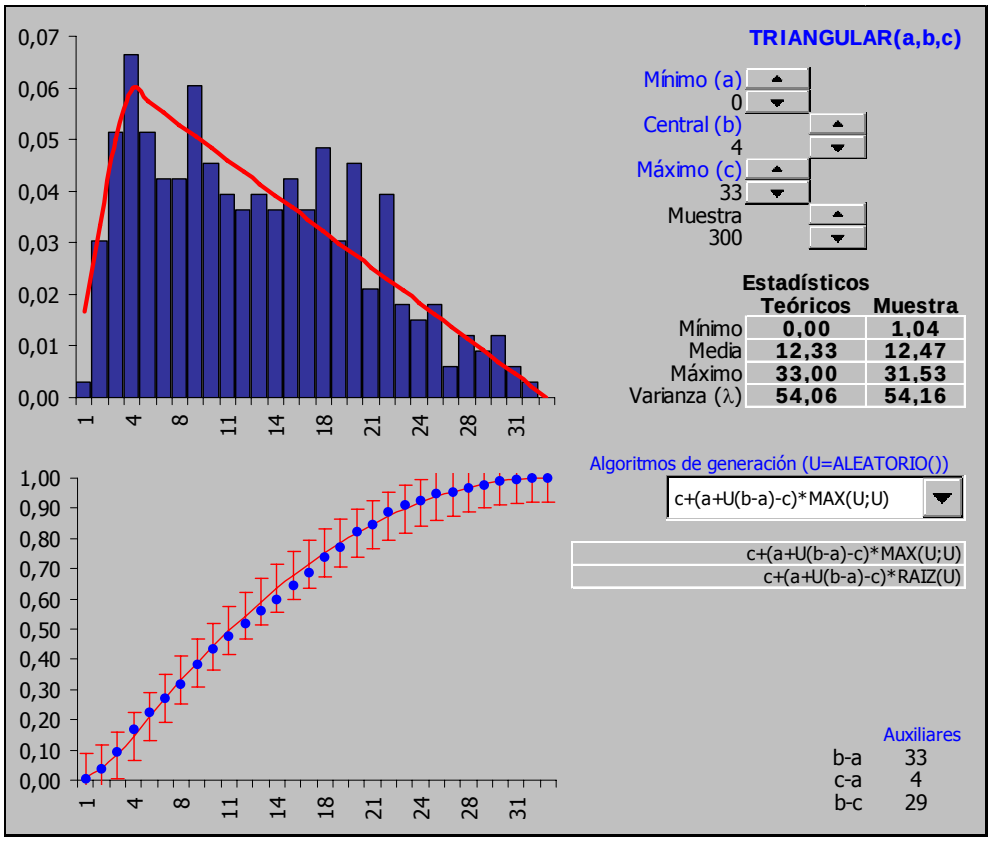
Excel no cuenta con una función para la inversa de la función de distribución, sin embargo, la generación de variables aleatorias puede hacerse utilizando cualquiera de las dos fórmulas siguientes:

c + (a + ALEATORIO()*(b-a)-c)* MAX(ALEATORIO();ALEATORIO())

c + (a + ALEATORIO()*(b-a)-c)* RAIZ(ALEATORIO())

Hoja de cálculo.

El fichero Triang.xls es una plantilla para la generación y análisis de la distribución Triangular en Excel. Su aspecto es el siguiente:



7.13 Uniforme

Su uso es como aproximación a la modelización de una magnitud aleatoria de la que no se cuenta con datos y únicamente puede aventurarse un mínimo y máximo absolutos no pudiéndose hacer conjeturas sobre su distribución dentro de ese intervalo. Por otra parte es la base de la generación del resto de variables aleatorias.

Notación y parámetros.

La notación habitual es X~U(a,b), el parámetro a es de *posición*, mientras que la cantidad b-a (b>a) determina la *escala* de la distribución.

Densidad y Distribución.

La función de densidad es:

$$f(x) = \frac{1}{b-a}$$

,la función de distribución es:

$$F(x) = \frac{X-a}{b-a}$$

Estadísticos.

La media y varianza son (respectivamente):

$$\frac{a+b}{2} \quad ; \quad \frac{(b-a)^2}{12}$$

el sesgo, la curtosis y el coeficiente de variación son (respectivamente):

$$0 \quad ; \quad \frac{9}{5} \quad ; \quad \frac{1}{\sqrt{3}}\frac{b-a}{a+b}$$

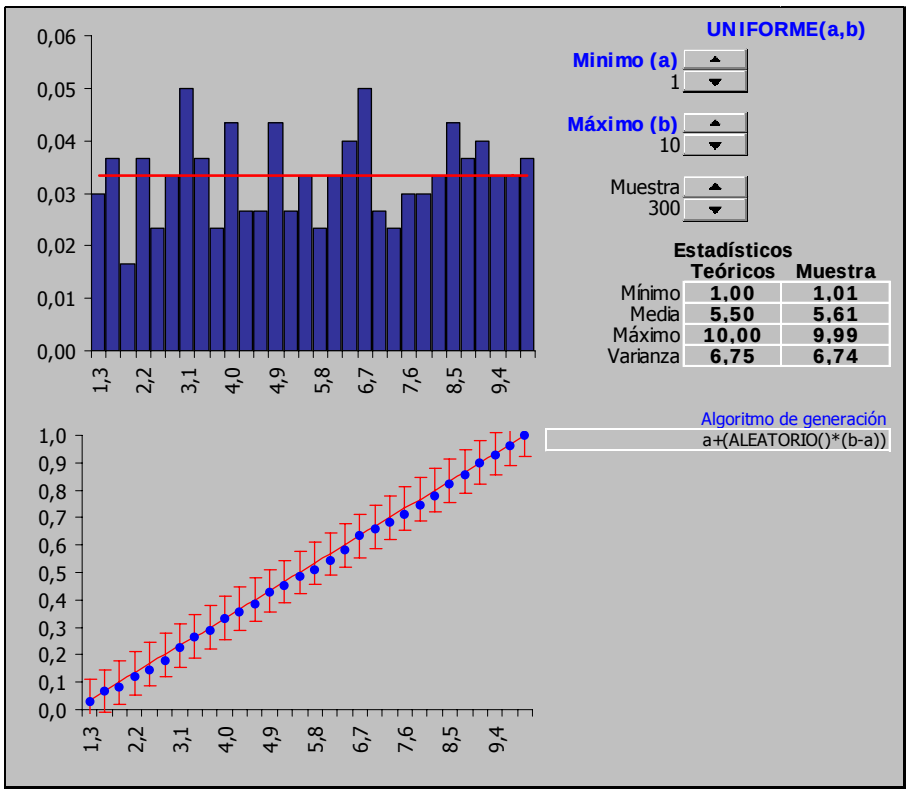
Generación.

Excel cuenta con una función para la generación de variables aleatorias uniformes, la v.a. U(0,1) se obtiene a través de la función **ALEATORIO()**, mientras que a partir de ésta puede obtenerse la de la U(a,b) sin más que usar la fórmula

a + (b-a)*ALEATORIO().

Hoja de cálculo.

El fichero Uniforme.xls es una plantilla para la generación y análisis de esta distribución en Excel. Su aspecto es el siguiente:



7.14 PROBLEMAS

7.14.1 Se sabe que el peso de un colectivo se distribuye con arreglo a una N(π = 100 Kg. ; σ=10 Kg.). ¿Cuál es la probabilidad de que un integrante de dicho colectivo pese más de 115 Kg. o menos de 85 Kg?.

7.14.2 Hacer un gráfico de la distribución Normal estándar (60 puntos).

7.14.3 Reproducir la tabla E.2.a del libro.

7.14.4 Reproducir la tabla E.2.b del libro.

7.14.5 Una persona espera un autobús desde las 12:00 horas hasta la 13:00. El autobús puede llegar en cualquier momento entre esos límites. Generar 100 valores aleatorios de otras tantas horas de llegada de un supuesto autobús y describir la muestra generada.

7.14.6 Sobre los datos anteriores contrastar los valores empíricos de la media, máximo, mínimo, primer, tercer cuartil y mediana con los esperados según la teoría.

7.14.7 Usando la fórmula del problema 3.2.2 generar 100 valores de una distribución N(12;2). Graficar los datos y comparar con lo esperado superponiendo la densidad de la normal teórica al histograma de los datos.

7.14.8 Comprobar el proceso de normalización (Ver que al normalizar una N(π;σ) obtenemos idénticos resultados por ambas funciones).

7.14.9 Sumar 3 v.a. N(0;1) elevadas al cuadrado y comprobar que dicha suma se distribuya según una Chi-cuadrado de 3 grados de libertad.

7.14.10 Comprobar empíricamente el Teorema Central del Límite.

7.14.11 ¿A partir de que valor de **n**, la distribución normal aproxima razonablemente bien una distribución binomial B(**n**,½)?.

7.14.12 Las puntuaciones de un determinado test se sabe que se distribuyen según una N(μ=950,σ=50). La última aplicación del test a un grupo de 18 personas dio el siguiente resultado.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Datos	875	933	1010	1007	1035	910	998	852	1063	981	957	1018	963	1048	1023	1010

a) ¿Hay motivos para pensar que los parámetros han variado?). Utilizar el test de bondad del ajuste K-S descrito a continuación).

7.14.13 Prueba de bondad del ajuste de Kolmogorov-Smirnov.

Este contraste, que es válido únicamente para variables continuas, compara la función de distribución (probabilidad acumulada) teórica con la observada, y calcula un valor de discrepancia, representado habitualmente como Dn, que corresponde a la discrepancia máxima en valor absoluto entre la distribución observada y la distribución teórica. Es un test independiente de la distribución concreta a la que se suponen se han de ajustar los datos.

Para la aplicación del este test es necesario determinar en primer lugar la Frecuencia observada acumulada en los datos **$S_n(x)$**. Para ello se ordena la muestra de menor a mayor y se calcula:

$$S_n(x) = \frac{i}{n+1}$$

En segundo lugar debemos ser capaces de obtener la frecuencia acumulada teórica para cada uno de los datos de la muestra **$F_0(x)$**.

Una vez determinadas ambas frecuencias, se obtiene el máximo de las diferencias entre ambas, en la i-ésima posición de orden, que se denomina **D_n**.

$$D_n = \max_x \left| S_n(x) - F_0(x) \right|$$

Finalmente, dado un valor para la significación del test, se recurre a la tabla de valores críticos de Dn en la prueba de bondad de ajuste de Kolmogorov-Smirnov, y considerando el tamaño de la muestra, se establece lo siguiente:

Si D_n > D tabla, se rechaza H_0

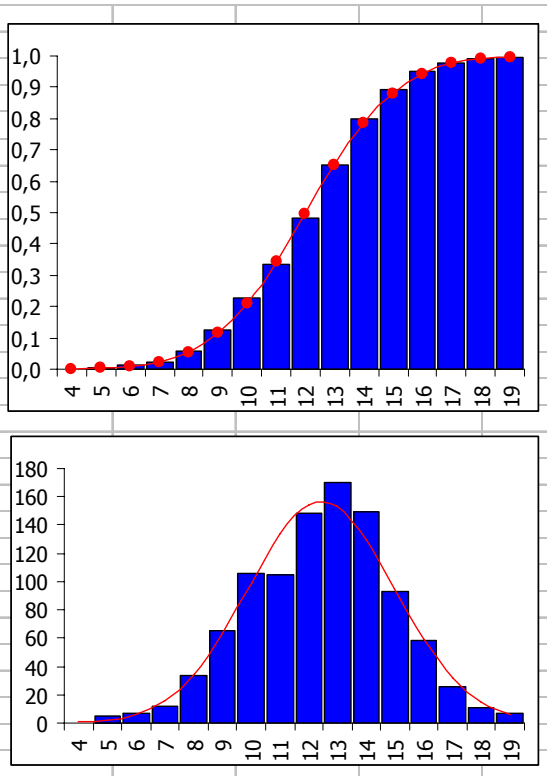
Distribución del estadístico de Kolmogorov-Smirnov-Lilliefors (D_n) para el contraste de normalidad.
Se tabula d tal que $P(D_n > d) = \alpha$.

	α					
n	0'2	0'15	0'1	0'05	0'01	0'001
4	0'303	0'321	0'346	0'376	0'413	0'433
5	0'289	0'303	0'319	0'343	0'997	0'439
6	0'269	0'281	0'297	0'323	0'371	0'424
7	0'252	0'264	0'280	0'304	0'351	0'402
8	0'239	0'250	0'265	0'288	0'333	0'384
9	0'227	0'238	0'252	0'274	0'317	0'365
10	0'217	0'228	0'241	0'262	0'304	0'352
11	0'208	0'218	0'231	0'251	0'291	0'338
12	0'200	0'210	0'222	0'242	0'281	0'325
13	0'193	0'202	0'215	0'234	0'271	0'314
14	0'187	0'196	0'208	0'226	0'262	0'305
15	0'181	0'190	0'201	0'219	0'254	0'296
16	0'176	0'184	0'195	0'213	0'247	0'287
17	0'171	0'179	0'190	0'207	0'240	0'279
18	0'167	0'175	0'185	0'202	0'234	0'273
19	0'163	0'170	0'181	0'197	0'228	0'266
20	0'159	0'166	0'176	0'192	0'223	0;260
25	0'143	0'150	0'159	0'173	0'201	0'236
30	0'131	0'138	0'146	0'159	0'185	0'217
> 30	$\frac{0'740}{\sqrt{n}}$	$\frac{0'770}{\sqrt{n}}$	$\frac{0'820}{\sqrt{n}}$	$\frac{0'890}{\sqrt{n}}$	$\frac{1'040}{\sqrt{n}}$	$\frac{1'220}{\sqrt{n}}$

7.14.14 Generar una muestra de 1000 valores de una distribución N(12;2,5). Hacer los siguiente:

a) Tabularla y obtener la distribución de frecuencias absolutas acumuladas y no acumuladas. Estimar su media y su desviación típica.
b) Representar en dos gráficos deferentes las dos tabulaciones de los datos con los valores teóricos según la estimación anterior de los parámetros.
c) Utilizar SOLVER para estima los parámetros que minimizan las diferencias según un test de bondad del ajuste basado en la Chi2.
d) Igual que el anterior pero basado en Kolmogorov-Smirnov. Comparar los resultados.

	12,444			12,020		
	2,548			2,523		
	1000					
Clase	Frec		Chi	Frec		Kol
4	0	0,646	0,646	0,0000	0,0007	0,0007
5	5	2,197	3,577	0,0050	0,0027	0,0023
6	7	6,400	0,056	0,0120	0,0085	0,0035
7	12	15,982	0,992	0,0240	0,0233	0,0007
8	34	34,218	0,001	0,0580	0,0555	0,0025
9	65	62,807	0,077	0,1230	0,1157	0,0073
10	106	98,830	0,520	0,2290	0,2117	0,0173
11	105	133,321	6,016	0,3340	0,3430	0,0090
12	148	154,184	0,248	0,4820	0,4968	0,0148
13	170	152,864	1,921	0,6520	0,6511	0,0009
14	149	129,928	2,799	0,8010	0,7837	0,0173
15	93	94,674	0,030	0,8940	0,8812	0,0128
16	58	59,141	0,022	0,9520	0,9426	0,0094
17	26	31,672	1,016	0,9780	0,9758	0,0022
18	11	14,541	0,862	0,9890	0,9911	0,0021
19	7	5,723	0,285	0,9960	0,9972	0,0012
			19,07			**0,02**



7.14.15 No sabemos si un determinado valor λ procede de una N(2;1) o de una N(3;2). Pero sabemos que los errores de imputación que cometemos se pagan con arreglo a la siguiente matriz de pagos:

PAGOS		Pero realmente el valor proviene de	
		A $\leftrightarrow$ N(2;1)	**B** $\leftrightarrow$ N(3;2)
Nosotros decimos que el valor λ proviene de	**A** $\leftrightarrow$ N(2;1)	10	-5
	B $\leftrightarrow$ N(3;2)	-4	10

a) Determinar la regla óptima de asignación Si $(\lambda \leq \Upsilon_0) \rightarrow \lambda \in A$; Si $(\lambda > \Upsilon_0) \rightarrow \lambda \in B$ y dibujar los pagos en $\pm 3\sigma$

8 Regresión lineal

8.1 Regresión

- **PENDIENTE** Calcula la pendiente (a) para un modelo y = ax + b +ε
- **INTERSECCION.EJE** Calcula el término (b) para un modelo y = ax + b +ε
- **TENDENCIA y PRONOSTICO** Ambas calculan el valor estimado, para un x dado, según un modelo lineal.
- **ESTIMACION.LINEAL** Devuelve los parámetros de una tendencia lineal
- **ESTIMACION.LOGARITMICA** Devuelve los parámetros de una tendencia exponencial

Existen varias posibilidades de realizar, a través de la hoja de cálculo Excel, la estimación por mínimos cuadrados de un modelo lineal con una única variable:

$$\hat{y}_i = \mathbf{a} \cdot x_i + \mathbf{b}$$

La más rápida y sencilla - quizás también la más completa - es a través de la opción **Análisis de Datos**, aunque en este documento utilizaremos también, a efectos de comprobar los resultados obtenidos mediante el método anterior, el cálculo directo realizado sobre la misma hoja.

Trabajaremos con el siguiente ejemplo:

X	Y
1	2550
2	2590
3	2640
4	2670
5	2750
6	2800
7	2850
8	2900

Una vez introducidos los datos en la hoja, llamaríamos al módulo de Análisis, pero antes de esto realizaremos algunos cálculos sobre estos valores. Los resultados que obtengamos serán los que determinen las características fundamentales del ajuste.

	A	B	C	D	E	F	G	H	I	J	K
1	X	Y	X-MedX	Y-MedY	(X-MedX)2	(Y-MedY)2	Yest	Yest-MedY	(Y-MedY)*(Yest-MedY)	C*D	(G-B)^2
2	1	2550	-3,500	-168,750	12,250	28476,563	2539,167	32250,198	30304,69875	590,625	117,363
3	2	2590	-2,500	-128,750	6,250	16576,563	2590,476	16454,193	16515,26463	321,875	0,227
4	3	2640	-1,500	-78,750	2,250	6201,563	2641,786	5923,519	6060,9465	118,125	3,188
5	4	2670	-0,500	-48,750	0,250	2376,563	2693,095	658,174	1250,676375	24,375	533,384
6	5	2750	0,500	31,250	0,250	976,563	2744,405	658,159	801,70625	15,625	31,309
7	6	2800	1,500	81,250	2,250	6601,563	2795,714	5923,473	6253,333125	121,875	18,369
8	7	2850	2,500	131,250	6,250	17226,563	2847,024	16454,116	16835,91	328,125	8,859
9	8	2900	3,500	181,250	12,250	32851,563	2898,333	32250,090	32549,43688	634,375	2,779
10					42,000	111287,500		110571,921	110571,973	2155,000	715,476
11											
12				PROMEDIO(A2:A9)	4,5	Media de X					
13				PROMEDIO(B2:B9)	2718,75	Media de Y					
14				J10/E10	**51,3095**	**Pendiente**					
15				F13-(F14*F12)	**2487,8571**	**Intercepción**					
16				RAIZ(K10/6)	**10,91998**	**Desviación estándar**					
17				F16/RAIZ(E10)	**1,685**	**Error estándar de la pendiente**					
18				H10^2	12226161103	Numerador					
19				D10*G10	12305272680	Denominador					
20				**D19/D20**	**0,9936**	**Coeficiente de Determinación**					

En la hoja anterior nos hemos limitado, simplemente, a calcular los parámetros de la recta de regresión:

- Pendiente: $r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} = \frac{\mathrm{cov}(x,y)}{s_x^2}$.
- Intersección: $b = \bar{y} - m\,\bar{x}$.

En la columna "J" y en la "E" hemos calculado, respectivamente:

$$\sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y}) \quad ; \quad \sum_{i=1}^{n} (x_i - \bar{x})^2$$

finalmente en la celda correspondiente calculamos, primero la pendiente:

PROMEDIO(A2:A9)	4,5	Media de X
PROMEDIO(B2:B9)	2718,75	Media de Y
J10/E10	**51,3095**	**Pendiente**
F13-(F14*F12)	2487,8571	Intercepción
RAIZ(K10/6)	10,91998	Desviación estándar
F16/RAIZ(E10)	1,685	Error estándar de la pendiente
H10^2	12226161103	Numerador
D10*G10	12305272680	Denominador
D19/D20	0,9936	Coeficiente de Determinación

y después, aprovechando la media de **x** e **y**, y la **pendiente** recién calculada, obtenemos el valor de la **intercepción:**

PROMEDIO(A2:A9)	4,5	Media de X
PROMEDIO(B2:B9)	2718,75	Media de Y
J10/E10	51,3095	Pendiente
F13-(F14*F12)	**2487,8571**	**Intercepción**
RAIZ(K10/6)	10,91998	Desviación estándar
F16/RAIZ(E10)	1,685	Error estándar de la pendiente
H10^2	12226161103	Numerador
D10*G10	12305272680	Denominador
D19/D20	0,9936	Coeficiente de Determinación

Finalmente, calculamos el valor de R^2 usando el resultado de las columnas "H", "D" y "G":

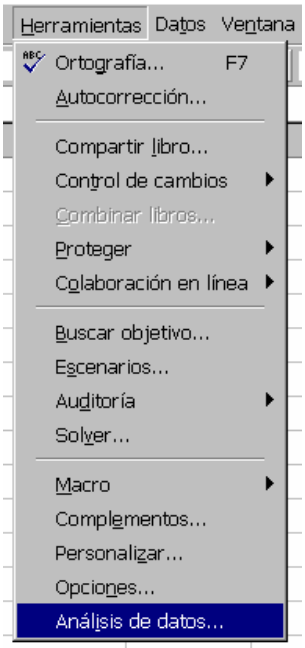
En general, el **coeficiente de determinación** R^2 se puede definir como el cuadrado de la correlación entre los valores de y_i y los valores estimados $\hat{y}_i$:

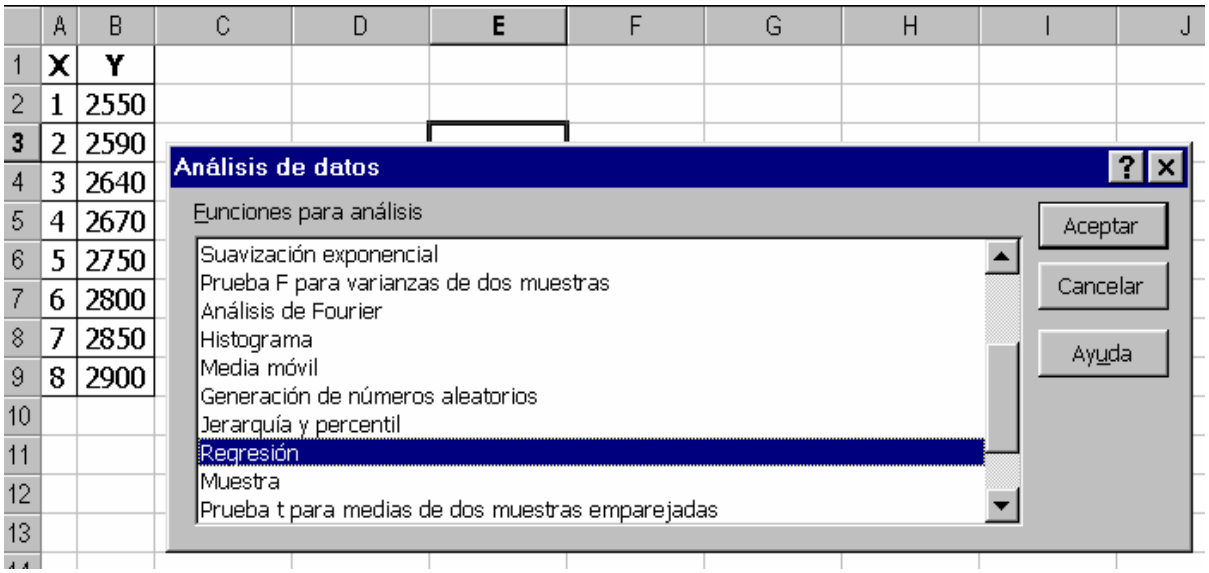
$$R^2 = \frac{[\sum_i (y_i - \bar{y})(\hat{y}_i - \bar{y})]^2}{\sum_i (y_i - \bar{y})^2 \sum_i (\hat{y}_i - \bar{y})^2}.$$

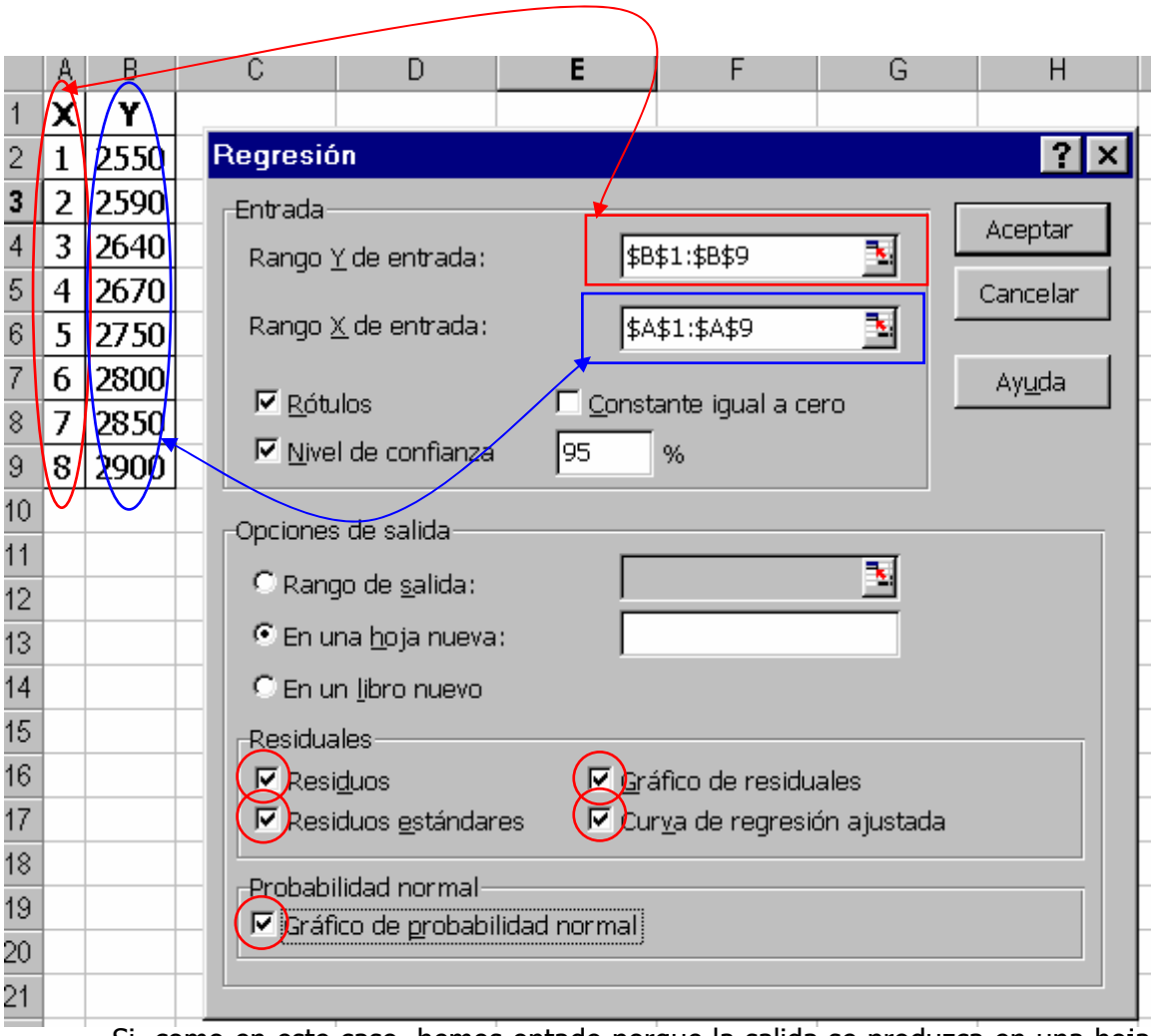
H10^2	12226161103	Numerador
D10*G10	12305272680	Denominador
D19/D20	**0,9936**	**Coeficiente de Determinación**

Una vez realizados estos cálculos, cuyo único será permitir la comprobación de los resultados que obtendremos a continuación, invocamos la opción de **Análisis de Datos**. Especificamos los rangos, tanto de la variable dependiente como de la inde-

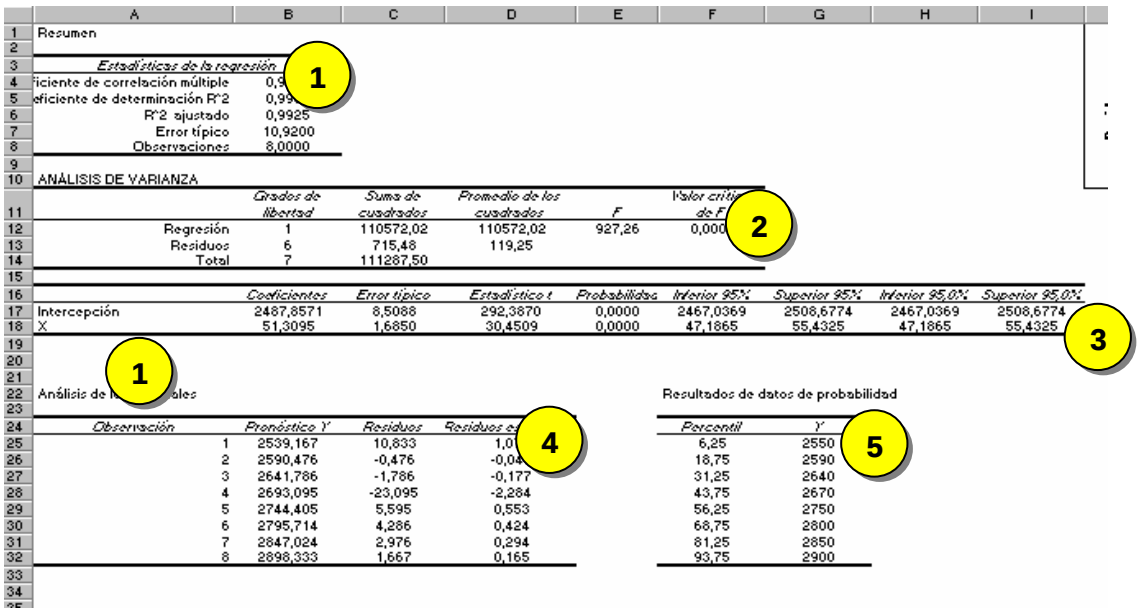
pendiente, marcando la opción *Rótulos* si éstos incluyen los nombres de las variables, e indicando el resto de las opciones deseadas:







Si, como en este caso, hemos optado porque la salida se produzca en una hoja nueva, ésta tendrá la forma siguiente:



1

	A	B
1	Resumen	
2		
3	*Estadísticas de la regresión*	
4	Coeficiente de correlación múltiple	0,9968
5	Coeficiente de determinación R^2	0,9936
6	R^2 ajustado	0,9925
7	Error típico	10,9200
8	Observaciones	8,0000
9		

Como vemos, el coeficiente de determinación coincide en su valor, con el que hemos obtenido previamente al hacer los cálculos directamente (no así con el valor 0,98858 que aparece en el material). También coincidirán los valores de los parámetros del modelo:

3

		Coeficientes	*Error típico*	*Estadístico t*	*Probabilidad*	*Inferior 95%*	*Superior 95%*	*Inferior 95,0%*	*Superior 95,0%*
16									
17	Intercepción	2487,8571	8,509	292,39	0,0000	2467,037	2508,6774	2467,0369	2508,6774
18	X	51,3095	1,685	30,45	0,0000	47,1865	55,4325	47,1865	55,4325

Es decir el modelo calculado es:

$$\hat{y}_i = 51{,}31\, x_i + 2487{,}86$$

el error estándar de β_1 es:

$$S_{\beta_1} = \frac{\sqrt{\left(\frac{1}{n-2}\right)\sum_1^n (y_i - \beta_o - \beta_1 \cdot x_i)^2}}{\sqrt{\sum_1^n (x_i - \bar{x})^2}} = \frac{10{,}92}{\sqrt{42}} = 1{,}685$$

Para calcular un intervalo de confianza del 95%, tomamos nuestra estimación de la pendiente, que era 51.31, como punto medio y calculamos el margen de error usando el error estándar y el valor crítico apropiado de la distribución t, con n -2 = 6 grados de libertad, y un nivel de significación del 5% $t_{(0.025,6)} = \pm 2{,}4469$, según vemos en las tablas de la **t** de Student.

$$51{,}31 \mp t(0{,}025{,}6) \cdot 1{,}685 = 51{,}31 \mp 4{,}123 = (47{,}187\ ;\ 55{,}433)$$

2.6. El contraste de hipótesis sobre la pendiente

Si la pendiente es cero, y es una constante y no hay una relación lineal entre y y x, la hipótesis nula natural que se contrastará es así: $H_0 : \beta_1 = 0$; en otras palabras, la hipótesis es que dentro de la población no hay ninguna relación entre la media y x, y la pendiente que se obtiene se debe a la variación aleatoria. La hipótesis alternativa puede ser unilateral o bilateral según el problema que se trata. Considerad una vez más el ejemplo del alza en el precio de la acción bursátil. Aunque es obvio que el alza será muy significativa, llevemos a cabo su prueba formal:

1) Las hipótesis nula y alternativa son: $H_0 : \beta_1 = 0$; $H_1 : \beta_1 > 0$

3

		Coeficientes	Error típico	Estadístico t	Probabilidad	Inferior 95%	Superior 95%	Inferior 95,0%	Superior 95,0%
16									
17	Intercepción	2487,8571	8,509	292,39	0,0000	2467,037	2508,6774	2467,0369	2508,6774
18	X	51,3095	1,685	30,45	0,0000	47,1865	55,4325	47,1865	55,4325

2) El estadístico de contraste es la pendiente estimada $\hat{\beta}_1$ estándar de este pendiente 49,52/2,18 = 22,7

$$t = \frac{51,31}{1,685} = 30,45$$

3) La distribución usada es la t con 6 grados de libertad y únicamente nos interesa la probabilidad de una de las colas.

4) En las tablas encontramos que el punto crítico por $\alpha = 0$ y 6 grados de libertad es 2,4469. Vemos que 22,7 supera ampliamente este punto crítico.

5) La regresión es altamente significativa y el precio de la acción bursátil muestra un incremento significativamente positivo.

El resultado siguiente es un listado de un programa de ordenador (MacAnova) obtenido con los datos del ejemplo.

Finalmente encontraremos muy útil la representación gráfica tanto del modelo construido como de los residuos de éste.

4

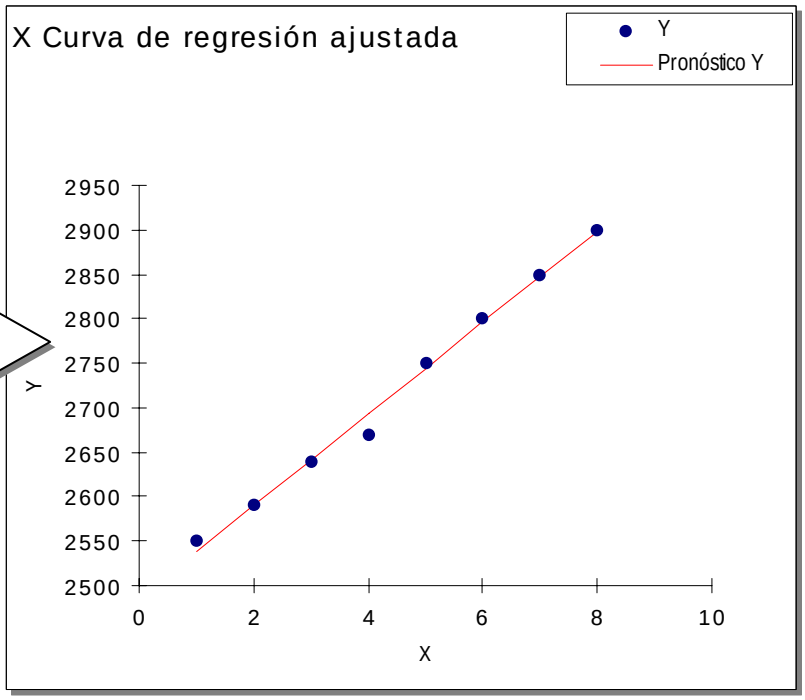
Análisis de los residuales

Observación	Pronóstico Y	Residuos	Residuos estándares
1	2539,167	10,833	1,072
2	2590,476	-0,476	-0,047
3	2641,786	-1,786	-0,177
4	2693,095	-23,095	-2,284
5	2744,405	5,595	0,553
6	2795,714	4,286	0,424
7	2847,024	2,976	0,294
8	2898,333	1,667	0,165

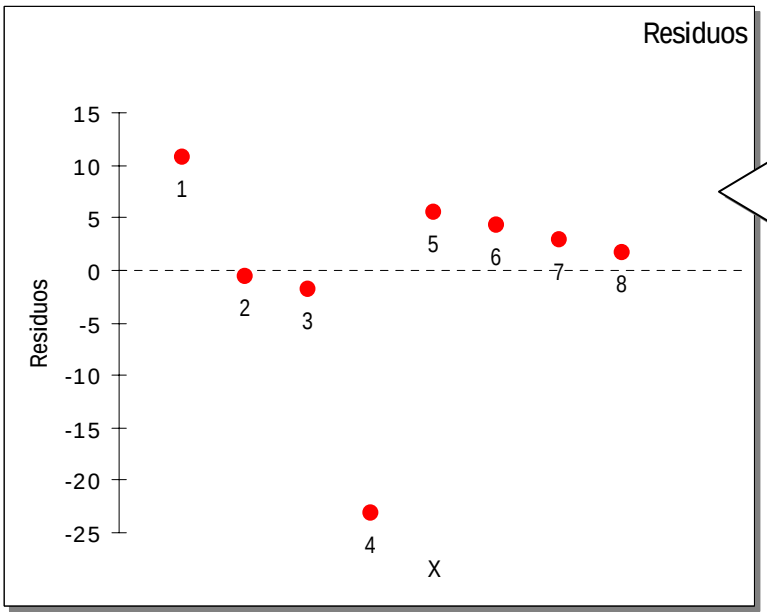
5

Resultados de datos de probabilidad

Percentil	Y
6,25	2550
18,75	2590
31,25	2640
43,75	2670
56,25	2750
68,75	2800
81,25	2850
93,75	2900



El ajuste (R^2=99%) es bastante aceptable



Los residuos no presentan un patrón claramente definido: el modelo lineal parece apropiado

9 Análisis de varianza

9.1 Resumen de los procedimientos

9.1.1 ANOVA unidireccional con muestras independientes

a) Introducimos los datos en celdas contiguas añadiendo los rótulos de los factores.
b) Elegimos Herramientas + Análisis de Datos.
c) Elegimos Análisis de varianza de un factor.
d) Elegimos como Rango de entrada el que contiene tanto a los datos (normalmente organizados en columnas) como a los rótulos (señalaremos también esta opción) y modificamos, en su caso, el valor de alfa.
e) Obtenemos los resultados en la forma descrita en el material de la asignatura.
f) Podemos realizar la prueba de Levene sin más que, en la misma hoja, realizar un nuevo análisis sobre las diferencias en valor absoluto respecto a las medias por factor.

9.1.2 ANOVA factorial con muestras independientes.

a) Introducimos los datos en celdas contiguas añadiendo los rótulos de los factores y de los grupos.
b) Elegimos Herramientas + Análisis de Datos. Elegimos Análisis de varianza de dos factores con varias muestras por grupo.
c) Rango de entrada el que contiene tanto a los datos (normalmente organizados en columnas) como a los rótulos de factores y grupos (señalamos también esta opción) y modificamos, en su caso, el valor de alfa.
d) Obtenemos los resultados en la forma descrita en el material de la asignatura, tanto en lo referente a la suma de cuadrados.
e) .... como en lo referente a los grados de libertad .....
f) .... como a las medias cuadráticas y los valores de F.

9.1.3 ANOVA unidireccional con muestras emparejadas.

a) Introducimos los datos en celdas contiguas añadiendo los rótulos de los factores y de los grupos.
b) Elegimos Herramientas + Análisis de Datos. Elegimos Análisis de varianza de dos factores con una sola muestra por grupo.
c) Elegimos como Rango de entrada el que contiene tanto a los datos (normalmente organizados en columnas) como a los rótulos (señalaremos también esta opción) y modificamos, en su caso, el valor de alfa. Obtenemos los resultados en la forma descrita en el material de la asignatura, tanto en lo referente a la suma de cuadrados.
d) .... como en lo referente a los grados de libertad .....
e) .... como a las medias cuadráticas y los valores de F.

9.2 ANOVA unidireccional con muestras independientes.

Realizaremos el siguiente ejemplo

Tabla 1: ejemplo

	A1 Antidepresivo	A2 Psicoterapia	A3 Sin tratamiento
	4	6	1
	7	8	–2
	4	5	0
	4	7	2
	6	9	–1
Σ	25	35	0

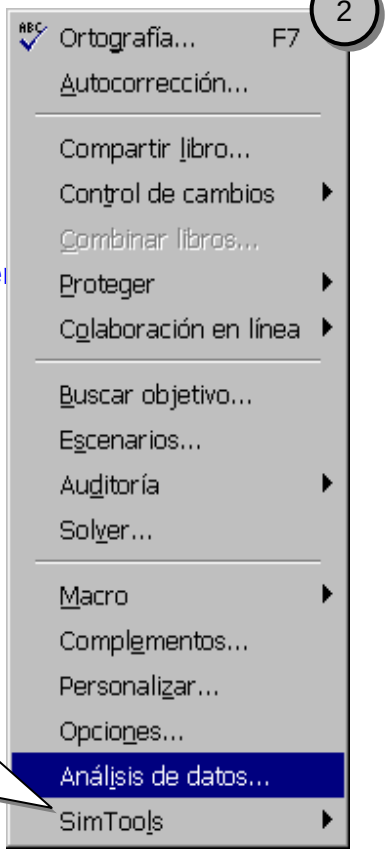
El primer paso consistirá en introducir los datos en la hoja de cálculo añadiendo los rótulos que permiten identificar los factores a analizar:

1

A1 Antidepresivo	A2 Psicoterapia	A3 Sin tratamiento
4	6	1
7	8	-2
4	5	0
4	7	2
6	9	-1

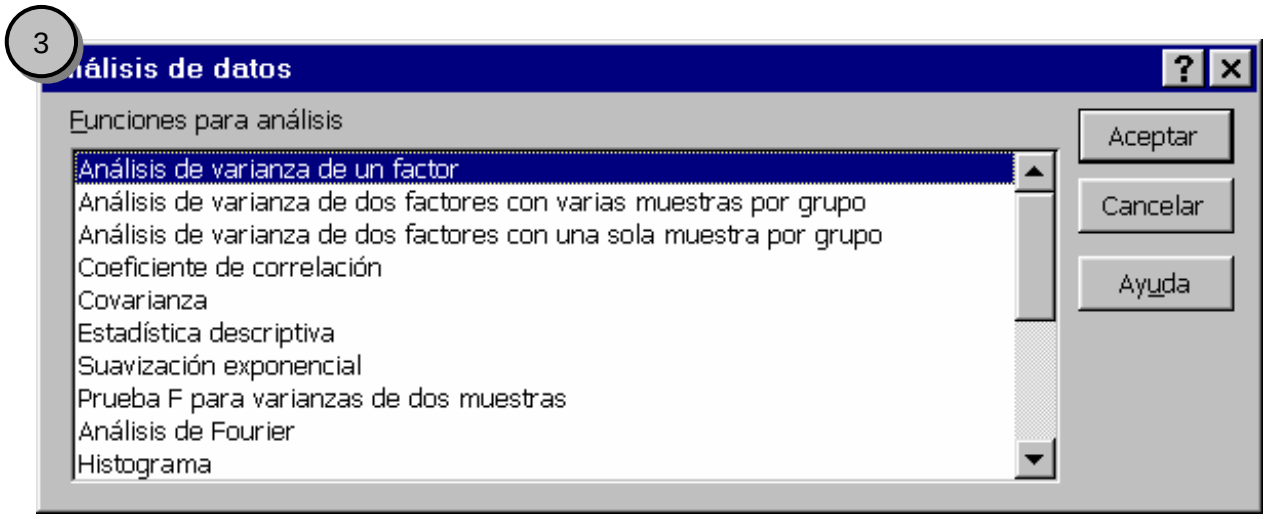
a continuación elegimos **Análisis de Datos**, del menú **He**

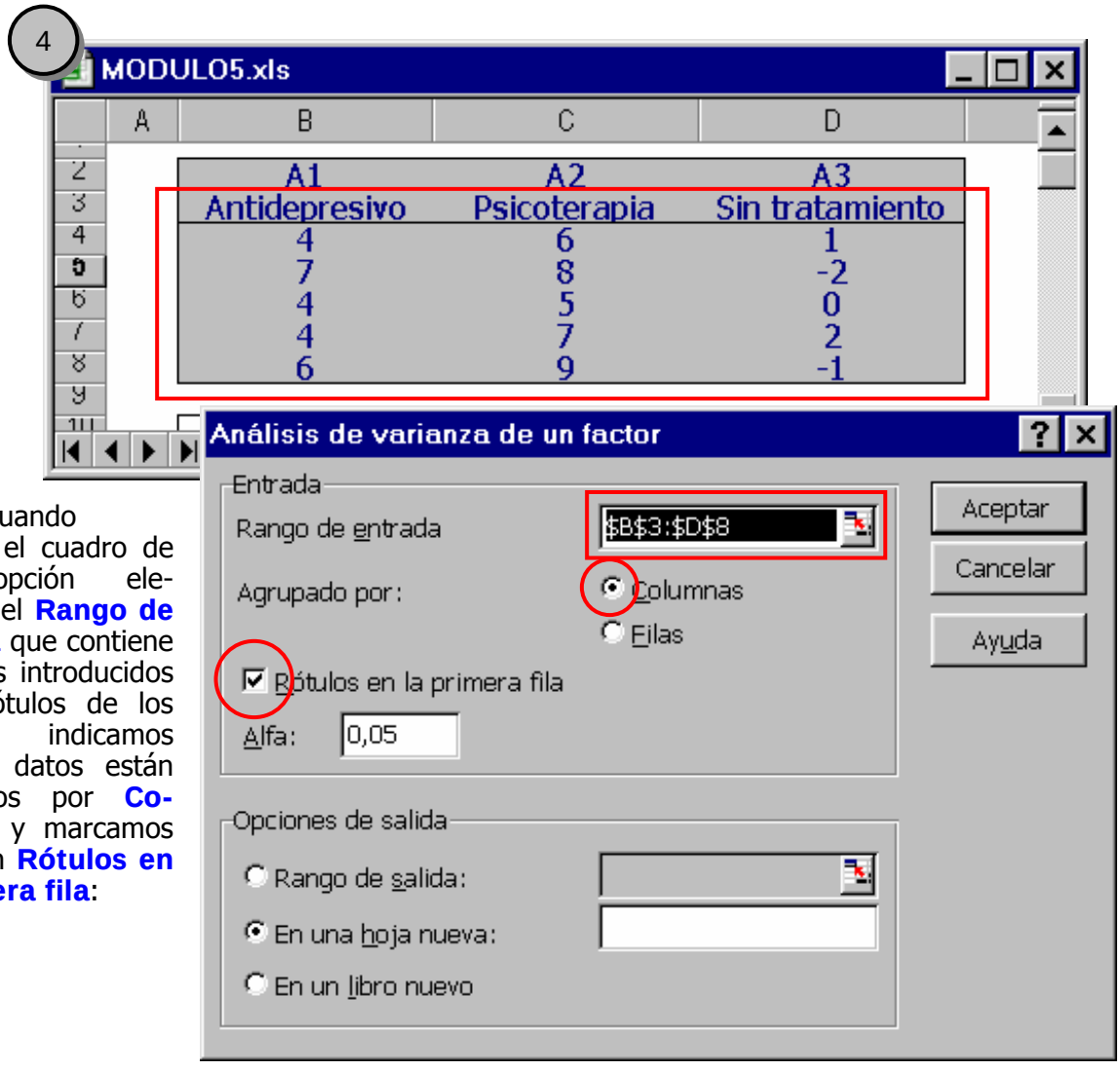
2



NOTA: SimTools es un "add-in" que ha cargado este usuario, pero no aparecerá en tu ordenador a menos que también hayas decido "bajártelo" de la WWW e instalarlo en tu ordenador.

En las Funciones para análisis elegimos la opción **Análisis de varianza de un factor**:





Cuando aparece el cuadro de esta opción elegiremos el **Rango de entrada** que contiene los datos introducidos y los rótulos de los factores; indicamos que los datos están agrupados por **Columnas** y marcamos la opción **Rótulos en la primera fila**:

El resultado será como el siguiente (aquí se presenta ligeramente modificado respecto del formato original con que lo hace Excel).

5

Análisis de varianza de un factor

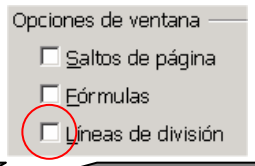
RESUMEN

Grupos	Cuenta	Suma	Promedio	Varianza
Antidepresivo	5	25	5	2
Psicoterapia	5	35	7	2,5
Sin tratamiento	5	0	0	2,5

ANÁLISIS DE VARIANZA

Origen de las variaciones	Suma de cuadrados	Grados de libertad	Promedio de los cuadrado	F	Probabilidad	Valor crítico para F
Entre grupos	130	2	65	27,8571	0,0000	3,8853
Dentro de los grupos	28	12	2,333			
Total	158	14				

NOTA: El usuario de este ordenador ha deshabilitado la opción Líneas de división, ese es el motivo por el cual no aparecen las características líneas de Excel"



Aunque la presentación del resumen es ligeramente diferente respecto a la descrita en el material de la asignatura:

1.3. Tabla resumen del análisis de la varianza

6

FV	SC	g. l.	MC	*F*	Significación
Entre grupos o tratamientos	$\Sigma_j \frac{(\Sigma_i Y_{ij})^2}{n_j} - \frac{(\Sigma_i \Sigma_j Y_{ij})^2}{N}$	$k-1$	$\frac{SC_{entre}}{k-1}$	$\frac{MC_{entre}}{MC_{intra}}$	
Intragrupos o error	$\Sigma_i \Sigma_j Y_{ij}^2 - \Sigma_j \frac{(\Sigma_i Y_{ij})^2}{n_j}$	$N-k$	$\frac{SC_{intra}}{N-k}$		
Total	$\Sigma_i \Sigma_j Y_{ij}^2 - \frac{(\Sigma_i \Sigma_j Y_{ij})^2}{N}$	$N-1$			

Los resultados, por supuesto, son los mismos

En nuestro ejemplo:

FV	SC	g. l.	MC	*F*	Significación
Entre grupos o tratamiento	130	2	65	27,857	$p < 0{,}0001$
Intragrupos o error	28	12	2,333		
Total	158	14			

Análisis de varianza de un factor

RESUMEN

Grupos	Cuenta	Suma	Promedio	Varianza
Antidepresivo	5	25	5	2
Psicoterapia	5	35	7	2,5
Sin tratamiento	5	0	0	2,5

ANÁLISIS DE VARIANZA

Origen de las variaciones	Suma de cuadrados	Grados de libertad	Promedio de los cuadrado	F	Probabilidad	Valor crítico para F
Entre grupos	130	2	65	27,8571	0,0000	3,8853
Dentro de los grupos	28	12	2,333			
Total	158	14				

Realizar la prueba de Levene exigirá que hagamos unos sencillos cálculo previos:

1.4.1. Prueba de Levene

La prueba de Levene es relativamente sencilla en su aplicación. Consiste en calcular para cada observación su diferencia (en valor absoluto) con respecto a la media de su grupo. A partir de estas puntuaciones de diferencia se hace un ANOVA unidireccional que permite obtener la probabilidad de cumplimiento de la homogeneidad de las varianzas.

En la misma hoja podemos:

1. copiar la tabla de los datos originales;
2. calcular las medias de las puntuaciones de cada factor de la tabla original;
3. construir la nueva tabla restando de las puntuaciones la medias recién calculadas;
4. aplicar los pasos anteriores para hacer un análisis sobre estos nuevos datos.

Antidepresivo	Psicoterapia	Sin tratamiento
5	7	0
-1	-1	1
2	1	-2
-1	-2	0
-1	0	2
1	2	-1

D11= PROMEDIO(D4:D8)

=D6-D$11

El resultado final podría ser como el siguiente:

A1 Antidepresivo	A2 Psicoterapia	A3 Sin tratamiento
4	6	1
7	8	-2
4	5	0
4	7	2
6	9	-1

Antidepresivo	Psicoterapia	Sin tratamiento
5	7	0
-1	-1	1
2	1	-2
-1	-2	0
-1	0	2
1	2	-1

Análisis de varianza de un factor (LEVENE)

RESUMEN

Grupos	Cuenta	Suma	Promedio	Varianza
Antidepresivo	5	0	0	2
Psicoterapia	5	0	0	2,5
Sin tratamiento	5	0	0	2,5

ANÁLISIS DE VARIANZA

Origen de las variaciones	Suma de cuadrados	Grados de libertad	Promedio de los cuadrados	F	Probabilidad	Valor crítico para F
Entre grupos	0	2	0,0000	0,0000	1,0000	3,8853
Dentro de los grupos	28	12	2,3333			
Total	28	14				

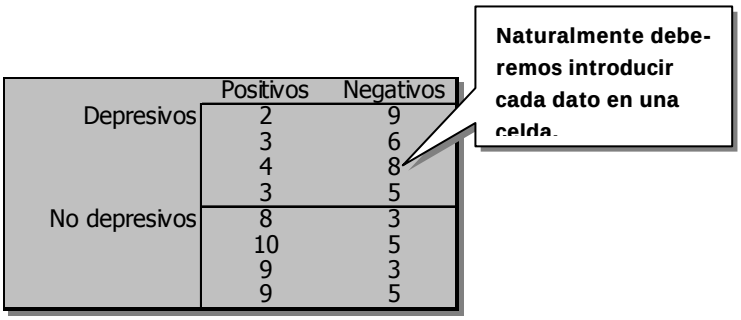
9.3 ANOVA factorial con muestras independientes.

Realizaremos el siguiente ejemplo:

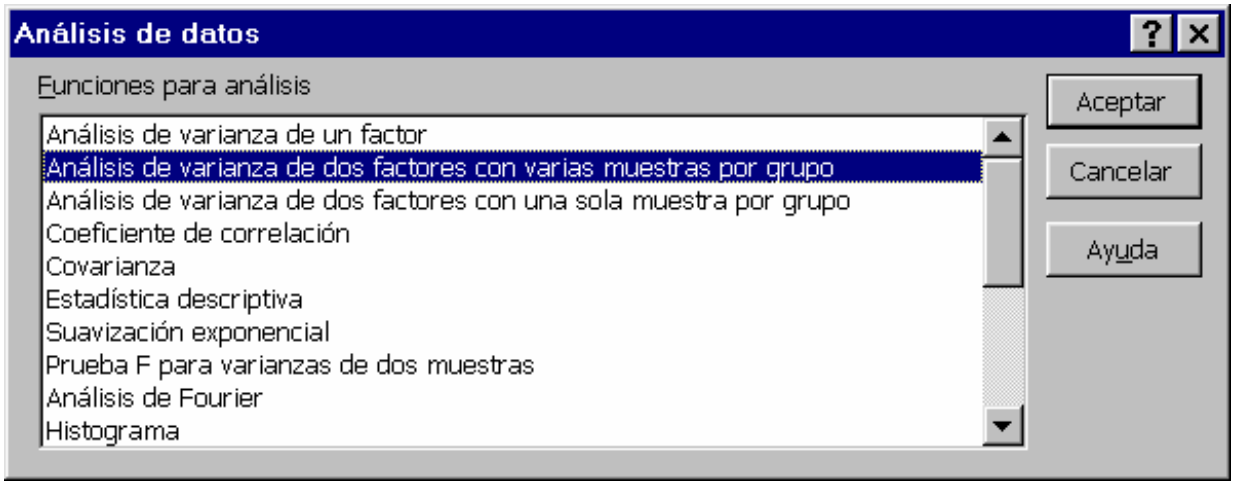
Tabla 2. Número de adjetivos recordados según la condición de estado de ánimo y tipo de adjetivo. Se han calculado los totales (T) en cada fila, columna, casilla y el total general.

	Adjetivos		
	Positivos	Negativos	
Depresivos	2 3 4 3 T = 12	9 6 8 5 T = 28	T = 40
No depresivos	8 10 9 9 T = 36	3 5 3 5 T = 16	T = 52
	T = 48	T = 44	T = 92

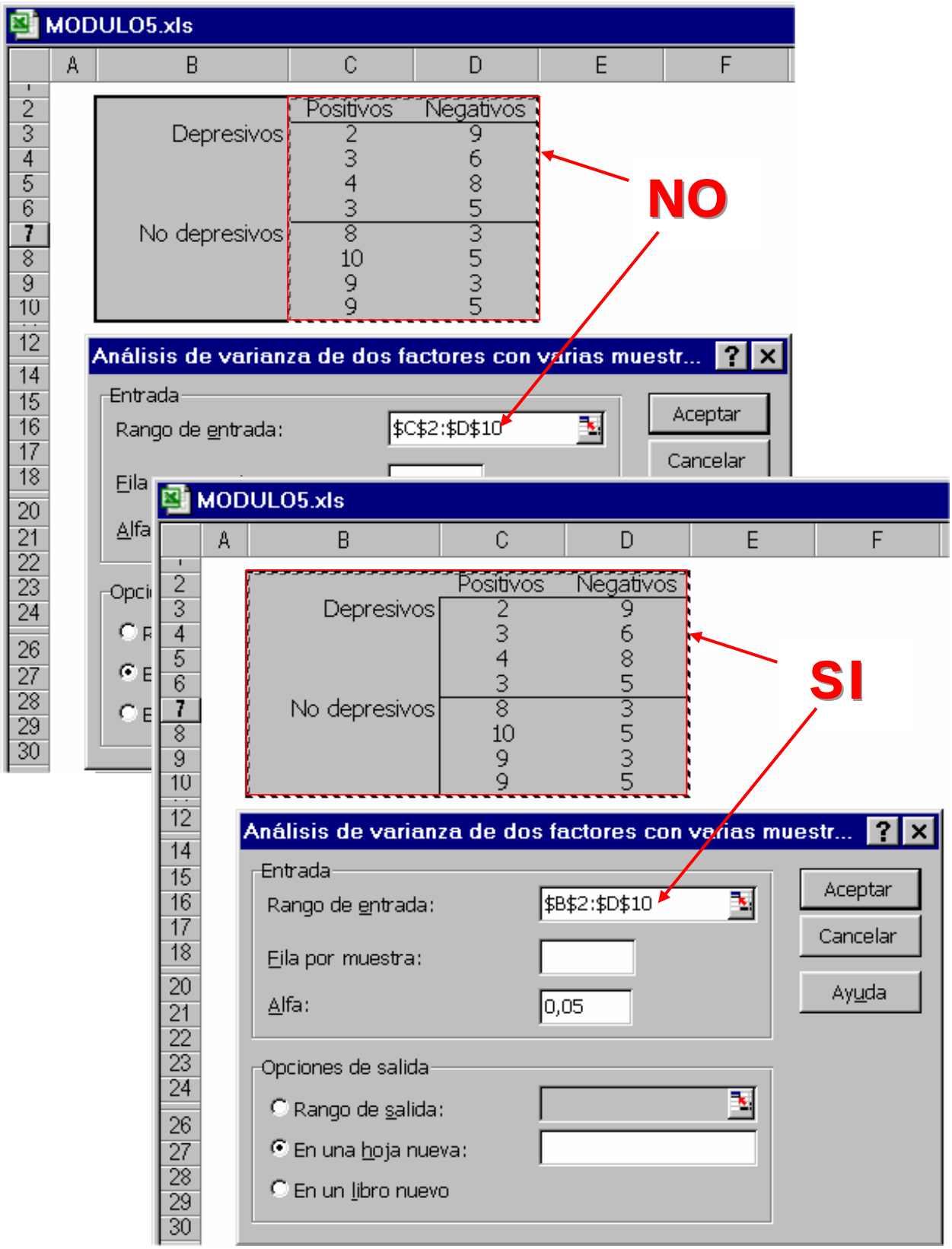
Como siempre, el primer paso consiste en introducir datos y rótulos en la hoja de cálculo:



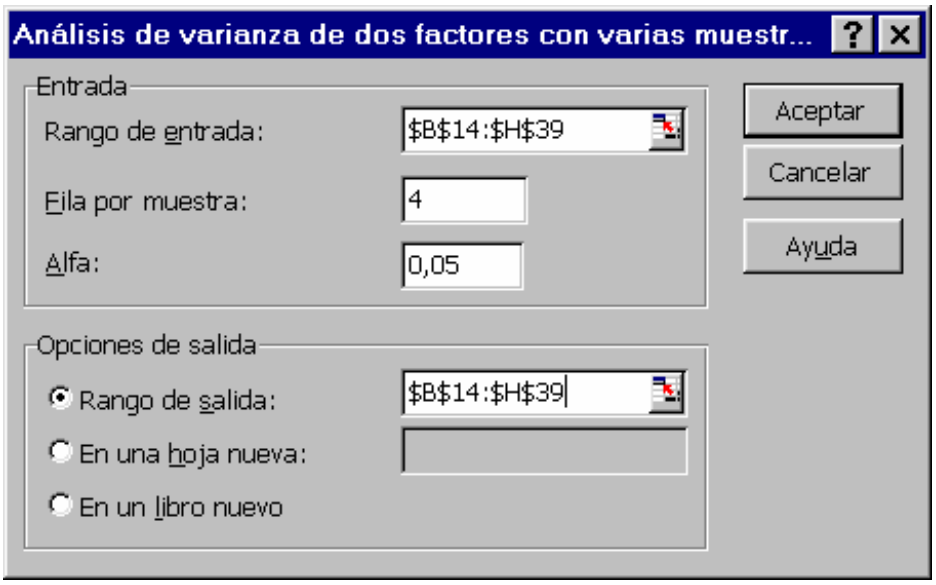
Abrimos de nuevo el menú **Herramientas** + **Análisis de Datos** y elegimos ahora la opción **Análisis de varianza de dos factores con varias muestras por grupo**.



A la hora de rellenar los campos de esta opción deberemos tener cuidado en elegir bien el rango de entrada. Éste deberá incluir los rótulos tanto de los factores, como de las muestras.



Deberemos indicar el número de filas que ocupan las muestras (4 en nuestro caso) y con el fin de mantener datos y resultados en la misma hoja señalar como rango de salida una porción desocupada de la hoja de cálculo en la que se volcará los resultados.



El resultado final aparecerá como el siguiente (aquí ligeramente modificado en su formato de presentación):

	Positivos	Negativos
Depresivos	2	9
	3	6
	4	8
	3	5
No depresivos	8	3
	10	5
	9	3
	9	5

Análisis de varianza de dos factores con varias muestras por grupo

Depresivos			
Cuenta	4	4	8
Suma	12	28	40
Promedio	3	7	5
Varianza	0,667	3,333	6,286

No depresivos			
Cuenta	4	4	8
Suma	36	16	52
Promedio	9	4	6,5
Varianza	0,667	1,333	8,000

Total		
Cuenta	8	8
Suma	48	44
Promedio	6	5,5
Varianza	10,8571	4,5714

ANÁLISIS DE VARIANZA

Origen de las variaciones	Suma de cuadrados	Grados de libertad	Promedio de los cuadrados	F	Probabilidad	Valor crítico para F
Muestra	**9**	1	9	6,000	0,031	4,747
Columnas	**1**	1	1	0,667	0,430	4,747
Interacción	81	1	81	54,000	0,000	4,747
Dentro del grupo	**18**	12	1,5			
Total	**109**	15				

dos c

ANÁLISIS DE VARIANZA						
Origen de las variaciones	Suma de cuadrados	Grados de libertad	Promedio de los cuadrados	F	Probabilidad	Valor crítico para F
Muestra	**9**	1	9	6,000	0,031	4,747
Columnas	**1**	1	1	0,667	0,430	4,747
Interacción	81	1	81	54,000	0,000	4,747
Dentro del grupo	**18**	12	1,5			
Total	**109**	15				

$$SC_{Total} = \Sigma_i\Sigma_j\Sigma_k Y_{ijk}^2 - \frac{(\Sigma_i\Sigma_j\Sigma_k Y_{ijk})^2}{N} = 2^2 + 3^2 + 4^2 + \ldots + 5^2 + 3^2 + 5^2 - \frac{92^2}{16} = 109$$

$$SC_{Depresión} = \Sigma_j \frac{(\Sigma_i\Sigma_k Y_{ijk})^2}{n_j} - \frac{(\Sigma_i\Sigma_j\Sigma_k Y_{ijk})^2}{N} = \frac{40^2}{8} + \frac{52^2}{8} - \frac{92^2}{16} = 9,$$

$$SC_{Adjectivo} = \Sigma_k \frac{(\Sigma_i\Sigma_j Y_{ijk})^2}{n_k} - \frac{(\Sigma_i\Sigma_j\Sigma_k Y_{ijk})^2}{N} = \frac{48^2}{8} + \frac{44^2}{8} - \frac{92^2}{16} = 1,$$

$$SC_{error} = \Sigma_i\Sigma_j\Sigma_k Y_{ijk}^2 - \left(\Sigma_j\Sigma_k \frac{\Sigma_i Y_{ijk}^2}{n_{jk}}\right) =$$

$$= 2^2 + 3^2 + 4^2 + \ldots + 5^2 + 3^2 + 5^2 - \left[\frac{12^2}{4} + \frac{28^2}{4} + \frac{36^2}{4} + \frac{16^2}{4}\right] = 18,$$

$$SC_{Depres.x.Adjet} = SC_{Total} - \lfloor SC_{Depresión} + SC_{Adjectivo} + SC_{error} \rfloor =$$

$$= 109 - [9 + 1 + 18] = 81$$

....en las **medias cuadráticas** y en los **grados de libertad**.....

$$MC_{Total} = \frac{SC_{Total}}{N-1} = \frac{109}{15} = 7{,}267$$

$$MC_A = \frac{SC_A}{a-1} = \frac{9}{1} = 9$$

$$MC_B = \frac{SC_B}{b-1} = \frac{1}{1} = 1$$

$$MC_{A\times B} = \frac{SC_{A\times B}}{(a-1)(b-1)} = \frac{81}{1} = 81$$

$$MC_{error} = \frac{SC_{error}}{N-a\times b} = \frac{18}{12} = 1{,}5$$

Suma de cuadrados	Grados de libertad	Promedio de los cuadrados
9	1	**9**
1	1	**1**
81	1	**81**
18	12	**1,5**
109	15	

Así, en el ejemplo tendremos lo siguiente:

Fuentes de variación	Grados de libertad
Depresión	$a-1 \rightarrow 1$
Tipo de adjetivo	$b-1 \rightarrow 1$
Interacción Depres. × Adjet.	$(a-1)(b-1) \rightarrow 1$
Error	$N-(a\times b) \rightarrow 12$
Total	$N-1 \rightarrow 15$

....en los valores de la **F de Snedecor**...

ANÁLISIS DE VARIANZA

Origen de las variaciones	Suma de cuadrados	Grados de libertad	Promedio de los cuadrados	F	Probabilidad	Valor crítico para F
Muestra	9	1	9	6,000	0,031	4,747
Columnas	1	1	1	0,667	0,430	4,747
Interacción	81	1	81	54,000	0,000	4,747
Dentro del grupo	18	12	1,5			
Total	109	15				

$$F_{variable.A} = \frac{MC_A}{MC_{Error}} = \frac{9}{1,5} = 6$$

$$F_{variable.B} = \frac{MC_B}{MC_{Error}} = \frac{1}{1,5} = 0,667$$

$$F_{Interaccion\ A\times B} = \frac{MC_{Interaccion.A\times B}}{MC_{Error}} = \frac{81}{1,5} = 54.$$

....como en las **probabilidades** asociadas a la hipótesis nula.

	F	Significación
	6	0,031
	0,667	0,430
	54	0,000

ANÁLISIS DE VARIANZA

Origen de las variaciones	Suma de cuadrados	Grados de libertad	Promedio de los cuadrados	F	Probabilidad	Valor crítico para F
Muestra	9	1	9	6,000	0,031	4,747
Columnas	1	1	1	0,667	0,430	4,747
Interacción	81	1	81	54,000	0,000	4,747
Dentro del grupo	18	12	1,5			
Total	109	15				

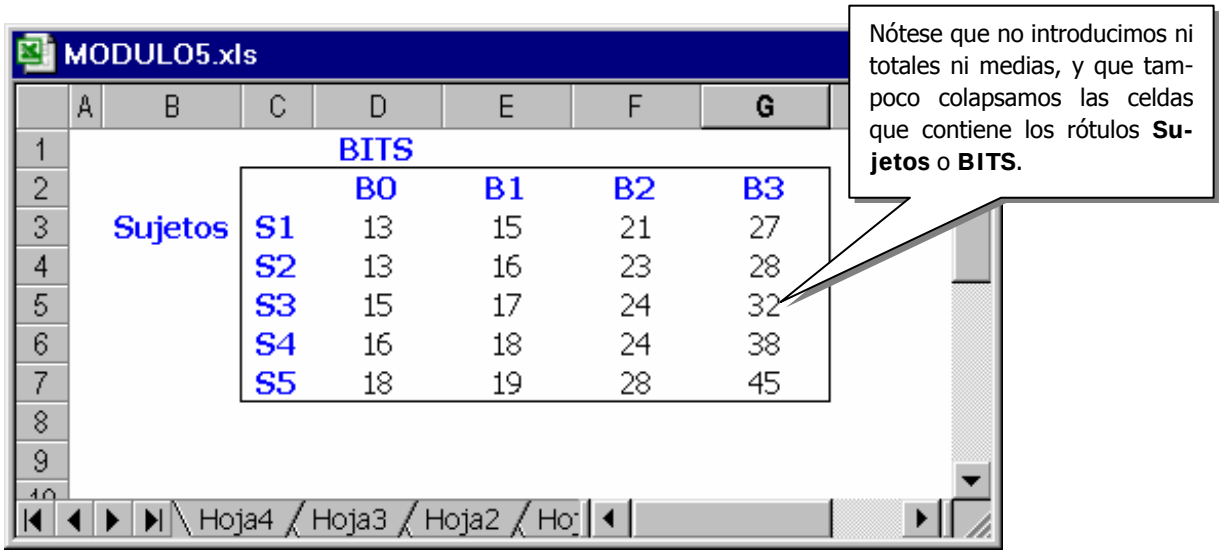
9.4 ANOVA unidireccional con muestras emparejadas.

Realizaremos el siguiente ejemplo

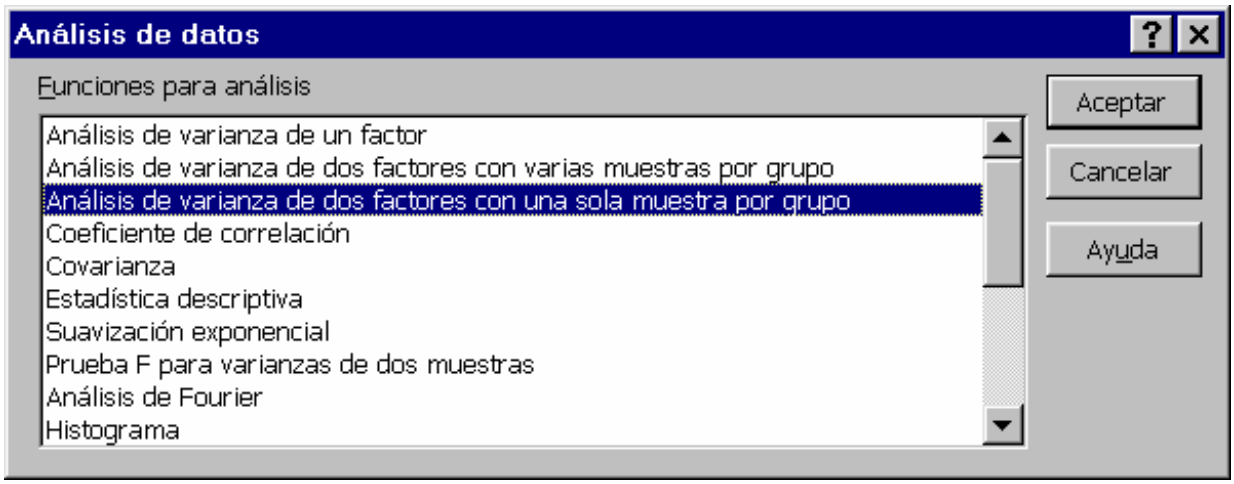
Tabla 3: TR (en décimas de segundo) de los cinco sujetos en cada uno de los niveles de la variable independiente (de 0 a 3 bits de información).

		Bits					
		0	1	2	3	Totales	Medias
Sujeto	1	13	15	21	27	76	19
"	2	13	16	23	28	80	20
"	3	15	17	24	32	88	22
"	4	16	18	24	38	96	24
"	5	18	19	28	45	110	27,5
Totales		75	85	120	170	450	
Medias		15	17	24	34		$\bar{Y}_{..}$ = 22,5

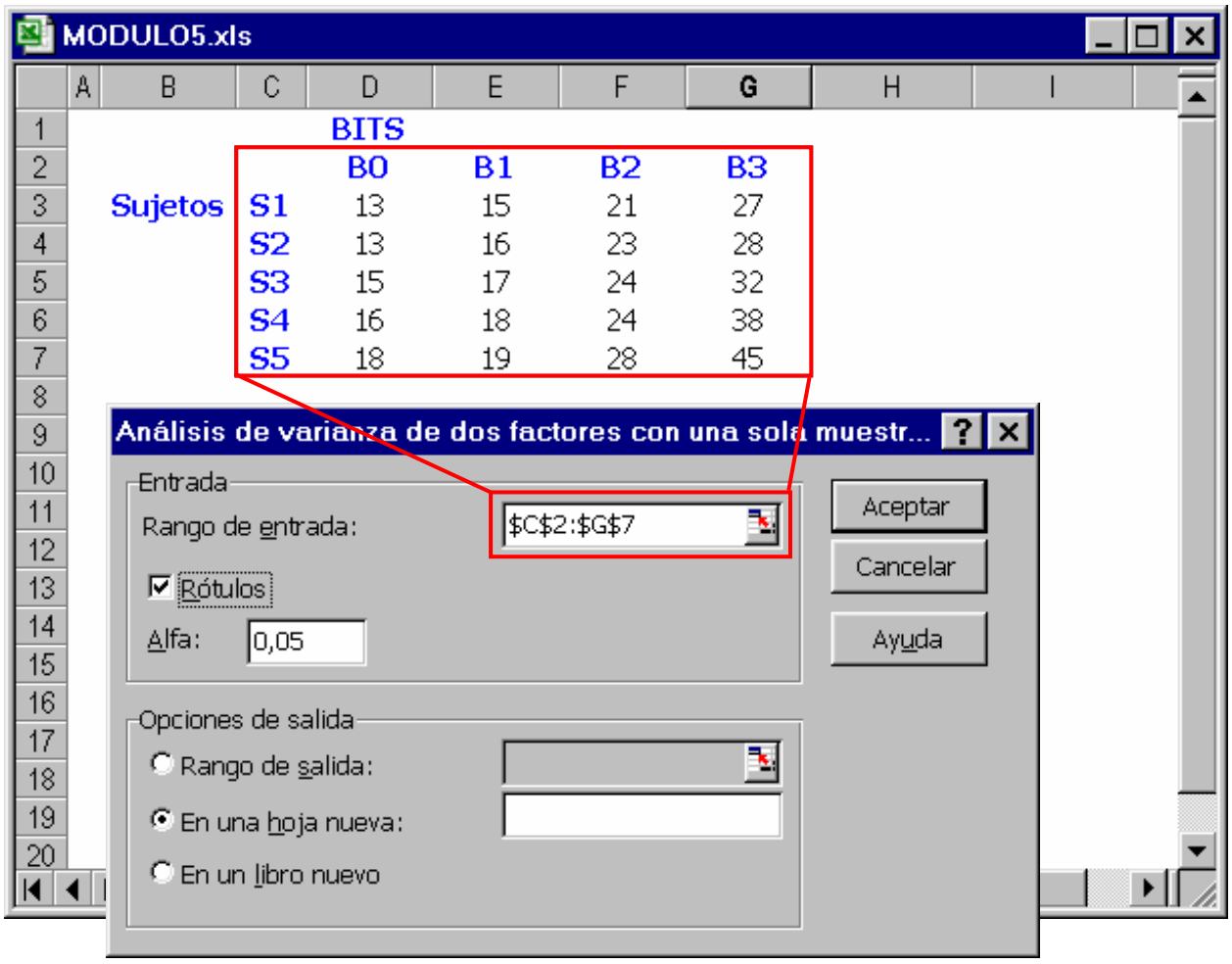
De nuevo comenzamos introduciendo datos y rótulos en la hoja de cálculo:



En **Herramientas**, **Análisis de Datos**, elegimos ahora la opción **Análisis de varianza de dos factores con una sola muestra por grupo**.



El **rango de entrada** contiene los rótulos de los niveles de cada factor (S_1,.. y B_0,..), pero no los de los factores (Sujetos y BITS).



De nuevo comprobamos que los resultados descritos en el material y los obtenidos por Excel coinciden plenamente.

$$SC_{Total} = \Sigma_i \Sigma_j Y_{ij} - \frac{(\Sigma_i \Sigma_j Y_{ij})^2}{N} =$$

$$= 13^2 + 13^2 + 15^2 + \ldots + 32^2 + 38^2 + 45^2 - \frac{450^2}{20} = 1385,$$

$$SC_{Tractament} = \Sigma_j \frac{(\Sigma_i Y_{ij})^2}{n} - \frac{(\Sigma_i \Sigma_j Y_{jk})^2}{N} =$$

$$= \frac{75^2}{5} + \frac{85^2}{5} + \frac{120^2}{5} + \frac{170^2}{5} - \frac{450^2}{20} = 1105$$

$$SC_{Subjectes} = \Sigma_k \frac{(\Sigma_j Y_{ij})^2}{k} - \frac{(\Sigma_i \Sigma_j Y_{jk})^2}{N} =$$

$$= \frac{76^2}{4} + \frac{80^2}{4} + \frac{88^2}{4} + \frac{96^2}{4} + \frac{110^2}{4} - \frac{450^2}{20} = 184$$

$$SC_{Error} = SC_{Total} - SC_{Tratamiento} + SC_{Sujetos} =$$

$$= 1385 - [1105 + 184] = 96$$

ANÁLISIS DE VARIANZA

Origen de las variaciones	Suma de cuadrados	Grados de libertad	Promedio de los cuadrados	F	Probabilidad	Valor crítico para F
Filas	**184**	4	46,0000	5,7500	0,0080	3,2592
Columnas	**1105**	3	368,3333	46,0417	0,0000	3,4903
Error	**96**	12	8,0000			
Total	**1385**	19				

Los grados de libertad son los siguientes:

Fuente de variación	Grados de libertad
Total	$N - 1 \rightarrow 19$
Tratamientos	$k - 1 \rightarrow 3$
Sujetos	$n - 1 \rightarrow 4$
Error	$(k - 1)(n - 1)$ 12

ANÁLISIS DE VARIANZA

Origen de las variaciones	Suma de cuadrados	Grados de libertad	Promedio de los cuadrados	F	Probabilidad	Valor crítico para F
Filas	184	4	46,0000	5,7500	0,0080	3,2592
Columnas	1105	3	368,3333	46,0417	0,0000	3,4903
Error	96	12	8,0000			
Total	1385	19				

10 Tablas de contingencia

10.1 Distribución de frecuencias observadas.

El único aspecto cuantificable en el análisis cualitativo es el número de individuos que presenta una combinación los niveles de los factores. Estos valores se recogen en tablas de contingencia. (frecuencias observadas de cada combinación).

Factores	Nivel 1º factor B	Nivel 2º factor B	$n_{i\bullet}$
factor A Nivel 1º	n_{11}	n_{12}	n_{1j}
factor A Nivel 2º	n_{21}	n_{22}	n_{2j}
$n_{\bullet j}$	$n_{\bullet 1}$	$n_{\bullet 2}$	$n = \sum\sum n_{ij}$

Los n_{ij} representan el número de individuos observados en cada combinación de los niveles de los factores A, B y se consideran como la realización de una v.a. con valores enteros y positivos. Nuestro objetivo principal es contrastar la independencia entre los factores en estudio.

Consideremos una tabla de contingencia IxJ y sea P_{ij} la probabilidad poblacional de que un individuo sea elegido en la casilla (i, j).

La hipótesis de independencia entre factores es:

$$P_{ij} = P_{i\bullet}P_{\bullet j} \quad \Leftrightarrow \quad \hat{m}_{ij} = \frac{n_{i\bullet}\, n_{\bullet j}}{n}$$

10.2 INDEPENDENCIA EN TABLAS DE CONTINGENCIA BIDIMENSIONALES.

Contrastación de la hipótesis de independencia en una tabla de contingencia bidimensional.

Contrastes de independencia exactos.

En caso de muestras pequeñas.

1. Determinar el espacio muestral del diseño empleado en la tabla observada, es decir todas las tablas posibles manteniendo constantes los marginales.
2. Seleccionar de todas estas tablas las que se alejan tanto o más de H_0 que la tabla observada en la dirección de H_1.
3. Calcular las probabilidades de ocurrencia bajo H_0 de dichas tablas.
4. Calcular el p-valor del test. (sumar las probabilidades de dichas tablas)
5. Comparar el p-valor con el nivel de significación α prefijado.
   - Si $p>\alpha$ aceptamos H_0.
   - Si $p\leq\alpha$ rechazamos H_0.

Inconvenientes:

- El cálculo de la probabilidad exacta de las tablas puede depender de parámetros desconocidos. Se soluciona estimando éstos.
- Cuando aumenta la muestra o los niveles de los factores el cálculo del p-valor es muy laborioso.

Contrastes de independencia asintóticos.

Contraste χ^2 de independencia.

Las hipótesis a contrastar son:

$$H_0 : P_{ij} = P_{i.}P_{.j}$$
$$H_1 : P_{ij} \neq P_{i.}P_{.j}$$

El estadístico propuesto para realizar este contraste es el siguiente:

$$\chi^2 = \sum_i \sum_j \frac{\left(n_{ij} - \hat{m}_{ij}\right)^2}{\hat{m}_{ij}}$$

teniendo en cuenta que, bajo H_0 tenemos

$$\hat{m}_{ij} = \frac{n_{i.}n_{.j}}{n}$$

Dicho estadístico se distribuye según una χ^2 con (I-1)(J-1) grados de libertad. Si el valor observado supera al esperado, rechazaremos H_0.

Corrección por continuidad (Yates).

El estadístico corregido tiene la siguiente expresión:

$$\chi_C^2 = \sum_i \sum_j \frac{\left(\left|n_{ij} - \hat{m}_{ji}\right| - 1/2\right)^2}{\hat{m}_{ij}}$$

y se distribuye según una χ^2con (I-1)(J-1) grados de libertad.

Análisis de residuos.

Si en una tabla de contingencia la hipótesis de independencia se ha visto rechazada, mediante el análisis de residuos podemos detectar los niveles de los factores que pueden ser los causantes de tal asociación.

Residuos estandarizados:

$$e_{ij} = \frac{n_{ij} - \hat{m}_{ij}}{\sqrt{\hat{m}_{ij}}}$$

La varianza estimada de los residuos:

$$\hat{V}_{(e_{ij})} = \left(1 - \frac{n_{i.}}{n}\right)\left(1 - \frac{n_{.j}}{n}\right)$$

Residuos ajustados:

$$d_{ij} = \frac{e_{ij}}{\sqrt{\hat{V}_{ij}}}$$

Se consideran significativos a un nivel de significación α aquellos que en valor absoluto superen el cuantil correspondiente a una N(0,1).

10.3 MEDIDAS DE ASOCIACIÓN EN TABLAS IxJ

Cuando la hipótesis de independencia es rechazada podemos plantearnos cuál es el grado de asociación y la dirección en que se produce tal. Las medidas de asociación son parámetros poblacionales que dependen de las probabilidades poblacionales Pij.

Éstas deben ser fácilmente interpretables y deben estar acotadas de manera que los factores indiquen asociación perfecta o falta de asociación. Suelen estar normalizadas tomando valores entre 0 y 1 ó entre –1 y 1, lo cual permite la comparaciones entre tablas de diferentes tamaños.

Medidas de asociación en tablas 2x2.

Cociente de probabilidad. Se define el cociente de probabilidad como:

$$\theta = \frac{w_1}{w_2} = \frac{\frac{p_{22}}{p_{21}}}{\frac{p_{12}}{p_{11}}} = \frac{p_{11}p_{22}}{p_{12}p_{21}}$$

Propiedades:

- $\theta \in [0,\infty]$
- no definido si p_{11} o p_{22} son 0.
- Si las dos son cero hay asociación perfecta estricta positiva.
- $\theta=0\rightarrow$ cuando p_{11} y/o p_{22} son nulas.
- $\theta=1\rightarrow$ dependencia entre los factores.
- $\theta>1\rightarrow$ asociación positiva.
- $\theta<1\rightarrow$ asociación negativa.
- Invariante frente a cambios de escala en filas y/o columnas.

El estimador de θ es:

$$\hat{\theta} = \frac{n_{11}n_{22}}{n_{12}n_{21}}$$

con similar interpretación.

Q de Yule. Definido como:

$$Q = \frac{p_{11}p_{22} - p_{12}p_{21}}{p_{11}p_{22} + p_{12}p_{21}} = \frac{\theta - 1}{\theta + 1}$$

se verifica que:

$$Q = 0 \text{ independencia}$$
$$Q > 0 \text{ asocic + si } \theta > 1$$
$$Q < 0 \text{ asocic - si } \theta < 1$$
$$Q = 1 \text{ asocic perf estrc +}$$
$$Q = -1 \text{ asocic perf estrc -}$$

valor muestral:

$$\hat{Q} = \frac{n_{11}n_{22} - n_{12}n_{21}}{n_{11}n_{22} + n_{12}n_{21}}$$

Medidas de asociación en tablas IxJ.

Medida ϕ^2 de Pearson.

Valor poblacional:

$$\phi^2 = \frac{1}{n}\sum_i\sum_j \frac{(P_{ij} - P_{i.}P_{.j})^2}{P_{i.}P_{.j}}$$

Valor estimado:

$$\hat{\phi}^2 = \frac{1}{n}\sum_i\sum_j \frac{(n_{ij} - \hat{m}_{ij})^2}{m_{ij}} = \frac{\chi^2}{n}$$

- Vale 0 sii independencia.
- Asociación perfecta estricta : vale 1
- Es simétrica y fácil de calcular.

Coeficiente de contingencia.

Valor poblacional:

$$C = \sqrt{\frac{\phi^2}{\phi^2 + 1}}$$

Valor estimado:

$$C = \sqrt{\frac{\left(\frac{\chi^2}{n}\right)}{\left(\frac{\chi^2}{n}\right) + 1}}$$

- Si vale cero hay independencia.
- No alcanza su valor máximo aún cuando hay asociación perfecta. Este depende del tamaño de la tabla.
- Para tablas cuadradas el valor máximo que puede tomar es el siguiente:

$$C_{max} = \sqrt{\frac{I-1}{I}}.$$

- En la práctica se utiliza el ajustado:

$$C_A = \frac{C}{C_{max}}$$

Medida T de Tschuprov.

Valor poblacional:

$$T = \sqrt{\frac{\phi^2}{\sqrt{(I-1)(J-1)}}}$$

Valor estimado:

$$\hat{T} = \sqrt{\frac{\chi^2}{n\sqrt{(I-1)(J-1)}}}$$

- Vale 0 cuando hay independencia.
- Vale 1 en caso de asociación perfecta estricta en tablas 2x2.

V de Cramer.

Valor poblacional:

$$V = \sqrt{\frac{\phi^2}{m}} \text{ con } m = \min\{(I-1),(J-1)\}$$

Valor estimado:

$$\hat{V} = \sqrt{\frac{\chi^2}{nm}}$$

- Vale 0 sii independencia.
- En asociación perfecta alcanza su valor máximo.
- En tablas cuadradas su valor coincide con T

10.4 Funciones relacionadas

- **DISTR.CHI** devuelve el complementario a la unidad de la función de distribución para un valor de x, es decir, la probabilidad de que la variable aleatoria distribuida según una χ2GL sea mayor que x.

La descripción de esta función que figura en la ayuda de Excel es la siguiente:

Devuelve la probabilidad de una variable aleatoria continua siguiendo una distribución chi cuadrado de una sola cola. La distribución chi cuadrado está asociada con la prueba chi cuadrado. Use la prueba chi cuadrado para comparar los valores observados con los esperados. Por ejemplo, un experimento genético podría estar basado en la hipótesis de que la próxima generación de plantas presentará un conjunto determinado de colores. Al comparar los resultados observados con los resultados esperados, puede decidir si su hipótesis original es válida.

DISTR.CHI (x ; grados_de_libertad)

- X es el valor al que desea evaluar la distribución.
- grados_de_libertad es el número de grados de libertad.
- Si uno de los argumentos no es numérico, DISTR.CHI devuelve el valor de error #¡VALOR!.
- Si el argumento x es negativo, DISTR.CHI devuelve el valor de error #¡NUM!.
- Si el argumento grados_de_libertad no es un entero, se trunca.
- Si el argumento grados_de_libertad < 1 o grados_de_libertad ≥ 10^10, DISTR.CHI devuelve el valor de error #¡NUM!
- DISTR.CHI se calcula como DISTR.CHI = P(X>x), donde X es una variable aleatoria de χ2.

- **PRUEBA.CHI.INV** Esta función devuelve los valores críticos para una distribución χ2GL, es decir fijada una probabilidad p, por ejemplo 0,05, y dados los grados de libertad GL correspondientes, la función devuelve el valor X de la variable aleatoria tal que

$$P(X \le \chi^2_{GL}) = p$$

Esto es, devuelve los valores que aparecen en las tablas y que se usarán normalmente para comprobar la significación de un resultado. La descripción de esta función que figura en la ayuda de Excel es la siguiente:

Devuelve el inverso de una probabilidad dada, de una sola cola, en una distribución chi cuadrado.
Devuelve para una probabilidad dada, de una sola cola, el valor de la variable aleatoria siguiendo una distribución chi cuadrado.
Si el argumento probabilidad = DISTR.CHI(x;...), entonces PRUEBA. CHI. INV(probabilidad,...) = x. Use esta función para comparar los resultados observados con los resultados esperados, a fin de decidir si la hipótesis original es válida.

PRUEBA.CHI.INV(probabilidad ; grados_de_libertad)

- Probabilidad es una probabilidad asociada con la distribución chi cuadrado.
- Grados_de_libertad es el número de grados de libertad.
- Si uno de los argumentos no es numérico, PRUEBA.CHI.INV devuelve el valor de error #¡VALOR!.
- Si el argumento probabilidad < 0 o probabilidad > 1, PRUEBA.CHI.INV devuelve el valor de error #¡NUM!.
- Si el argumento grados_de_libertad no es un entero, se trunca.
- Si el argumento grados_de_libertad < 1 o grados_de_libertad ≥ 10^10, PRUEBA.CHI.INV devuelve el valor de error #¡NUM!.
- PRUEBA.CHI.INV usa una técnica iterativa para calcular la función. Dado un valor de probabilidad, PRUEBA.CHI.INV reitera hasta que el resultado tenga una exactitud de ± 3x10^-7. Si PRUEBA.CHI.INV no converge después de 100 iteraciones, la función devuelve el valor de error #N/A.

- **PRUEBA.CHI** Finalmente, la función Prueba.chi, devuelve la probabilidad asociada a un contraste (tanto de independencia como de bondad del ajuste) cuando como argumentos se le suministran las frecuencias observadas y las esperadas. La descripción de esta función que figura en la ayuda de Excel es la siguiente:

Devuelve la prueba de independencia. PRUEBA.CHI devuelve el valor de la distribución Chi cuadrado para la estadística y los grados de libertad apropiados. Las pruebas Chi cuadrado pueden usarse para determinar si un experimento se ajusta a los resultados teóricos.

PRUEBA.CHI(rango_actual ; rango_esperado)

- Rango_actual es el rango de datos que contiene observaciones para probar frente a valores esperados.
- Rango_esperado es el rango de datos que contiene la relación del producto de los totales de filas y columnas con el total global.
- Si rango_actual y rango_esperado tienen un número diferente de puntos dedatos, PRUEBA.CHI devuelve el valor de error #N/A.
- La prueba Chi cuadrado primero calcula una estadística Chi cuadrado y después suma las diferencias entre los valores reales y los valores esperados.
- PRUEBA.CHI devuelve la probabilidad para una estadística Chi cuadrado y grados de libertad, gl, donde gl = (r - 1)(c - 1).

Será con esta función con la que llevaremos a cabo los contrastes de independencia, para ello será necesario primero calcular los valores de las frecuencias esperadas bajo la hipótesis nula de independencia. Veremos cómo hacer esto con unos sencillos ejemplos.

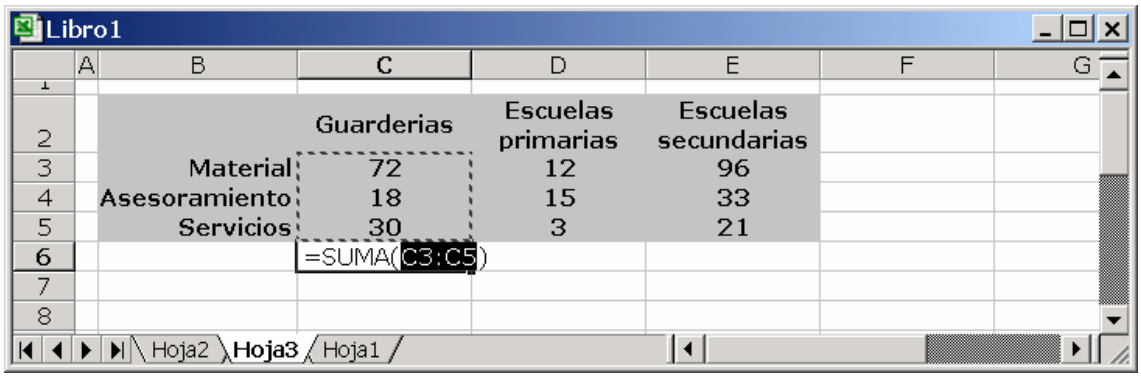
Resolveremos el siguiente ejemplo:

Servicios Psicopedagógicos Telerín				
	Guarderías	Escuelas primarias	Escuelas secundarias	Total
Material	72	12	96	180
Asesoramiento	18	15	33	66
Servicios	30	3	21	54
Total	120	30	150	300

introduciendo en la hoja de cálculo los datos (sin totales):

	Guarderias	Escuelas primarias	Escuelas secundarias
Material	72	12	96
Asesoramiento	18	15	33
Servicios	30	3	21

dejando a Excel la responsabilidad de calcular los totales:



para obtener la tabla completa de las frecuencias observadas

	Guarderias	Escuelas primarias	Escuelas secundarias	
Material	72	12	96	180
Asesoramiento	18	15	33	66
Servicios	30	3	21	54
	120	30	150	300

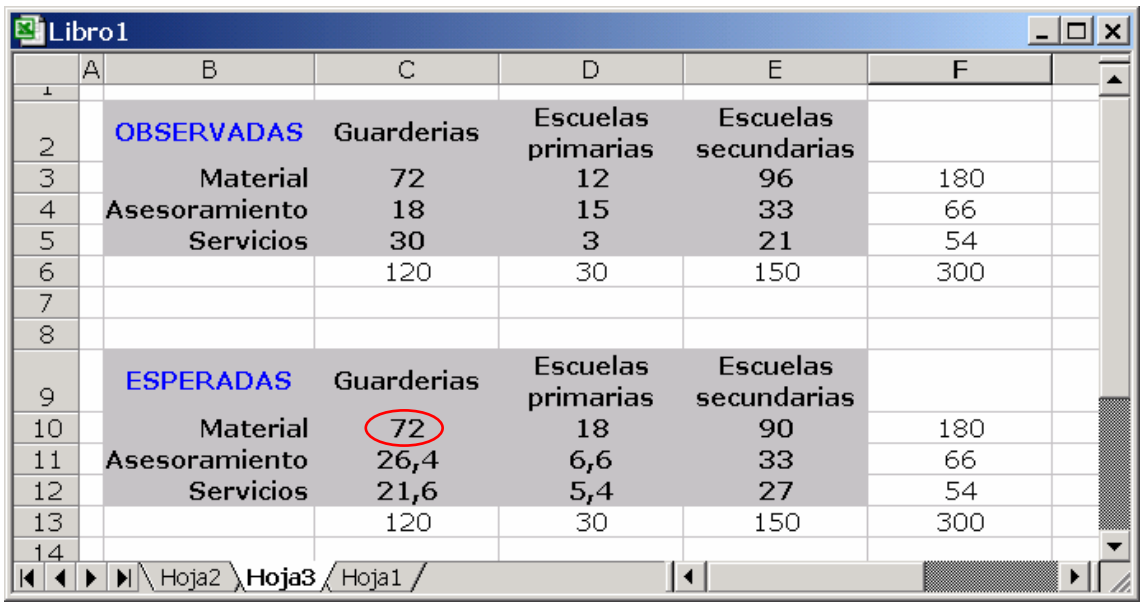
Lo mejor que podemos hacer para construir la tabla de **frecuencias esperadas** es

- copiar la tabla anterior unas cuantas líneas más abajo;
- copiar sobre ella misma sólo los valores (de esa manera se mantendrán los valores de las frecuencias marginales cuyas fórmulas suma... habrán desaparecido manteniéndose los valores calculados anteriormente);
- borrar los contenidos de las celdas correspondientes a las frecuencias observadas;

- calcular las nuevas aplicando las fórmulas dadas por la teoría de probabilidades:
- El resultado de las acciones anteriores, al rellenar la nueva tabla con la fórmula genérica:

=C$13*$F10/F13

Debería ser el siguiente

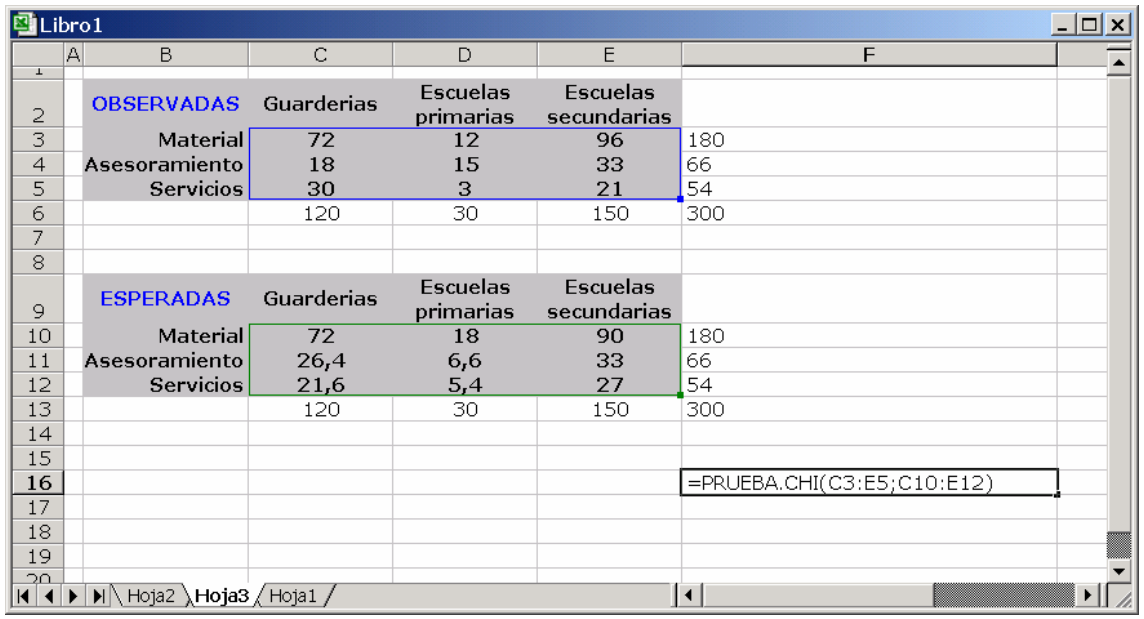


Libro1

	A	B	C	D	E	F
2		OBSERVADAS	Guarderias	Escuelas primarias	Escuelas secundarias	
3		Material	72	12	96	180
4		Asesoramiento	18	15	33	66
5		Servicios	30	3	21	54
6			120	30	150	300
7						
8						
9		ESPERADAS	Guarderias	Escuelas primarias	Escuelas secundarias	
10		Material	72	18	90	180
11		Asesoramiento	26,4	6,6	33	66
12		Servicios	21,6	5,4	27	54
13			120	30	150	300
14						

Hoja2 Hoja3 Hoja1

Una vez construidas las dos tablas basta con aplicar la función anterior



Libro1

	A	B	C	D	E	F
2		OBSERVADAS	Guarderias	Escuelas primarias	Escuelas secundarias	
3		Material	72	12	96	180
4		Asesoramiento	18	15	33	66
5		Servicios	30	3	21	54
6			120	30	150	300
7						
8						
9		ESPERADAS	Guarderias	Escuelas primarias	Escuelas secundarias	
10		Material	72	18	90	180
11		Asesoramiento	26,4	6,6	33	66
12		Servicios	21,6	5,4	27	54
13			120	30	150	300
14						
15						
16						=PRUEBA.CHI(C3:E5;C10:E12)
17						
18						
19						

Hoja2 Hoja3 Hoja1

y obtendríamos la probabilidad asociada a la hipótesis nula de independencia

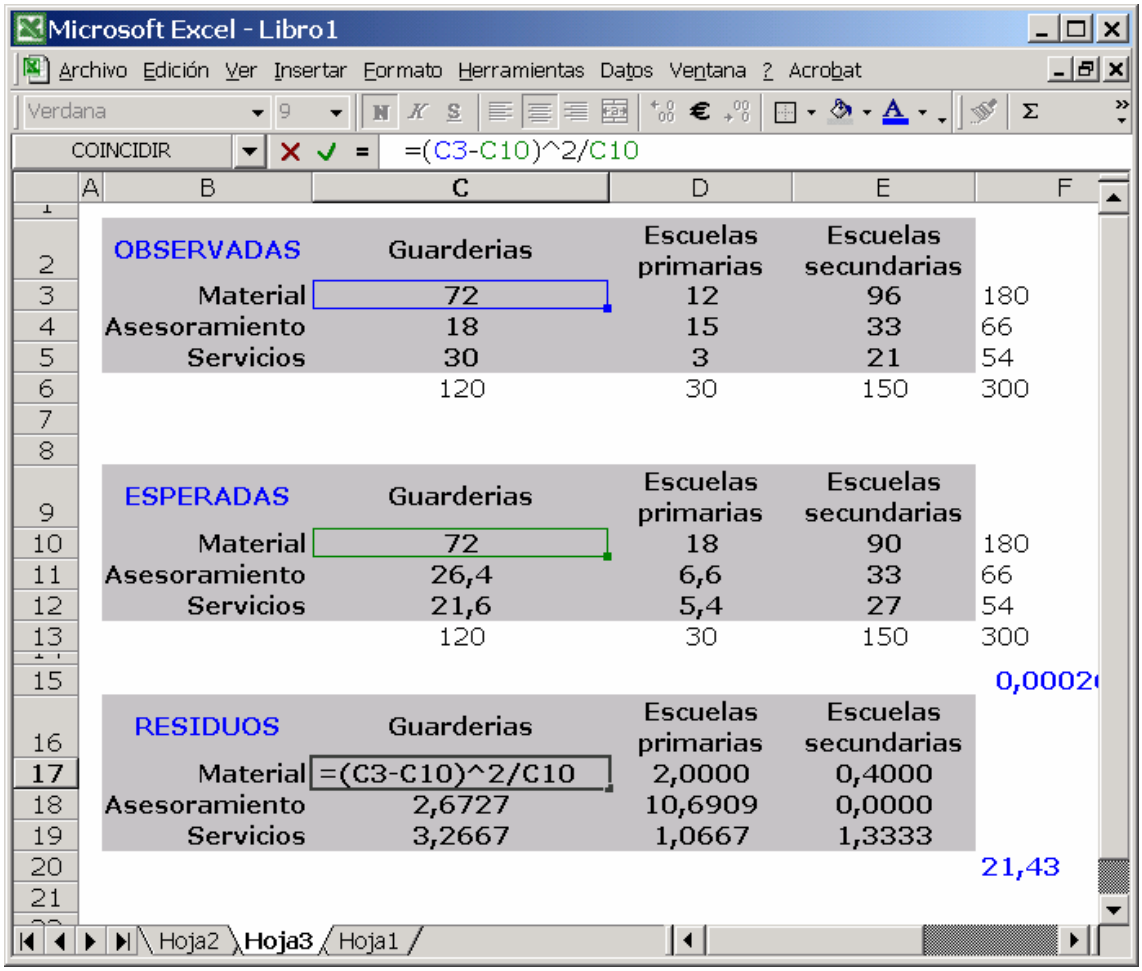
ESPERADAS	Guarderias	Escuelas primarias	Escuelas secundarias	
Material	72	18	90	180
Asesoramiento	26,4	6,6	33	66
Servicios	21,6	5,4	27	54
	120	30	150	300

0,0002601

Notemos que, a diferencia de la resolución "manual", tal como está descrita en el material de la asignatura, lo que obtenemos de este modo es el p.valor de la prueba, y no el valor del estadístico de contraste que habría que comparar después con el valor crítico de la tabla. No obstante, si quisiéramos obtener el valor del estadístico χ^2 (lo cual es aconsejable por los motivos que veremos a continuación), deberíamos construir una tercera tabla sobre la que calcular los residuos, esto es, los sumandos de la fórmula:

$$\chi^2 = \sum \frac{(Obs_i \quad Esp_i)^2}{Esp_i}$$

El proceso de construcción de esta tercera tabla sería idéntico al anterior con la diferencia de que ahora los valores corresponden a los sumandos del estadístico.



El resultado final, suma de los valores de la nueva tabla:

$$\chi^2 = \frac{(72-72)^2}{72} + \frac{(12-18)^2}{18} + \frac{(96-90)^2}{90} + \frac{(18-26{,}4)^2}{26{,}4} + \frac{(15-6{,}6)^2}{6{,}6} +$$

$$+ \frac{(33-33)^2}{33} + \frac{(30-21{,}6)^2}{21{,}6} + \frac{(3-5{,}4)^2}{5{,}4} + \frac{(21-27)^2}{27} = 21{,}43$$

Sin ninguna modificación importante es posible hacer también los contrastes de bondad del ajuste. Resolvamos el siguiente ejemplo:

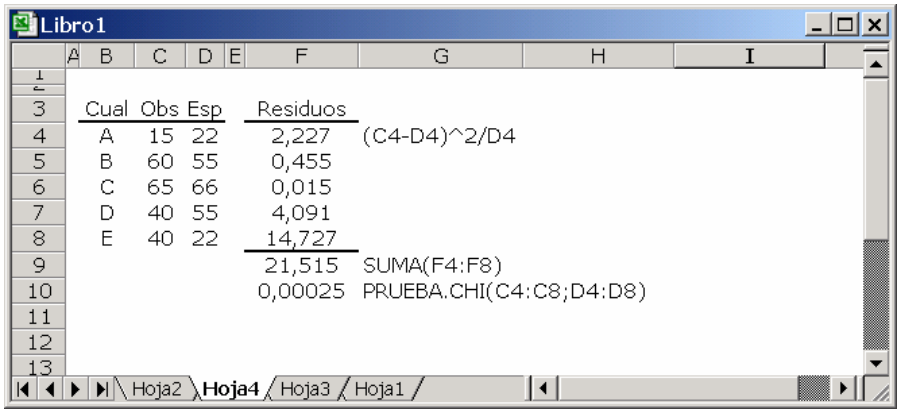
Cualificación	Distribución real	Distribución teórica
A	15	22
B	60	55
C	65	66
D	40	55
E	40	22
Total	**220**	**220**

cuya solución es la siguiente:

Puesto que tenemos cinco categorías, entonces $\nu = 5 - 1 = 4$ grados de libertad. Si consultamos el programa Excel, vemos que en el caso de ser cierta la hipótesis de nulidad, la probabilidad de obtener un valor 21,514 es igual a 0,0002.

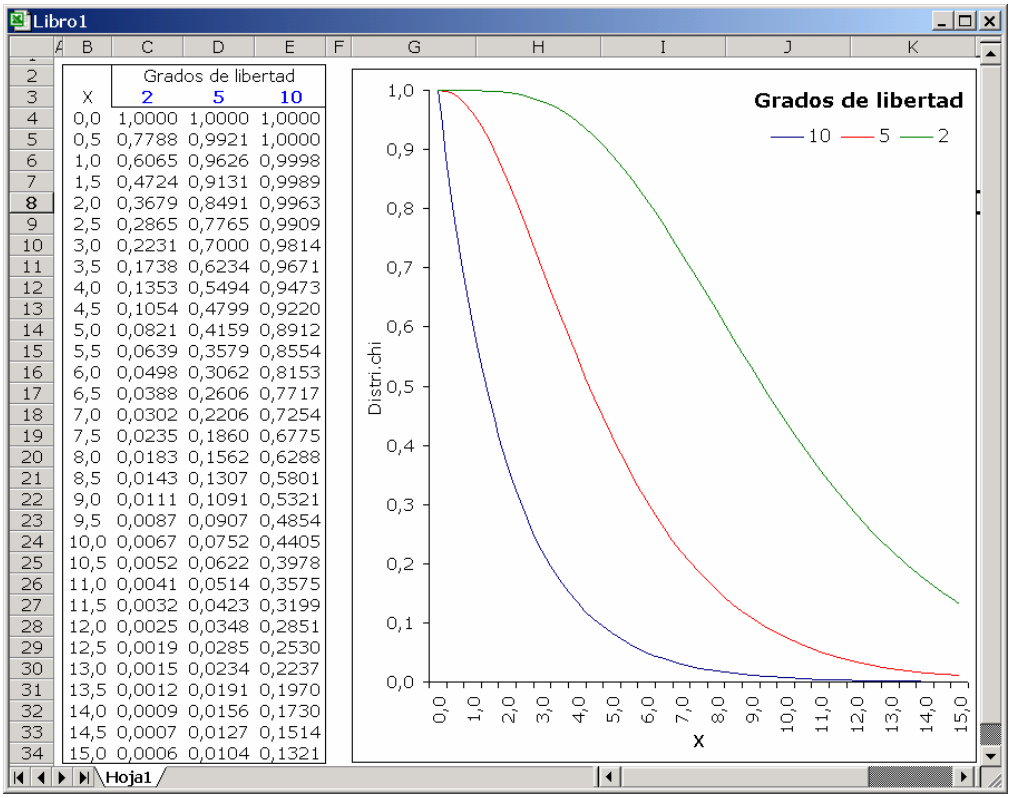
Puesto esta probabilidad es muy pequeña, decidimos rechazar la H_0, o lo que es lo mismo, suponemos que la distribución empírica no se ajusta de forma adecuada a la distribución teórica prefijada.

que coincide con la obtenida a través de la hoja de cálculo:



10.5 PROBLEMAS

10.5.1 Tabular y graficar las funciones de distribución de la Chi2 para 2, 5 y 10 grados de libertad



X	Grados de libertad 2	5	10
0,0	1,0000	1,0000	1,0000
0,5	0,7788	0,9921	1,0000
1,0	0,6065	0,9626	0,9998
1,5	0,4724	0,9131	0,9989
2,0	0,3679	0,8491	0,9963
2,5	0,2865	0,7765	0,9909
3,0	0,2231	0,7000	0,9814
3,5	0,1738	0,6234	0,9671
4,0	0,1353	0,5494	0,9473
4,5	0,1054	0,4799	0,9220
5,0	0,0821	0,4159	0,8912
5,5	0,0639	0,3579	0,8554
6,0	0,0498	0,3062	0,8153
6,5	0,0388	0,2606	0,7717
7,0	0,0302	0,2206	0,7254
7,5	0,0235	0,1860	0,6775
8,0	0,0183	0,1562	0,6288
8,5	0,0143	0,1307	0,5801
9,0	0,0111	0,1091	0,5321
9,5	0,0087	0,0907	0,4854
10,0	0,0067	0,0752	0,4405
10,5	0,0052	0,0622	0,3978
11,0	0,0041	0,0514	0,3575
11,5	0,0032	0,0423	0,3199
12,0	0,0025	0,0348	0,2851
12,5	0,0019	0,0285	0,2530
13,0	0,0015	0,0234	0,2237
13,5	0,0012	0,0191	0,1970
14,0	0,0009	0,0156	0,1730
14,5	0,0007	0,0127	0,1514
15,0	0,0006	0,0104	0,1321

10.5.2 Reproducir la tabla E.4 del texto

10.5.3 Resolver el problema 8.42 (Pág. 448) del texto.

		Cargos Si	No	**Total**
Objetivo	GI	17	26	**43**
	IL	14	17	**31**
	MC	7	10	**17**
	SC	16	16	**32**
	TK	6	11	**17**
	Total	**60**	**80**	**140**

10.5.4 Supóngase la siguiente tabla de contingencia

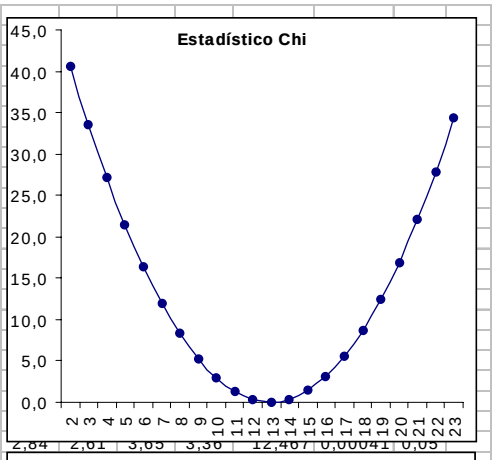
	C_1	C_2
F_1	12	15
F_2	11	10

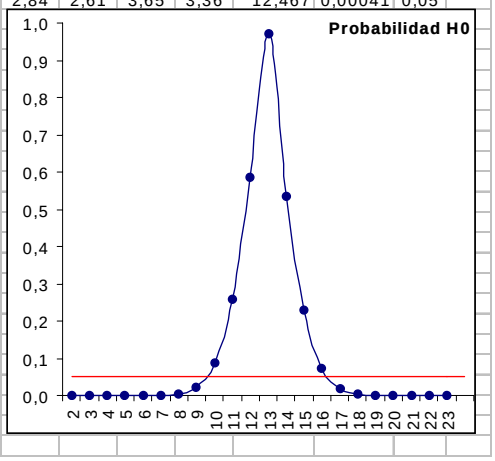
a) Generar todas las posibles tablas para los 48 datos (los marginales han de estar fijos).
b) Calcular el estadístico χ^2 de cada una.
c) Calcular la probabilidad asociada, bajo la hipótesis de independencia, de cada tabla.
d) Indicar cuáles de ellas son compatibles (al 95%) con la hipótesis de independencia

	C1	C2	
F1	12	15	27
F2	11	10	21
	23	25	48

	C1	C2	
F1	12,9	14,1	27
F2	10,1	10,9	21
	23	25	48

12,9	14,1	10,1	10,9	48	
F1C1	F1C2	F2C1	F2C2	SUMA	Chi
0	27	23	-2	48	56,777
1	26	22	-1	48	48,339
2	25	21	0	48	40,580
3	24	20	1	48	33,499
4	23	19	2	48	27,096
5	22	18	3	48	21,372
6	21	17	4	48	16,326
7	20	16	5	48	11,959
8	19	15	6	48	8,270
9	18	14	7	48	5,259
10	17	13	8	48	2,927
11	16	12	9	48	1,273
12	15	11	10	48	0,298
13	14	10	11	48	0,001
14	13	9	12	48	0,383
15	12	8	13	48	1,443
16	11	7	14	48	3,181
17	10	6	15	48	5,598
18	9	5	16	48	8,694
19	8	4	17	48	12,467
20	7	3	18	48	16,920
21	6	2	19	48	22,050
22	5	1	20	48	27,859
23	4	0	21	48	34,347
24	3	-1	22	48	41,513
25	2	-2	23	48	49,357
26	1	-3	24	48	57,880
27	0	-4	25	48	67,081
28	-1	-5	26	48	76,960





10.5.5 Realizar el Test exacto de Fisher para tablas 2x2 para la tabla siguiente:

10	17
13	8

Método:

Sea la tabla

a	b	a+b
c	d	c+d
a+c	b+d	N=a+b+c+d

1. Determinar todas las tablas posibles manteniendo constantes los marginales.
2. Seleccionar de todas estas tablas las que se alejan tanto o más de H_0 que la tabla observada en la dirección de H_1.
3. Calcular las probabilidades de ocurrencia bajo H_0 de dichas tablas.

$$P_{(a,b,c,d)} = \frac{(a+b)!\,(a+b)!\,(a+b)!\,(a+b)!}{a!\,b!\,c!\,d!\,N!}$$

4. Calcular el p-valor del test sumando las probabilidades de dichas tablas.
5. Comparar el p-valor con el nivel de significación α prefijado.
   - Si $p>\alpha$ aceptamos H_0.
   - Si $p\leq\alpha$ rechazamos H_0.

a+b 108888694504184000000000000
c+d 51090942171709400000
a+c 25852016738885000000000
b+d 15511210043331000000000000

a	b	
c	d	
10	17	27
13	8	21
23	25	48

13	14
10	11

		12,9	14,1	11,0	10,0							
a	b	a	b	10,1	10,9	S					χ2	Fisher
12	15	11	10	48	2,27324E+19	39070080	6,47648E+14	4,27447E+18	0,13	0,19806
13	14	10	11	48	1,74865E+18	586051200	7,12412E+15	3,88589E+17	0,19	0,22853
11	16	12	9	48	2,72789E+20	2441880	5,39706E+13	4,27447E+19	0,75	0,12379
14	13	9	12	48	1,24903E+17	8204716800	7,12412E+16	3,23824E+16	0,93	0,19044
10	**17**	**13**	**8**	48	3,00068E+21	143640	4,15159E+12	3,84703E+20	2,04	0,05545
15	12	8	13	48	8,3269E+15	1,06661E+11	6,41171E+17	2,49095E+15	2,35	0,11427
9	18	14	7	48	3,00068E+22	7980	2,96542E+11	3,07762E+21	4,02	0,01760
16	11	7	14	48	5,20431E+14	1,27994E+12	5,12937E+18	1,77925E+14	4,45	0,04897
8	19	15	6	48	2,70061E+23	420	19769460480	2,15433E+22	6,67	0,00389
17	10	6	15	48	3,06136E+13	1,40793E+13	3,59056E+19	1,18617E+13	7,22	0,01479
7	20	16	5	48	2,16049E+24	21	1235591280	1,2926E+23	10,00	0,00058
18	9	5	16	48	1,70076E+12	1,40793E+14	2,15433E+20	7,41355E+11	10,68	0,00308
6	21	17	4	48	1,51234E+25	1	72681840	6,463E+23	14,02	0,00006
19	8	4	17	48	89513424000	1,26714E+15	1,07717E+21	43609104000	14,81	0,00043
5	22	18	3	48	9,07406E+25	0,045454545	4037880	2,5852E+24	18,70	0,00000
20	7	3	18	48	4475671200	1,01371E+16	4,30867E+21	2422728000	19,62	0,00004
4	23	19	2	48	4,53703E+26	0,001976285	212520	7,75561E+24	24,07	0,00000
21	6	2	19	48	213127200	7,09596E+16	1,2926E+22	127512000	25,11	0,00000
3	24	20	1	48	1,81481E+27	8,23452E-05	10626	1,55112E+25	30,12	0,00000
22	5	1	20	48	9687600	4,25758E+17	2,5852E+22	6375600	31,28	0,00000
2	25	21	0	48	5,44443E+27	3,29381E-06	506	1,55112E+25	36,84	0,00000
23	4	0	21	48	421200	2,12879E+18	2,5852E+22	303600	38,13	0,00000
												1,00000

11 Estimación por intervalos.

11.1 Intervalos de estimación más utilizados.

11.1.1 Media de una población normal de σ conocida:

$$\bar{x} \mp Z_{(\alpha/2)} \cdot \frac{\sigma}{\sqrt{n}}$$

podemos hacer los cálculos directamente:

DISTR.NORM.ESTAND.INV(α+(1-α)/2)* Sigma/RAIZ(n)

para calcular el error típico de la estimación (ETE), y obtener los límites mediante:

[PROMEDIO(Dat) - ETE ; PROMEDIO(Dat)+ ETE]

o bien usar directamente la función

- **INTERVALO.CONFIANZA** Devuelve el intervalo de confianza para la media de una población.

INTERVALO.CONFIANZA(alfa;desv_estándar;tamaño)

- **Alfa** es el nivel de significación empleado para calcular el nivel de confianza. El nivel de confianza es igual a 100(1 - alfa)%, es decir, un alfa de 0,05 indica un nivel de confianza de 95%.
- **Desv_estándar** es la desviación estándar de la población y se asume que es conocida.
- **Tamaño** es el tamaño de la muestra.

Observaciones

- Si uno de los argumentos no es numérico, INTERVALO.CONFIANZA devuelve el valor de error #¡VALOR!.
- Si el argumento alfa ≤ 0 o alfa ≥ 1, INTERVALO.CONFIANZA devuelve el valor de error #¡NUM!.
- Si el argumento desv_estándar ≤ 0, INTERVALO.CONFIANZA devuelve el valor de error #¡NUM!.
- Si el argumento tamaño no es un entero, se trunca.
- Si el argumento tamaño < 1, INTERVALO.CONFIANZA devuelve el valor de error #¡NUM!.
- Si suponemos que el argumento alfa es igual a 0,05, se tendrá que calcular el área debajo de la curva normal estándar que es igual a (1 - alfa) o 95%.

Ejemplo

De una población de varillas de hierro se ha extraído un muestra de 64 y calculado su media de resistencia a la rotura que resultó ser 1012kgf/cm^2. Se sabe por experiencia que para este tipo de varillas σ=25. Hallar los límites de confianza de μ al 95%.

n	64	
Alfa	0,95	
Media	1012	
Sigma	25,00	
Z	1,960	DISTR.NORM.ESTAND.INV(Alfa+(1-Alfa)/2)
ETE	6,125	Z*Desv/RAIZ(n)
	6,125	INTERVALO.CONFIANZA(1-Alfa;Sigma;n)
LI	1.005,875	
LS	1.018,125	

11.1.2 Media de una población normal de σ desconocida:

El intervalo para α=95% es:

$$\bar{x} \mp t_{(\alpha/2,n-1)} \frac{Sx}{\sqrt{n}}$$

Podemos usar la combinación de instrucciones:

DISTR.T.INV(1-α;n-1)*DESVEST(Datos)/RAIZ(CONTAR(Datos)))

para calcular el error típico de la estimación (ETE), y obtener los límites mediante:

[PROMEDIO(Dat) - ETE ; PROMEDIO(Dat)+ ETE]

Ejemplo

Con el fin de investigar un nuevo tipo de combustible para cohetes, se depararon cuatro unidades obteniéndose las siguientes velocidades iniciales:

19600
20300
20500
19800

obtener un intervalo de estimación de la media de las velocidades para a un nivel de confianza del 95%.

Datos			
19600	**Alfa**	0,95	
20300	**Media**	20050	
20500	**Desv**	420,32	
19800	**t**	3,182	DISTR.T.INV(1-Alfa;n-1)
	ETE	668,819	
	LI	19.381,181	
	LS	20.718,819	

11.1.3 Varianza de una población normal:

El intervalo para α=95% es:

$$\left[\frac{\left(\frac{S_x}{\sqrt{n-1}} \right)}{\sqrt{\chi^2_{(\alpha/2,n-1)}}} ; \frac{\left(\frac{S_x}{\sqrt{n-1}} \right)}{\sqrt{\chi^2_{(1+\alpha/2,n-1)}}} \right]$$

podemos usar la combinación de instrucciones:

RAIZ(n-1)*DESVEST/RAIZ(Chi1) ; RAIZ(n-1)*DESVEST/RAIZ(Chi2)

siendo:

Chi1 = PRUEBA.CHI.INV(α/2;n-1)

Chi2 =PRUEBA.CHI.INV(α+(α/2);n-1)

Ejemplo

Un fabricante de relojes deseaba calcular un intervalo de estimación de la desviación típica de los tiempos marcados en 100 horas por todos los relojes del mimo modelo. Para ello puso en marcha 10 relojes obteniendo una cuasidesviación típica de los tiempos marcados por cada uno de 50 segundos. Suponiendo normalidad, estimar la desviación de la población al 99%.

n	10	
Alfa	0,99	
Alfa/2	0,005	
Alfa+(Alfa/2)	0,995	
Desv	50	
Chi1	23,589	PRUEBA.CHI.INV(Alfa/2;n-1)
Chi2	1,735	PRUEBA.CHI.INV(Alfa+(Alfa/2);n-1)
LI	30,8840	RAIZ(n-1)*Desv/RAIZ(Chi1)
LS	113,8814	RAIZ(n-1)*Desv/RAIZ(Chi2)

11.1.4 Desviación típica de una población normal:

$$\left(\sqrt{\frac{(n-1)S_x^2}{\chi^2_{(\alpha/2,n-1)}}} \; ; \sqrt{\frac{(n-1)S_x^2}{\chi^2_{(1-\alpha/2,n-1)}}} \right)$$

11.1.5 Parámetro **p** de una distribución binomial (n·p grande)

$$\hat{p} \mp Z_{(\alpha/2)} \cdot \sqrt{\frac{\hat{p} \cdot (1-\hat{p})}{n}}$$

siendo $\hat{p} = x/n$

11.1.6 Parámetro **p** de una distribución binomial (sin condiciones)

Sabemos que el intervalo exacto viene dado por:

$$\left(\frac{x}{x + (n-x+1) \cdot F_{\alpha/2;2(n-x+1);2x}} \quad ; \quad \frac{(x+1) \cdot F_{\alpha/2;2(x+1);2(n-x)}}{(n-x) + (x+1) \cdot F_{\alpha/2;2(x+1);2(n-x)}} \right)$$

Usaremos la función:

DISTR.F.INV(n; GL_1; GL_2)

11.1.7 Parámetro de una distribución de Poisson

$$\left[\hat{\lambda} \pm Z_{\alpha/2} \sqrt{\frac{\hat{\lambda}}{n}} \right]$$

siendo $\hat{\lambda} = \dfrac{\sum x_i}{n}$

11.1.8 Diferencia de dos proporciones

$$\left[(\hat{p}_1 - \hat{p}_2) \pm Z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \right]$$

11.2 PROBLEMAS

11.2.1 Tomamos un muestra aleatoria de tamaño 16, procedente de una distribución normal de desviación típica 6, y obtenemos una media muestral de valor 25. Hallar un intervalo de confianza del 90% para la media poblacional.

Contestamos usando la fórmula directamente:

$$\bar{x} \mp Z_{(\alpha/2)} \cdot \frac{\sigma}{\sqrt{n}} \quad \Rightarrow \quad 25 \mp Z_{(\alpha/2)} \frac{6}{\sqrt{16}} \quad \Rightarrow \quad 25 \mp Z_{(\alpha/2)} \frac{6}{\sqrt{16}}$$

y haciendo los cálculos sobre la hoja:

Med	25			
n	16	Z(Alfa/2)	1,6449	DISTR.NORM.ESTAND.INV(Alfa+(Alfa/2))
Sigma	6	s/raiz(n)	1,500	n/RAIZ(Sigma)
Alfa	0,9		2,4673	ETE
(1-Alfa)/2	0,05	**L. Inf**	**22,5327**	Med-ETE
Alfa+(Alfa/2)	0,95	**L. Sup**	**27,4673**	Med+ETE
1-Alfa	0,10			

o bien usando la función **INTERVALO.CONFIANZA** para obtener directamente el Error típico de la estimación (ETE)

Med	25			
n	16	Z(Alfa/2)	1,6449	DISTR.NORM.ESTAND.INV(Alfa+(Alfa/2))
Sigma	6	s/raiz(n)	1,500	n/RAIZ(Sigma)
Alfa	0,9		2,4673	ETE
(1-Alfa)/2	0,05	**L. Inf**	**22,5327**	Med-ETE
Alfa+(Alfa/2)	0,95	**L. Sup**	**27,4673**	Med+ETE
1-Alfa	0,10			
			2,4673	INTERVALO.CONFIANZA(1-Alfa;Sigma;n)

11.2.2 Una muestra aleatoria de seis vehículos tienen los siguientes consumos (en Km/l).

{18,6 ; 18,4 ; 19,2 ; 20,8 ; 19,4 ; 20,5}

a) Calcular un intervalo de confianza del 90% para el consumo medio poblacional.
b) Generalizar para 80%,90%,95%,99%.

El intervalo viene dado por:

$$\bar{x} \mp t_{(\alpha/2,n-1)} \cdot \frac{S_x}{\sqrt{n}}$$

Podemos utilizar el módulo Análisis de Datos (Estadística descriptiva)

Cons
18,6
18,4
19,2
20,8
19,4
20,5
19,48

Cons	
Nivel de confianza(90,0%)	0,80671904
L. Inf	18,68
L. Sup	20,29

O bien usar la fórmula directamente

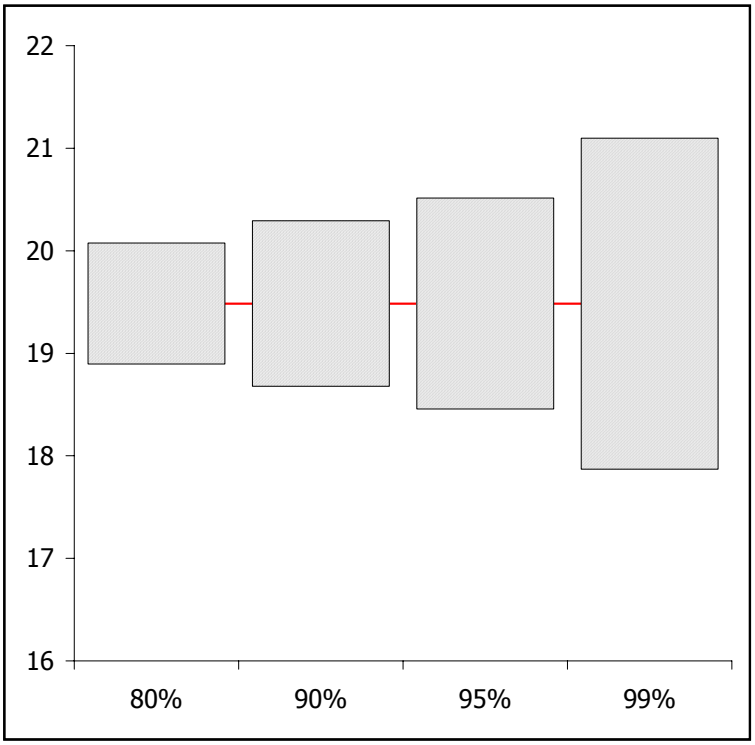
Cons
18,6
18,4
19,2
20,8
19,4
20,5
19,48

Alfa	0,9

Cons	
Nivel de confianza(90,0%)	0,806719037
L. Inf	**18,68**
L. Sup	**20,29**
Sx/Raiz(n)	0,40035
t(1-Alfa)	2,01505
ETE	0,80672
L. Inf	**18,68**
L. Sup	**20,29**

Las fórmulas nos permiten hacer fácilmente la generalización

Alfa	**0,8**	**0,9**	**0,95**	**0,99**
t(1-Alfa)	1,47588	2,01505	2,57058	4,03212
ETE	0,59087	0,80672	1,02912	1,61425
L. Inf	**18,89**	**18,68**	**18,45**	**17,87**
L. Sup	**20,07**	**20,29**	**20,51**	**21,10**



11.2.3 Para el siguiente conjunto de datos (que supondremos proviene de una población normal.

6,80	6,78	6,77	6,80	6,78	6,80	6,82	6,81	6,80	6,79

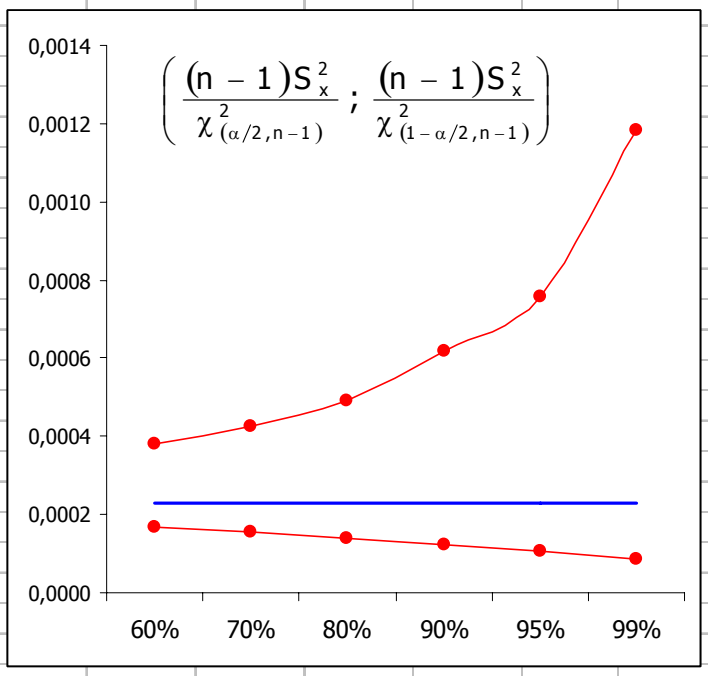
se pide estimar la media y varianza al 95% y al 65%

W				
6,80				
6,78	MEDIA(a)95%	6,78420	6,80580	
6,77	VAR(a)95%	0,00011	0,00076	
6,80				
6,78	MEDIA(a)65%	6,79030	6,79970	
6,80	VAR(a)65%	0,00016	0,00040	
6,82				
6,81				
6,80				
6,79				

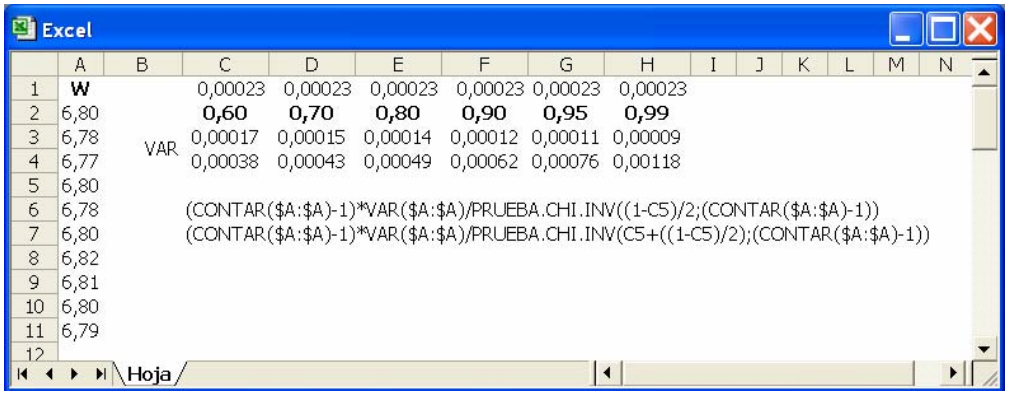
11.2.4 Para el conjunto de datos anterior representar gráficamente el intervalo de estimación de la varianza a los siguientes niveles de confianza

0,60 0,70 0,80 0,90 0,95 0,99

W		0,00023	0,00023	0,00023	0,00023	0,00023	0,00023
6,80		**0,60**	**0,70**	**0,80**	**0,90**	**0,95**	**0,99**
6,78	VAR	0,00017	0,00015	0,00014	0,00012	0,00011	0,00009
6,77		0,00038	0,00043	0,00049	0,00062	0,00076	0,00118
6,80							
6,78							
6,80							
6,82							
6,81							
6,80							
6,79							



las fórmulas empleadas son las siguientes:



11.2.5 El tamaño muestral necesario para conseguir una estimación que verifique que:

$$\left|\bar{x} - \mu\right| \leq E$$

viene dado por la expresión:

$$n = \left(\frac{z_{(\alpha/2)} \cdot \sigma}{E} \right)^2$$

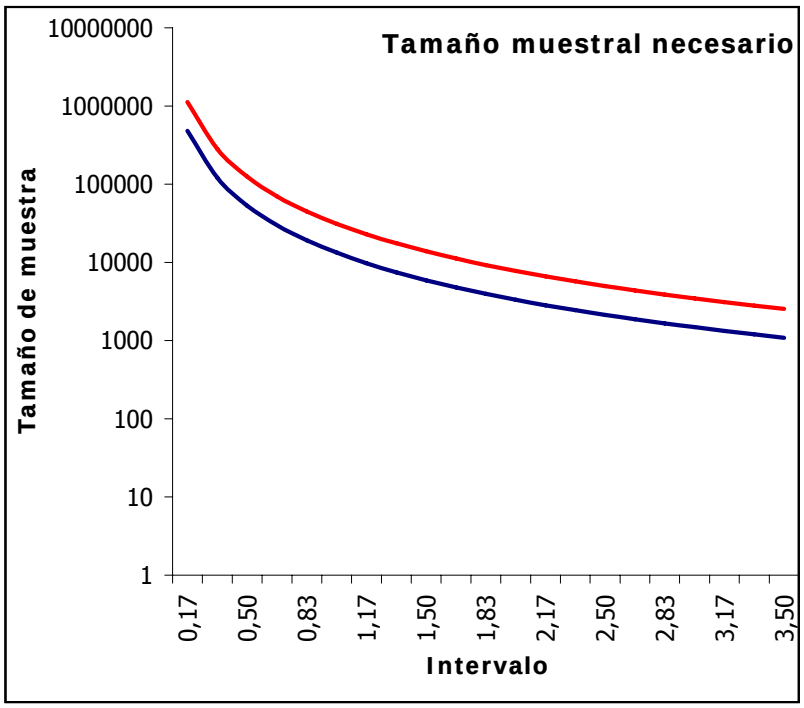
por otra parte, cualquier estimación tiene un precisión definida por

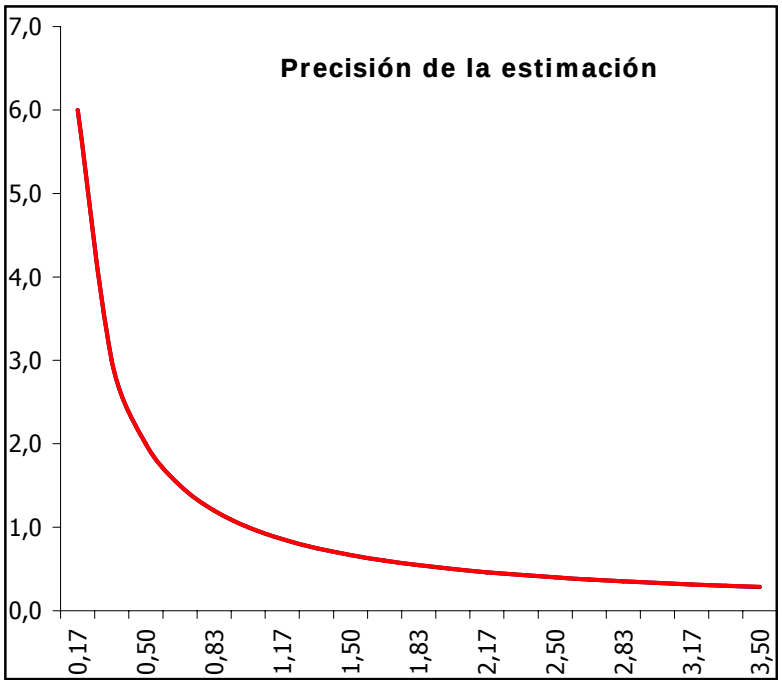
$$PRE = \frac{1}{z_{(\alpha/2)} \cdot \left(\frac{\sigma}{\sqrt{n}} \right)}$$

a) Obtener los valores de n y PRE para una población de s = 90 al 80% y 95%.
b) Graficar ambos valores.

Sigma 90

	Alfa	80%		95%	
	Z	1,282		1,960	
	E	**N**	**PRE**	**N**	**PRE**
1	0,17	478915	6,000	1120166	6,000
2	0,33	119728	3,000	280041	3,000
3	0,50	53212	2,000	124462	2,000
4	0,67	29932	1,500	70010	1,500
5	0,83	19156	1,200	44806	1,200
6	1,00	13303	1,000	31115	1,000
7	1,17	9773	0,857	22860	0,857
8	1,33	7483	0,750	17502	0,750
9	1,50	5912	0,667	13829	0,667
10	1,67	4789	0,600	11201	0,600
11	1,83	3957	0,545	9257	0,545
12	2,00	3325	0,500	7778	0,500
13	2,17	2833	0,461	6628	0,462
14	2,33	2443	0,429	5715	0,429
15	2,50	2128	0,400	4978	0,400
16	2,67	1870	0,375	4375	0,375
17	2,83	1657	0,353	3876	0,353
18	3,00	1478	0,333	3457	0,333
19	3,17	1326	0,316	3102	0,316
20	3,33	1197	0,300	2800	0,300
21	3,50	1085	0,286	2540	0,286

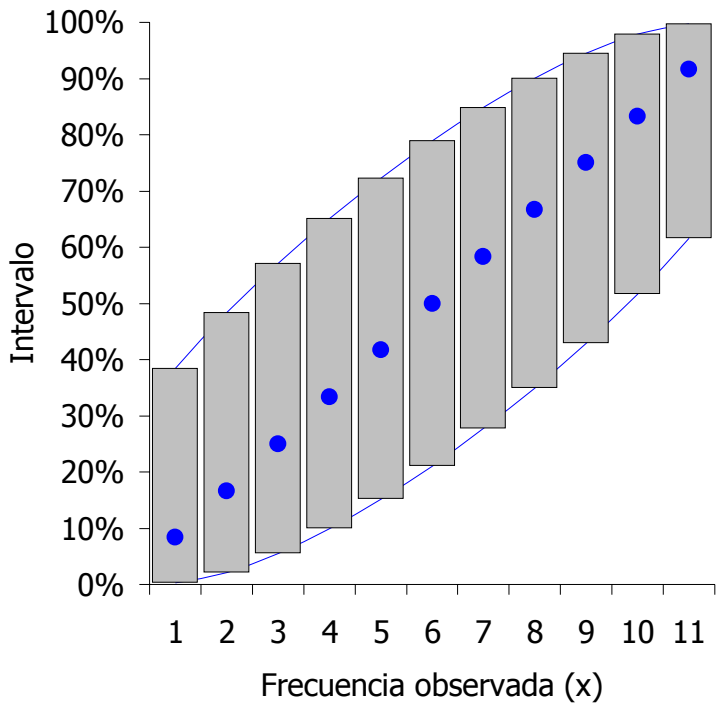

Tamaño muestral necesario
Tamaño de muestra
10000000
1000000
100000
10000
1000
100
10
1
0,17
0,50
0,83
1,17
1,50
1,83
2,17
2,50
2,83
3,17
3,50
Intervalo


Precisión de la estimación
7,0
6,0
5,0
4,0
3,0
2,0
1,0
0,0
0,17
0,50
0,83
1,17
1,50
1,83
2,17
2,50
2,83
3,17
3,50

11.2.6 Un botánico quiere investigar la fracción de plantas obtenidas mediante cierto cruce que presentan el carácter A. Para ello observó que, de 12 plantas, 3 de ellas presentaron dicho carácter. A partir de estos datos, determinar límites de confianza al 95% de la fracción p de plantas de la población que poseen el carácter **A**.

a) Generalizar los resultados suponiendo que el número x de plantas observadas son el carácter A es x∈ {1,2,...,11}.
b) Obtener un gráfico de los limites de estimación de p.
c) Comprobar los resultados obtenidos (para p=0,5) mediante una generación de un numero suficiente de v.a. binomiales B(n=12;p=0,5) y halar los límites al 95% de confianza.

n	12										
Alfa	0,95										
Alfa/2	0,025										
x	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
2(n-x+1)	24	22	20	18	16	14	12	10	8	6	4
2x	2	4	6	8	10	12	14	16	18	20	22
2(x+1)	4	6	8	10	12	14	16	18	20	22	24
2(n-x)	22	20	18	16	14	12	10	8	6	4	2
F1	39,457	8,533	5,168	4,034	3,496	3,206	3,050	2,986	3,005	3,128	3,440
F2	3,440	3,128	3,005	2,986	3,050	3,206	3,496	4,034	5,168	8,533	39,457
LI	0,2%	2,1%	5,5%	9,9%	15,2%	21,1%	27,7%	34,9%	42,8%	51,6%	61,5%
p	8,3%	16,7%	25,0%	33,3%	41,7%	50,0%	58,3%	66,7%	75,0%	83,3%	91,7%
LS	38,5%	48,4%	57,2%	65,1%	72,3%	78,9%	84,8%	90,1%	94,5%	97,9%	99,8%



Contrastes de hipótesis

11.3 Contrastes más usuales.

11.3.1 Contraste de la media de una población normal con varianza conocida:

$$\frac{\bar{x}-\mu_0}{\left(\frac{\sigma}{\sqrt{n}}\right)} \approx \mathbf{N}_{(0,1)}$$

Contraste bilateral

Hipótesis nula $H_0 : \mu = \mu_0$

Hipótesis alternativa $H_1 : \mu \neq \mu_0$

Se mantiene H_0 sí

$$\frac{|\bar{x}-\mu_0|}{\left(\frac{\sigma}{\sqrt{n}}\right)} \leq Z_{\alpha/2}$$

Se rechaza H_0 sí

$$\frac{|\bar{x}-\mu_0|}{\left(\frac{\sigma}{\sqrt{n}}\right)} > Z_{\alpha/2}$$

Contraste unilateral

Hipótesis nula $H_0 : \mu \leq \mu_0$

Hipótesis alternativa $H_1 : \mu > \mu_0$

Se mantiene H_0 si

$$\frac{|\bar{x}-\mu_0|}{\left(\frac{\sigma}{\sqrt{n}}\right)} \leq Z_{\alpha}$$

Se rechaza H_0 si

$$\frac{|\bar{x}-\mu_0|}{\left(\frac{\sigma}{\sqrt{n}}\right)} > Z_{\alpha}$$

11.3.2 Media de normal respecto a un valor nominal con varianza desconocida

$$\frac{\bar{x}-\mu_0}{\left(\frac{S_x}{\sqrt{n}}\right)} \approx t_{n-1}$$

Contraste bilateral

Hipótesis nula $H_0 : \mu = \mu_0$

Hipótesis alternativa $H_1 : \mu \neq \mu_0$

Se mantiene H_0 si

$$\frac{|\bar{x} - \mu_0|}{\left(\frac{S_x}{\sqrt{n}}\right)} \leq t_{(\alpha/2,n-1)}$$

Se rechaza H_0 si

$$\frac{|\bar{x} - \mu_0|}{\left(\frac{S_x}{\sqrt{n}}\right)} > t_{(\alpha/2,n-1)}$$

Contraste unilateral

Hipótesis nula $H_0 : \mu \leq \mu_0$

Hipótesis alternativa $H_1 : \mu > \mu_0$

Se acepta H0 si

$$\frac{|\bar{x} - \mu_0|}{\left(\frac{S_x}{\sqrt{n}}\right)} \leq t_{(\alpha,n-1)}$$

Se rechaza H_0 si

$$\frac{|\bar{x} - \mu_0|}{\left(\frac{S_x}{\sqrt{n}}\right)} > t_{(\alpha,n-1)}$$

11.3.3 Contraste de igualdad de medias de dos poblaciones normales de varianzas desconocidas:

Muestras grandes n1+ n2>30; $n_1 \approx n_2$

Hipótesis nula $H_0 : \mu = \mu_0$

Hipótesis alternativa $H_1 : \mu \neq \mu_0$

Se acepta H_0 si

$$\frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \leq Z_{\alpha/2}$$

Se rechaza H_0 si

$$\frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} > Z_{\alpha/2}$$

Muestras pequeñas: $n_1 + n_2 \leq 30$

Varianzas desconocidas pero iguales: $\sigma_1^2 = \sigma_2^2$

Hipótesis nula $H_0 : \mu = \mu_0$

Hipótesis alternativa $H_1 : \mu \neq \mu_0$

Se acepta H_0 si

$$\frac{\bar{x} - \bar{y}}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \leq t_{\alpha/2,(n_1+n_2-2)}$$

Se rechaza H_0 si

$$\frac{\bar{x} - \bar{y}}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} > t_{\alpha/2,(n_1+n_2-2)}$$

Muestras pequeñas: $n_1 + n_2 \leq 30$.

Varianzas desconocidas y distintas:

Se acepta H_0 si

$$\frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \leq t_{\alpha/2,f}$$

Se rechaza H_0 si

$$\frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} > t_{\alpha/2,f}$$

11.3.4 Varianza de normal respecto a un valor nominal

$$\frac{(n-1)\cdot S^2}{\sigma^2} \approx \chi^2_{n-1}$$

11.3.5 Igualdad de varianzas

$$\frac{S_1^2}{S_2^2} \approx \mathbf{F}_{(n-1,n-2)}$$

11.3.6 Proporción respecto a un valor nominal

$$\frac{\hat{p} - p_0}{\sqrt{\frac{p_0 \cdot (1 - p_0)}{n}}} \approx \mathbf{N}_{(0,1)}$$

Contraste bilateral

Hipótesis nula $H_0 : p = p_0$

Hipótesis alternativa $H_1 : p \neq p_0$

Se acepta H_0 si

$$\frac{\left|\hat{p}-p_0\right|}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \le Z_{\alpha/2}$$

Contraste unilateral

Hipótesis nula $H_0 : p \le p_0$

Hipótesis alternativa $H_1 : p > p_0$

Se mantiene H_0 si

$$\frac{\left|\hat{p}-p_0\right|}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \le Z_{\alpha}$$

Se rechaza H_0 si

$$\frac{\left|\hat{p}-p_0\right|}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} > Z_{\alpha}$$

11.3.7 Igualdad de proporciones

$$\frac{\hat{p}_1-\hat{p}_2}{\sqrt{\frac{\hat{p}_1\cdot(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2\cdot(1-\hat{p}_2)}{n_2}}} \approx \mathbf{N}_{(0,1)}$$

11.4 Funciones de Excel relacionadas

Para calcular los valores críticos:

- **DISTR.NORM.ESTAND.INV**
- **DISTR.NORM.INV**
- **DISTR.INV.F**
- **DISTR.T.INV**
- **PRUEBA.CHI.INV**

Para calcular el p.valor

- **DISTR.CHI**
- **DISTR.F**
- **DISTR.T**
- **DISTR.NORM**
- **DISTR.NORM.ESTAND**
- **PRUEBA.CHI**

11.5 PROBLEMAS

11.5.1 El encargado de la sección de camisería de un gran almacén desea comprobar que las camisas que le suministra el mayorista, y en cuya etiqueta figura "33 centímetros de manga", cumplen realmente esa especificación. Toma una muestra de 100 camisas y obtiene una media muestral de 34,0 centímetros y una (cuasi)desviación típica de 2 centímetros. ¿Qué puede afirmar al respecto? (α=0,0025).

11.5.2 Un sociólogo desea demostrar que el salario medio de un tipo de trabador es de 600€ semanales, tal como indica la prensa. Toma una muestra de 100 trabajadores de dicho sector y obtiene una media de 657€ y una (cuasi)desviación típica de 22 €. ¿Qué puede afirmar? (α=0,05).

11.5.3 El gerente de una fábrica tiene la impresión de que el coste mensual del mantenimiento de sus equipos no es de 500€ por máquina tal como se había proyectado en un principio. Toma una muestra de 32 máquinas y obtiene un coste medio de 592€ y una (cuasi)desviación típica de 101€. ¿Qué puede afirmar? (α=0,02).

11.5.4 El encargado de la compra de materia prima de una fábrica de salsa de tomate desea probar si es cierto, tal como dicen sus suministradores, que el 80% de los tomates que éstos le envían es de calidad "superior". Al analizar una muestra de 100 tomates, encuentra que 72 de ellos poseen dicha calidad, siendo el resto de una calidad inferior. ¿Qué puede afirmar? (α=0,05).

11.5.5 Un anunciante desea confirmar la afirmación del editor de una revista cuando éste dice que "el 25% de sus lectores son estudiantes universitarios". Toma una muestra de 200 lectores de los que 38, resultan ser estudiantes universitarios. Contrastar la hipótesis del editor de la revista. (α=0,01).

11.5.6 Un investigador médico desea saber si las ratas de laboratorio pueden vivir normalmente con sangre artificial. Experimenta con 16 ratas (cuya vida media se sabe perfectamente que sigue una distribución normal de media 5 meses) y obtiene una vida media de 4,1 meses y una (cuasi)desviación típica de 1,6 meses. ¿Qué se puede afirmar? (α=0,001).

11.5.7 La vida media de las bombillas de una fábrica es, teóricamente, de 190 meses. Se prueban 25 bombillas y se obtiene una media de 193 meses y una (cuasi)desviación típica de 3 meses. ¿Qué se puede afirmar? (α=0,05).

11.5.8 Un contable afirma que el tiempo medio que ciertas empresas tardan en pagar sus deudas es superior a 3 meses, exactamente afirma que "el 80% de las empresas tardan más de 3 meses en pagar". Elegidas 50 empresas encuentra que 20 de ellas pagaron antes de esa fecha. ¿Qué se puede afirmar? (α=0,001).

11.5.9 Los siguientes datos corresponden a la longitud en cm de 18 pedazos de cable sobrantes en cada rollo utilizado en un tipo de tarea.

9.0	3.41	6.13	1.99	6.92	3.12	7.86	2.01	5.98
4.15	6.87	1.97	4.01	3.56	8.04	3.24	5.05	7.37

Basándonos en estos datos, ¿podemos afirmar que la longitud media de los pedazos sobrantes es superior a 4 cm? (α=0,05).

11.5.10 El peso de los pollos de una granja se distribuye de forma normal, con media 2.6 Kg. y desviación típica 0,5 Kg. Se experimenta un nuevo tipo de alimentación con 50 crías, que al llegar a adultos alcanzan un peso medio de 2,78 Kg. ¿Qué puede decirse de la nueva alimentación? (α=0,01).

11.5.11 En un medio de comunicación se asegura que la cuota de mercado de una conocida cadena de comida rápida es del 30%. El director de la compañía no está de acuerdo con esta afirmación y decide encargar una encuesta. De 400 consumidores que fueron entrevistados, 140 aseguraron

que eran clientes de dicha cadena. ¿Debe el director aceptar los datos publicados? (α=5%).

11.5.12 Una compañía aérea quiere saber si el tiempo medio de los retrasos en los vuelos Paris-Madrid, que hasta la fecha había sido de 20 minutos, ha aumentado en los últimos meses. Toma una muestra de 21 vuelos y obtiene una media muestral de 22 minutos y una (cuasi)desviación típica de 5. (α=5%).

11.5.13 Generar 100 valores de una N(μ,σ), elegir un alfa.

a) Estimar μ y σ

b) Fijar un valor nominal y realizar los tres contrastes respecto del valor nominal de la media

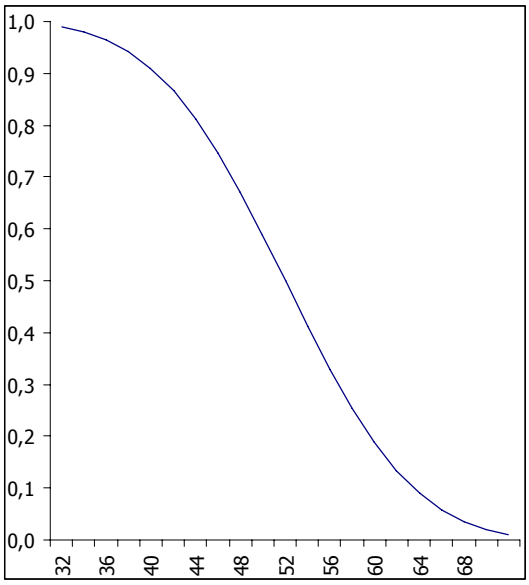
c) Obtener el p.valor

7,173					
13,893	μ	12			
9,927	σ	2			
13,188					
10,062	Alfa	0,95			
12,479	Media	11,35			
11,390	Desviación	2,35			
9,448					
13,843	NOMINAL		12		
13,006	Discrep		-2,2925		
7,846	Unilateral	-1,64			**RECHAZA**
13,303	Bilateral	-1,96		1,96	**RECHAZA**
15,065	Unilateral			1,64	**ACEPTA**
10,801	p.valor		0,011		
12,586					

11.5.14 De una muestra de 150 hombres, 75 resultaron poseer cierta característica genética. ¿Cuántas mujeres, de un grupo de 100, deberían poseer como mínimo dicha característica para que no rechazáramos la hipótesis de igualdad de proporciones entre géneros?

	Si	n	p	Num	Den1	Den2	Dis	P.valor
1. Hombres	75	150	0,5					
2. Mujeres	30	100	0,3	0,2	0,0408	0,0458	2,3081	0,990
	32		0,320	0,180		0,0466	2,0578	0,980
	34		0,340	0,160		0,0474	1,8141	0,965
	36		0,360	0,140		0,0480	1,5761	0,943
	38		0,380	0,120		0,0485	1,3428	0,910
	40		0,400	0,100		0,0490	1,1134	0,867
	42		0,420	0,080		0,0494	0,8871	0,812
	44		0,440	0,060		0,0496	0,6633	0,746
	46		0,460	0,040		0,0498	0,4412	0,670
	48		0,480	0,020		0,0500	0,2203	0,587
	50		0,500	0,000		0,0500	0,0000	0,500
	52		0,520	-0,020		0,0500	-0,2203	0,413
	54		0,540	-0,040		0,0498	-0,4412	0,330
	56		0,560	-0,060		0,0496	-0,6633	0,254
	58		0,580	-0,080		0,0494	-0,8871	0,188
	60		0,600	-0,100		0,0490	-1,1134	0,133
	62		0,620	-0,120		0,0485	-1,3428	0,090
	64		0,640	-0,140		0,0480	-1,5761	0,057
	66		0,660	-0,160		0,0474	-1,8141	0,035
	68		0,680	-0,180		0,0466	-2,0578	0,020
	70		0,700	-0,200		0,0458	-2,3081	0,010

$$\frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1 \cdot (1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2 \cdot (1-\hat{p}_2)}{n_2}}} \approx \mathbf{N}_{(0,1)}$$



12 Series temporales (Tratamiento clásico)

12.1 Introducción

Extraído de **Pepió M.** *"Series temporales"*. Ediciones UPC, 2001.

Una serie temporal es un conjunto de observaciones ordenadas en el tiempo o, también, la evolución de un fenómeno o variable a lo largo de él. Esta variable puede ser económica (ventas de una empresa, consumo de cierto producto, evolución de los tipos de interés,...), física (evolución del caudal de un río, de la temperatura de una región, etc.) o social (número de habitantes de un país, número de alumnos matriculados en ciertos estudios, votos a un partido,...).

El objetivo del análisis de una serie temporal, de la que se dispone de datos en períodos regulares de tiempo, es el conocimiento de su patrón de comportamiento para prever la evolución futura, siempre bajo el supuesto de que las condiciones no cambiarán respecto a las actuales y pasadas.

Si al conocer la evolución de la serie en el pasado se pudiese predecir su comportamiento futuro sin ningún tipo de error, estaríamos frente a un fenómeno determinista cuyo estudio no tendría ningún interés especial.

En general, las series de interés llevan asociados fenómenos aleatorios, de forma que el estudio de su comportamiento pasado sólo permite acercarse a la estructura o modelo probabilístico para la predicción del futuro.

12.2 Análisis de una Serie Temporal

Antes de abordar cualquier estudio analítico de una serie temporal, se impone una representación gráfica de la misma y la observación detenida de su aspecto evolutivo. Para estudiar el comportamiento de cualquier serie temporal, y predecir los valores que puede tomar en un futuro, puede hablarse de distintas metodologías, que denominaremos modelización por componentes y enfoque Box-Jenkins.

12.3 Modelización por componentes

Este método consiste en identificar, en la serie Y_t, cuatro componentes teóricas, que no tienen por qué existir todas, y que son:

1. Tendencia: T_t.
2. Estacionalidad: E_t.
3. Ciclos: C_t.
4. Residuos: R_t.

Cada una de estas componentes es una función del tiempo y el análisis consistirá en la separación y obtención de cada una de ellas, así como en determinar de qué forma se conjugan para dar lugar a la serie original.

La **tendencia** es la componente general a largo plazo y se suele expresar como una función del tiempo de tipo polinómico o logarítmico

Las **variaciones estacionales** son oscilaciones que se producen, y repiten, en períodos de tiempo cortos. Pueden estar asociadas a factores dinámicos, por ejemplo la ocupación hotelera, la venta de prendas de vestir, de juguetes, etc., cuya evolución está claramente ligada a la estacionalidad climática, vacacional, publicitaria, etc.

Las **variaciones cíclicas** se producen a largo plazo y suelen ir ligadas a etapas de prosperidad o recesión económica. Suelen ser tanto más difíciles de identificar cuanto más largo sea su período, debido, fundamentalmente, a que el tiempo de recogida de información no aporta suficientes datos, por lo que a veces quedarán confundidas con las otras componentes.

La componente **residual** es la que recoge la aportación aleatoria de cualquier fenómeno sujeto al azar.

Para evaluar las distintas componentes se utilizan técnicas estadísticas tales como modelo lineal, medias móviles, diferencias finitas, etc.

Admitiendo que el componente aleatorio (residuo) es aditivo, una vez identificadas las otras componentes surge un nuevo problema que es el cómo conjugar tendencia, estacionalidad y ciclos para dar lugar a la serie definitiva.

Así se proponen, entre otros, modelos genéricamente denominados aditivos y multiplicativos.

- Modelo aditivo: **Y = T+E+C + R**
- Modelo multiplicativo: **Y = T·E·C + R**

Para una primera identificación visual del caso, se puede considerar que si el patrón estacional se mantiene con amplitud constante se tratará de modelo aditivo; cuando dicho patrón se vaya amplificando con el tiempo, será multiplicativo.

12.4 Descomposición de una serie temporal

Este método, también denominado **sistema clásico**, descompone la serie en tendencia, estacionalidad, ciclos y residuos Una vez decidida la conjunción entre ellos, aditiva o multiplicativa, se obtiene el modelo con el que hacer previsiones. La tendencia es la componente más importante de la serie, al definir lo que se podría interpretar como comportamiento a largo plazo.

Cada observación va ligada a un valor del tiempo, lo que permite plantear un modelo del tipo

$$Y_{(t)} = \phi_{(t)} + \varepsilon$$

donde la función $\phi_{(t)}$ puede ser:

- lineal: $\phi_{(t)} = \alpha_0 + \alpha_1 t$
- polinómica: $\phi_{(t)} = \alpha_0 + \alpha_1 t + \alpha_1 t^2 + \alpha_1 t^3 + ...$
- exponencial: $\phi_{(t)} = \alpha_0 e^{\alpha 1 t}$

Si la serie no presenta estacionalidad, el método de estimación mínimo-cuadrática y todas las pruebas de hipótesis relativas a la explicación del modelo y a la significación de los coeficientes estimados, propios del modelo lineal ordinario, permiten estimar los coeficientes del modelo de tendencia sobre los datos directos.

Caso de existir componente estacional, para que ésta no enmascare la tendencia, es necesario estabilizar previamente la serie.

12.4.1 Medias móviles: tendencia

Con este método se consiguen suavizar tanto las oscilaciones periódicas de una serie como las aleatorias. Su aplicación requiere decidir, previamente, el período en que se repite cierto patrón de comportamiento, que pueda atribuirse a variaciones estacionales; la observación de la evolución gráfica de la serie puede ayudar a tomar la decisión.

Una vez fijado el período **p**, se calculan las medias de los valores de la serie tomados de **p** en **p**, sucesivamente desde el inicio. Asociando cada una de estas medias al valor del tiempo del punto central del período estudiado, se obtiene una nueva serie de valores mucho más estables, debido, por una parte, a la reducción de la variabilidad ocasionada al promediar y, por otra, a que, si el período escogido es el correcto, al pasar de una media móvil a la siguiente, el nuevo dato incorporado es del mismo comportamiento que el dato saliente.

12.4.2 Estacionalidad

La componente estacional, que provoca una oscilación sistemática de período corto, generalmente no superior al año, puede enmascarar la evolución a largo plazo, tendencia, si no se aísla convenientemente.

Se entiende como componente estacional, en modelos aditivos, la diferencia entre el valor de la estación y la media de todas las estaciones componentes del periodo; en modelos multiplicativos igual pero el cociente en vez de diferencia.

El análisis de la estacionalidad queda ligado al método que se decida emplear para modelizar la tendencia; así, en este punto estudiaremos la situación para el caso de trabajar con medias móviles.

Para calcular los valores de los Índices estacionales hay que seguir la siguiente sistemática:

1. Calcular las **medias móviles**, sobre los datos de la serie original, tomando el período de agrupación, **p**, que se considere oportuno.
2. Proponer un modelo de agrupación de las componentes, **aditivo** o **multiplicativo**.
3. Separar la parte explicada por la tendencia. Supuesto el modelo aditivo, esto equivale a calcular la **diferencia** (W) entre los valores originales y el resultado de aplicarle la media móvil; si fuese multiplicativo, en lugar de diferencias serian **cocientes**. Hay que destacar que en W están incluidas las componentes asociadas a la estacionalidad, los ciclos y los residuos.
4. Asumiendo que los residuos son variables aleatorias de media nula y que la componente cíclica, caso de existir, es de período suficientemente largo como para no ser recogida por los datos, se procede a evaluar la estacionalidad asociada a cada componente del período. Para ello se calculan los **promedios** de W de la misma estación y se **resta** después a cada uno de ellos la estacionalidad media en el caso aditivo, o el **cociente** en el multiplicativo.

12.5 Suavizado exponencial

Cuando la serie presenta componente estacional y tendencia que se mantienen de forma sostenida a lo largo de todo el período de recogida de datos, se han expuesto dos formas de modelizarla y poder hacer previsiones: la descomposición clásica y las variables categóricas.

Sin embargo, son frecuentes las situaciones en que la tendencia, caso de existir, puede ser difícil modelizarla a través de un simple modelo polinómico de menor o mayor grado. Podría entonces pensarse en un modelo de evolución que cambiase a lo largo del tiempo; en estos casos las técnicas asociadas a la metodología de la ponderación exponencial son útiles para hacer previsiones sobre la evolución futura.

12.5.1 Suavizado exponencial

La ponderación exponencial, o suavizado exponencial, es otra técnica destinada también a estabilizar la serie, eliminando en lo posible la influencia del componente aleatorio. Para ello se construye una nueva serie, la serie suavizada S_t, a partir de los datos iniciales, Y_t, de manera que:

$$S_t = \lambda\, Y_t + (1-\lambda)\, S_{t-1} \qquad \text{con } 0 < \lambda < 1$$

Para que la serie suavizada quede definida, es necesario concretar los valores de S_0, que generalmente se considera igual a Y_1, y el del coeficiente de ponderación λ. En la selección del valor de λ se pueden emplear distintos criterios de minimización de errores, que se expondrán a continuación.

Teniendo en cuenta que tal como hemos definido S_t, tendremos que:

$$S_{t-1} = \lambda\, Y_t + (1-\lambda)\, S_{t-2}$$
$$S_{t-2} = \lambda\, Y_{t-2} + (1-\lambda)\, S_{t-3}$$
$$\ldots\ldots\ldots\ldots$$
$$S_1 = \lambda\, Y_1 + (1-\lambda)\, S_0$$
$$S_0 = Y_1$$

sustituyendo repetitivamente S_{t-1}, S_{t-2},... por su expresión de St, se obtiene:

$$S_t = \lambda\, Y_t + (1-\lambda)\,[\lambda\, Y_{t-1} + (1-\lambda)\,[\lambda\, Y_{t-2} + (1-\lambda)\,\ldots..\,[\lambda\, Y_1 + (1-\lambda)\, Y_1]]]]]$$

El valor de S_t es la previsión para el tiempo siguiente, es decir:

$$\hat{Y}_{(t+1)} = S_t$$

El análisis de la expresión anterior permite interpretar este tipo de suavizado, de forma que el valor de Y previsto para el período t+1, es decir S_t, se obtenga como promedio ponderado de los valores reales que ha presentado la serie cronológica desde el inicio de la recogida de información. La discrepancia entre los valores obtenidos y los previstos, Y_{t+1}-S_t, es atribuible en parte al componente aleatorio y, posiblemente, a cambios bruscos en el comportamiento de la serie.

El coeficiente de ponderación λ juega el siguiente papel: cuanto mayor sea su valor, tanto más peso se dará a los valores recientes, en detrimento de los antiguos; mientras que valores de λ próximos a cero dan gran peso a la historia y poca importancia a los valores próximos.

Así, si la serie se mantiene estable, serán interesantes valores pequeños del coeficiente de ponderación ya que amortiguarán fuertemente la oscilación aleatoria, mientras que si la serie presentara cambios bruscos, la serie suavizada tardaría mucho en detectarlos si su λ fuese pequeña, mientras que respondería prontamente a ellos con valores altos del coeficiente λ.

Analizando la expresión del valor suavizado, para distintos valores de λ, se puede escribir, por ejemplo,

$$(\lambda = 0{,}10) \Rightarrow \hat{Y}_5 = S_4 = 0{,}10\, Y_4 + 0{,}09\, Y_3 + 0{,}081\, Y_2 + 0{,}729\, Y_1$$
$$(\lambda = 0{,}50) \Rightarrow \hat{Y}_5 = S_4 = 0{,}50\, Y_4 + 0{,}25\, Y_3 + 0{,}125\, Y_2 + 0{,}125\, Y_1$$
$$(\lambda = 0{,}90) \Rightarrow \hat{Y}_5 = S_4 = 0{,}90\, Y_4 + 0{,}09\, Y_3 + 0{,}009\, Y_2 + 0{,}001\, Y_1$$

Es decir, con un valor del factor de ponderación de 0,10, la previsión para t = 5 está constituida por un 10% del valor observado en t = 4, un 9% del de t = 3, un 8,1% del de t = 2 y un 72,9 % del de t = 1; o sea, con un valor pequeño de λ, la previsión está constituida mayoritariamente por el valor más antiguo.

Cuando λ es igual a 0,50, los pesos aplicados a cada valor recogido están más uniformemente repartidos y, cuando λ es grande, por ejemplo 0,90, el mayor componente de la previsión es el último valor observado; los demás tendrán un valor de ponderación tanto más pequeño cuanto más alejados estén en el tiempo.

El suavizado exponencial puede verse como un método alternativo a las medias móviles, con sus ventajas e inconvenientes.

Entre las primeras hay que citar que con la ponderación exponencial no se pierde ninguna información, al contrario que con las medias móviles, pues cuanto mayor era la longitud del período a promediar, tanta más información se perdía, en el inicio y en el fin de la serie.

Además una serie con cambios de tendencia, más o menos bruscos, se puede modelizar por suavizado exponencial y no podría hacerse ni por descomposición ni por variables categóricas. Por el contrario, si la serie presenta estacionalidad con las medias móviles, siempre que se escoja correctamente el período, ésta desaparece totalmente y da lugar a una serie estabilizada que permite modelizar directamente la tendencia, hecho que no ocurre con la ponderación exponencial simple, que no es capaz de suavizar la oscilación debida a la estacionalidad.

Para solucionar este inconveniente, se han desarrollado técnicas basadas en el suavizado exponencial, que permiten incorporar un modelo de tendencia o bien una componente estacionaria; éstas son las técnicas de Brown, para el primer caso, o de Winters para el segundo.

12.5.2 Selección del factor de ponderación

Tal como se ha expuesto, en función del valor de λ, se puede dar mayor o menor peso a la historia, y detectar con más o menos rapidez cambios bruscos en la serie; es por ello que la selección del valor más adecuado para el factor de ponderación es crucial en el éxito de la modelización de la serie y la previsión de valores futuros.

Todos los métodos utilizados para esta selección se basan en minimizar alguna función de los errores de ponderación.

Los errores más destacables son:

- *Error cuadrático medio*: promedio de los cuadrados de los errores de previsión:

$$MSE = \frac{\sum_{t=2}^{n} (Y_t - S_t)^2}{n-1}$$

- *Error absoluto medio*: promedio de los valores absolutos de los errores de previsión:

$$MAE = \frac{\sum_{t=2}^{n} |Y_t - S_t|}{n-1}$$

Hay que insistir en que en una serie en la que el tiempo es t = 1, 2,..., n, el suavizado exponencial no ofrece ninguna previsión para t = 1, y, por tanto, no existe error de previsión en este punto; consecuentemente, en este caso los errores siempre son promedios de n -1 valores.

En general, se selecciona aquel valor de l para el cual los valores del error absoluto medio y del cuadrático medio, MAE y MSE, alcancen los valores más bajos.

12.6 PROBLEMAS

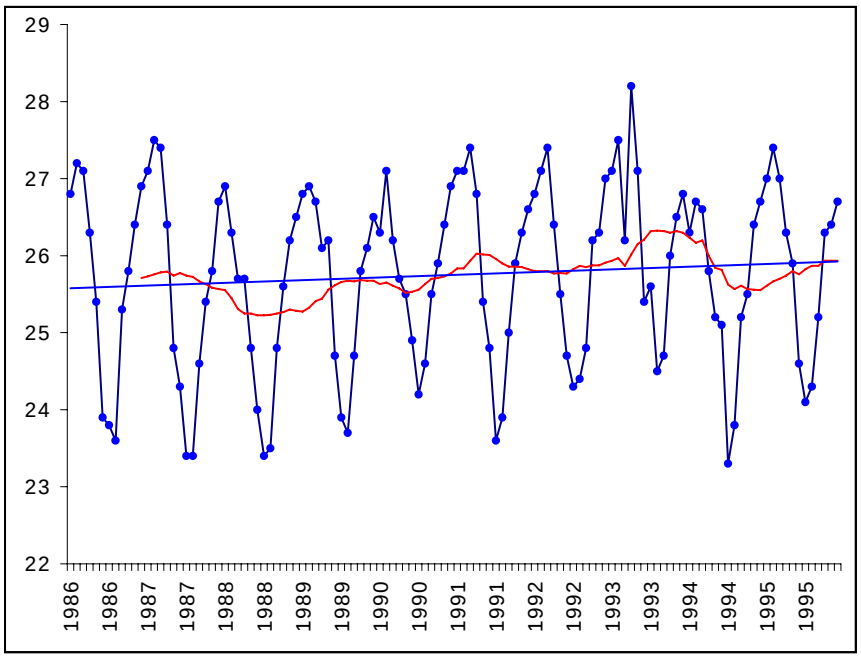
12.6.1 Ajustar un modelo aditivo a las siguientes series de datos correspondientes a las temperaturas mensuales de una ciudad del hemisferio sur.

	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995
Enero	26,8	27,1	26,9	26,8	26,3	27,1	26,8	27,1	26,3	27,0
Febrero	27,2	27,5	26,3	26,9	27,1	27,1	27,1	27,5	26,7	27,4
Marzo	27,1	27,4	25,7	26,7	26,2	27,4	27,4	26,2	26,6	27,0
Abril	26,3	26,4	25,7	26,1	25,7	26,8	26,4	28,2	25,8	26,3
Mayo	25,4	24,8	24,8	26,2	25,5	25,4	25,5	27,1	25,2	25,9
Junio	23,9	24,3	24,0	24,7	24,9	24,8	24,7	25,4	25,1	24,6
Julio	23,8	23,4	23,4	23,9	24,2	23,6	24,3	25,6	23,3	24,1
Agosto	23,6	23,4	23,5	23,7	24,6	23,9	24,4	24,5	23,8	24,3
Septiembre	25,3	24,6	24,8	24,7	25,5	25,0	24,8	24,7	25,2	25,2
Octubre	25,8	25,4	25,6	25,8	25,9	25,9	26,2	26,0	25,5	26,3
Noviembre	26,4	25,8	26,2	26,1	26,4	26,3	26,3	26,5	26,4	26,4
Diciembre	26,9	26,7	26,5	26,5	26,9	26,6	27,0	26,8	26,7	26,7

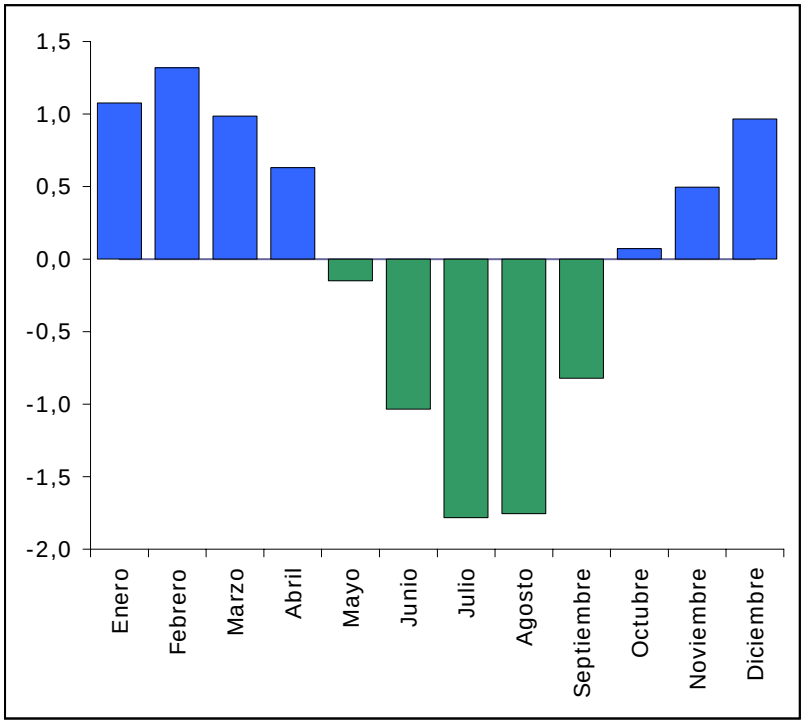
Los pasos a seguir serán

1. Representar la serie;
2. Confirmar la idea de que se trata de un modelo aditivo y no multiplicativo;
3. Aislar el componente estacional mensual representándolo gráficamente.
4. Decidir si sobre la serie suavizada por medias móviles parece existir una tendencia. En su caso modelizarla.
5. Construir el modelo
6. Calcular los residuos y representarlos
7. Juzgar la validez del modelo.
8. Predecir valores para el año siguiente

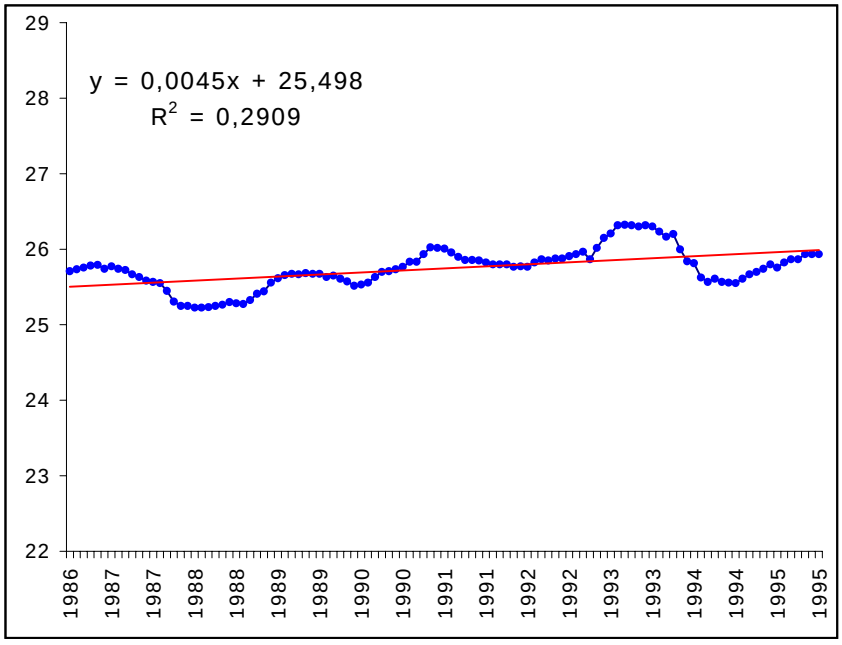
Representación de la serie



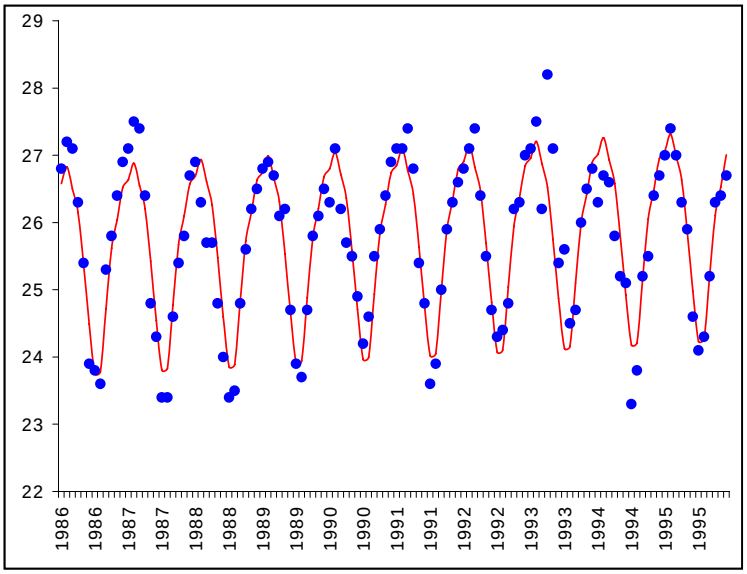
Componente estacional



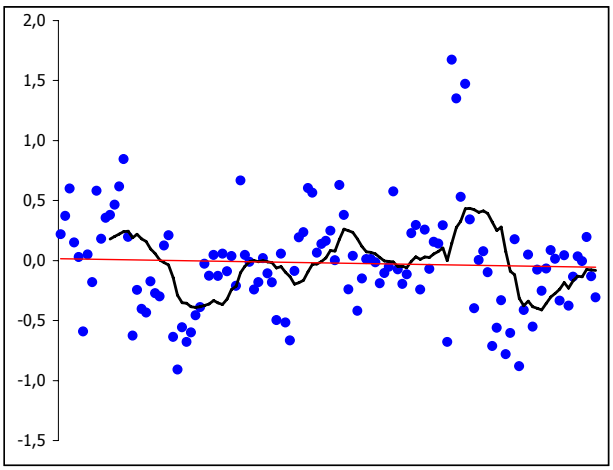
Se observa una tendencia en la serie desestacionalizada



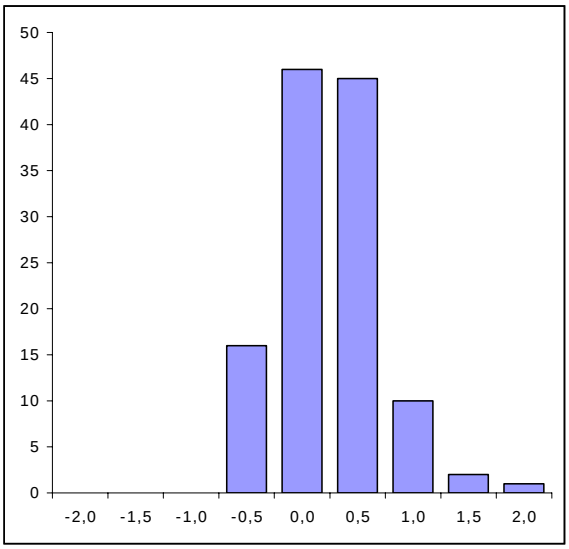
El modelo final es bueno



Los residuos no muestran patrón apreciabla



Los residuos son pequeños



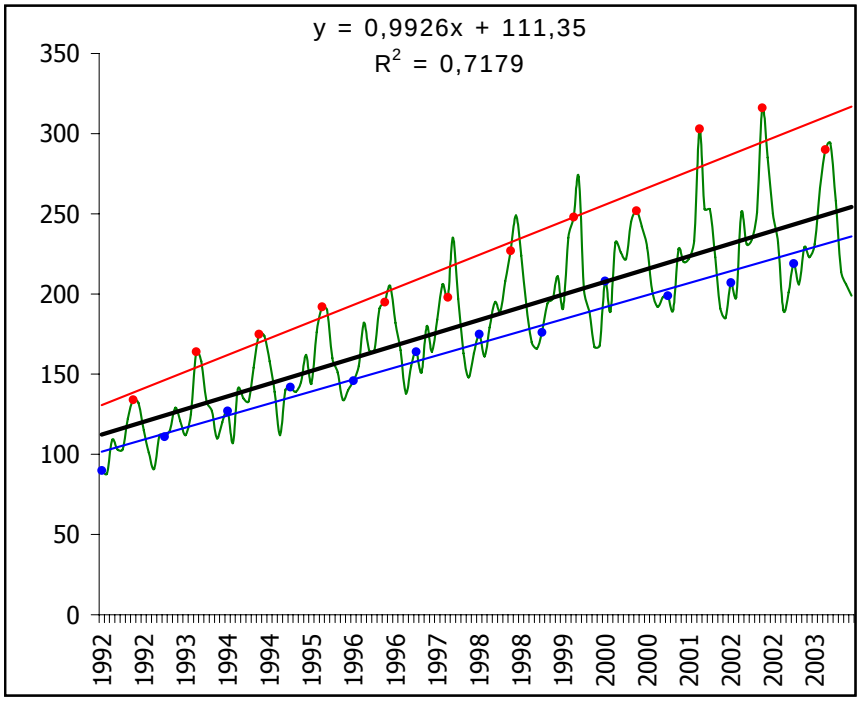
12.6.2 Ajustar un modelo a la siguiente serie de datos

	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003
Enero	90	111	127	142	146	164	175	176	208	199	207	219
Febrero	88	115	107	139	155	151	161	194	189	190	198	206
Marzo	109	129	141	145	182	180	179	197	232	228	251	229
Abril	103	121	135	162	165	164	195	211	226	220	231	223
Mayo	103	112	133	144	165	184	189	191	222	222	234	231
Junio	122	125	154	176	191	206	208	235	245	233	251	266
Julio	134	164	175	192	195	198	227	248	252	303	316	290
Agosto	132	158	174	190	205	235	249	273	242	253	285	294
Septiembre	115	133	158	160	182	197	224	202	229	253	250	258
Octubre	101	127	139	151	165	163	193	189	202	223	232	214
Noviembre	91	110	112	134	138	148	170	167	192	191	190	206
Diciembre	112	120	140	140	155	163	166	168	198	185	201	199

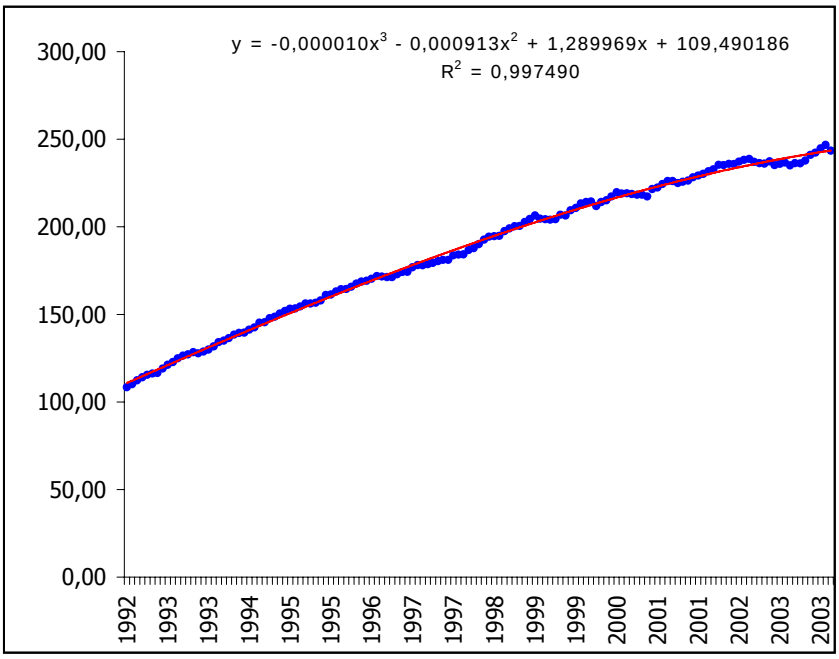
Los pasos a seguir serán

1. Representar la serie;
2. Incluir en el gráfico anterior los valores de los meses de Enero y Julio por separado junto con el total de los datos.
3. Proponer un modelo aditivo o multiplicativo;
4. Aislar el componente estacional mensual representándolo gráficamente.
5. Decidir si sobre la serie suavizada por medias móviles parece existir una tendencia. En su caso modelizarla.
6. Construir el modelo
7. Calcular los residuos y representarlos
8. Juzgar la validez del modelo.

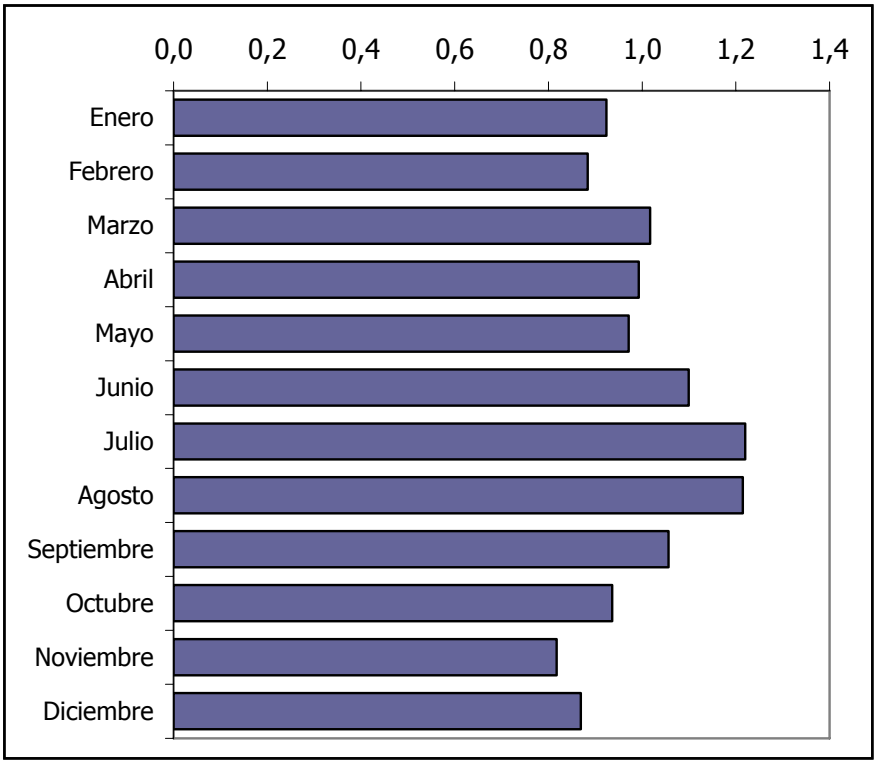
Representación gráfica



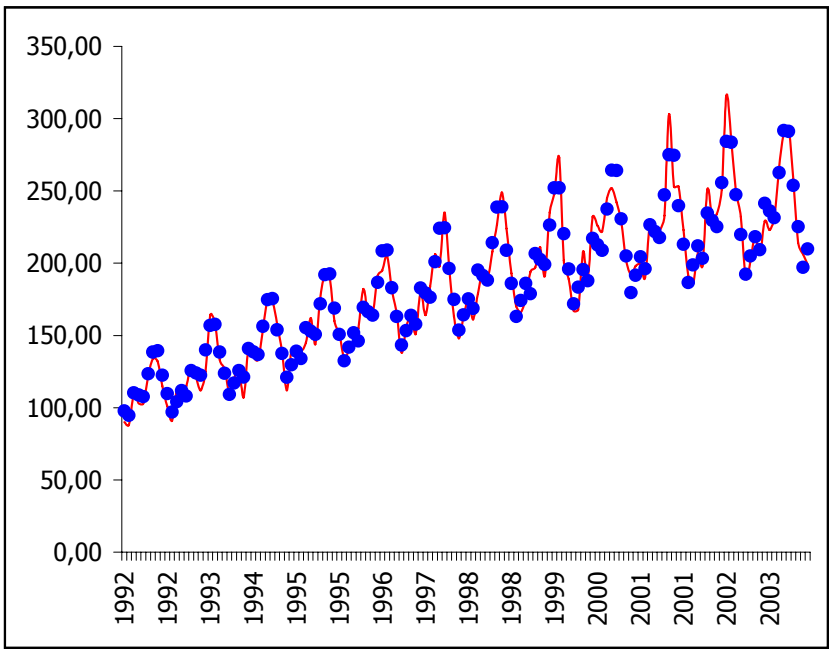
Tendencia polinómica



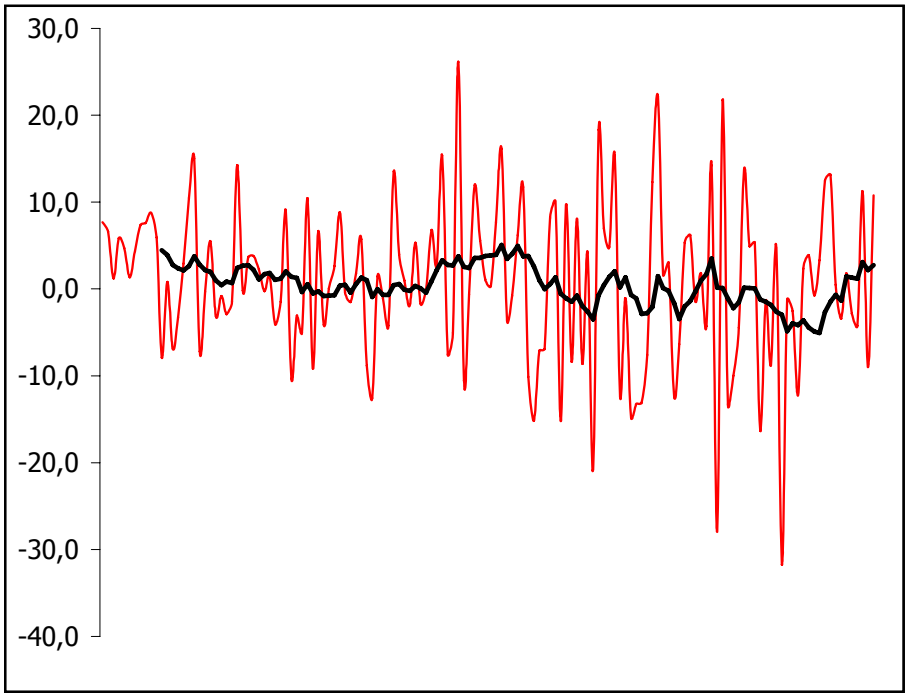
Componentes estacionales



Modelo final



Residuos del modelo



12.6.3 Construir un modelo para los siguientes datos de ventas

AÑO	CUATRIMESTRE	VENTAS
1990	1	40,22
1990	2	54,89
1990	3	63,51
1990	4	111,4
1991	1	46,95
1991	2	51,62
1991	3	61,47
1991	4	108,6
1992	1	41,38
1992	2	65,3
1992	3	64,25
1992	4	113,8
1993	1	53,34
1993	2	59,37
1993	3	66,15
1993	4	121,5
1994	1	67,38
1994	2	56,09
1994	3	75,11
1994	4	124,4
1995	1	55,9
1995	2	61,25
1995	3	75,44
1995	4	126,5

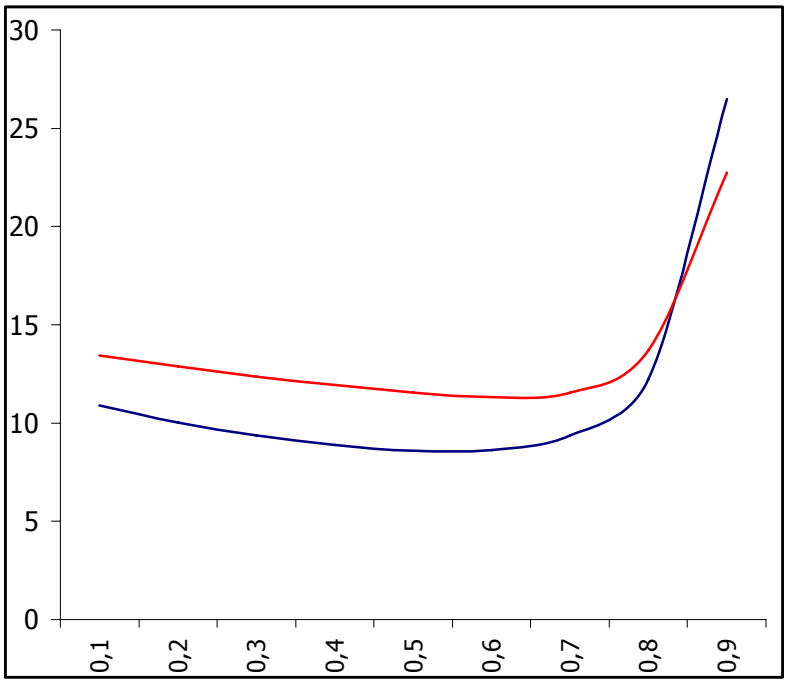
12.6.4 Construir un modelo para los siguientes datos de periodicidad semanal

Sem	Día	Y	Sem	Día	Y	Sem	Día	Y
1	lunes	3968	5	viernes	3618	9	jueves	2979
1	martes	4572	5	lunes	4427	9	viernes	3971
1	miércoles	3964	5	martes	4314	9	lunes	3291
1	jueves	6326	5	miércoles	5616	9	martes	5336
1	viernes	9673	5	jueves	8778	9	miércoles	8392
1	lunes	8307	5	viernes	7322	9	jueves	6790
2	martes	3593	6	lunes	2899	10	viernes	3539
2	miércoles	5367	6	martes	4918	10	lunes	4694
2	jueves	3763	6	miércoles	4226	10	martes	3120
2	viernes	6703	6	jueves	6025	10	miércoles	6026
2	lunes	9485	6	viernes	8712	10	jueves	7792
2	martes	8207	6	lunes	7685	10	viernes	7294
3	miércoles	3717	7	martes	3408	11	lunes	3254
3	jueves	4712	7	miércoles	4869	11	martes	4725
3	viernes	3538	7	jueves	3589	11	miércoles	3227
3	lunes	5758	7	viernes	5437	11	jueves	5588
3	martes	9112	7	lunes	8239	11	viernes	8320
3	miércoles	7501	7	martes	7360	11	lunes	6995
4	jueves	3108	8	miércoles	2915	12	martes	3229
4	viernes	4771	8	jueves	4237	12	miércoles	4648
4	lunes	3643	8	viernes	3679	12	jueves	3450
4	martes	6616	8	lunes	6060	12	viernes	5129
4	miércoles	8907	8	martes	8755	12	lunes	8159
4	jueves	7993	8	miércoles	7475	12	martes	6923

12.6.5 Para el siguiente conjunto de datos

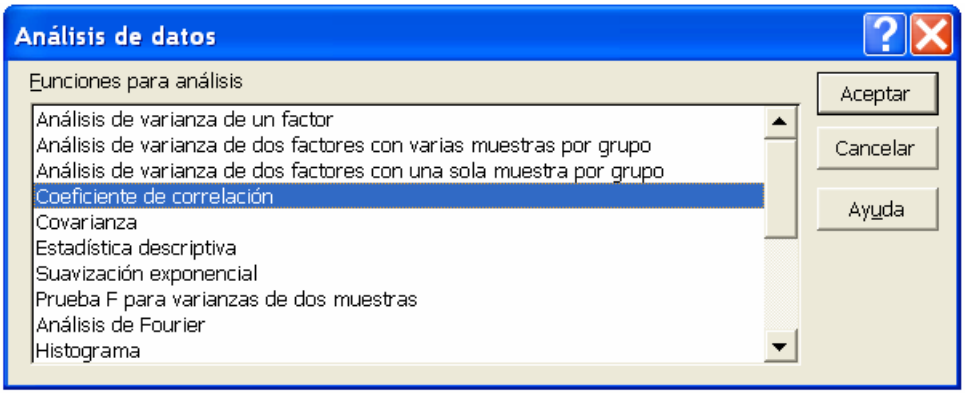
9,958	12,674	18,748	27,317	31,732
10,096	17,504	20,800	26,122	31,538
11,552	13,462	21,683	29,837	32,175
9,113	16,945	27,069	28,854	35,543
13,898	18,653	23,728	27,129	35,534
11,487	18,942	24,890	30,194	37,336
11,114	15,084	26,132	34,104	
9,505	16,568	24,663	28,448	
17,934	20,733	25,217	35,726	
12,339	26,267	24,653	30,602	
16,510	20,401	28,062	31,011	

a) Calcular el modelo de suavizado exponencial para $\lambda \in \{0,1 ; 0,2 ; ....; 0,9\}$
b) Encontrar el valor de λ que minimiza el error MSE para $\lambda \in \{0,1 ; 0,2 ; ....; 0,9\}$
c) Encontrar el valor de λ que minimiza el error MAE para $\lambda \in \{0,1 ; 0,2 ; ....; 0,9\}$
d) Encontrar el valor de λ que minimiza el error MAE,MSE para $0 \leq \lambda \leq 1$



13 Herramientas de análisis estadístico

Excel proporciona un conjunto de herramientas para el análisis de los datos denominado **Análisis de Datos** que podrá utilizar para ahorrar pasos en el desarrollo de análisis estadísticos. Cuando utilice una de estas herramientas, deberá proporcionar los datos y parámetros para cada análisis; la herramienta utilizará las funciones de macros estadísticas o técnicas correspondientes y, a continuación, mostrará los resultados en una tabla de resultados. Algunas herramientas generan gráficos además de tablas de resultados.



Para ver una lista de las herramientas de análisis, elija Análisis de datos en el menú Herramientas. Si este comando no está en el menú, ejecute el programa de instalación para instalar las Herramientas para análisis de la forma siguiente :

Activar las Herramientas para análisis

a) En el menú **Herramientas**, elija Macros automáticas. Si en la lista del cuadro de diálogo Macros automáticas no aparece Herramientas para análisis, haga clic en el botón "Examinar" y busque la unidad, directorio y archivo de la macro automática Herramientas para análisis, o bien ejecute el programa de instalación.
b) Seleccione la casilla de verificación **"Herramientas para análisis"**. Las macros automáticas que instale en Microsoft Excel permanecerán activas hasta que las quite.

13.1 Descripción de las herramientas

13.1.1 Análisis de la Varianza

Las herramientas de análisis de varianza proporcionan distintos tipos de análisis de la varianza. La herramienta que vaya a usarse dependerá del número de factores y del número de muestras que tenga de la población que desee comprobar.

- Varianza de un factor Esta herramienta realiza un análisis simple de varianza, comprobando la hipótesis según la cual dos o más muestras (extraídas de poblaciones con la misma media) son iguales. Esta técnica profundiza en las pruebas para dos medias como, por ejemplo, la Prueba t.
- Varianza de dos factores con varias muestras por grupo Esta herramienta de análisis realiza una extensión del análisis de la varianza de un factor que contiene más de una muestra por cada grupo de datos.

- Varianza de dos factores con una sola muestra por grupo Esta herramienta de análisis realiza un análisis de varianza de dos factores con una sola muestra por grupo, comprobando la hipótesis según la cual, las medias de dos o más muestras son iguales (extraídas de poblaciones con la misma media). Esta técnica profundiza en las pruebas para dos medias como, por ejemplo, la Prueba t.

13.1.2 Correlación

La herramienta de análisis Correlación mide la relación entre dos conjuntos de datos medidos para que sean independientes de la unidad de medida. El cálculo de la correlación de población devuelve la covarianza de dos conjuntos de datos dividida por el producto de sus desviaciones estándar

Puede utilizar la herramienta de análisis de correlación para determinar si dos conjuntos de datos varían conjuntamente, es decir, si los valores altos de un conjunto están asociados con los valores altos del otro (correlación positiva), si los valores bajos de un conjunto están asociados con los valores bajos del otro (correlación negativa) o si los valores de ambos conjuntos no están relacionados (correlación con tendencia a cero).

13.1.3 Covarianza

La covarianza es una medida de la relación entre dos rangos de datos. La herramienta de análisis Covarianza, devuelve el promedio de los productos entre las desviaciones de los puntos de datos con respecto a sus medias respectivas.

13.1.4 Estadística descriptiva

La herramienta de análisis Estadística descriptiva genera un informe estadístico de una sola variable para los datos del rango de entrada, y proporciona información acerca de la tendencia central y dispersión de los datos.

13.1.5 Suavización exponencial

La herramienta de análisis Suavización exponencial predice un valor basándose en el pronóstico del período anterior, ajustándose al error en ese pronóstico anterior. La herramienta utiliza la constante de suavización *a*, cuya magnitud determina la exactitud con la que los pronósticos responden a los errores en el pronóstico anterior

13.1.6 Prueba t para varianzas de dos muestras

La herramienta de análisis Prueba t para varianzas de dos muestras ejecuta una Prueba t de dos muestras para comparar dos varianzas de población.

13.1.7 Análisis de Fourier

La herramienta Análisis de Fourier resuelve problemas de sistemas lineales y analiza datos periódicos, transformándolos mediante el método de transformación rápida de Fourier (FFT, Fast Fourier Transform). Esta herramienta también realiza transformaciones inversas, en las que el inverso de los datos transformados devuelve los datos originales.

13.1.8 Histograma

La herramienta de análisis Histograma calcula las frecuencias individuales y acumulativas de rangos de celdas de datos y de clases de datos. Esa herramienta genera datos acerca del número de apariciones de un valor en un conjunto de datos.

13.1.9 Media móvil

La herramienta de análisis Media móvil proyecta valores en el período de pronósticos, basándose en el valor promedio de la variable calculada durante un número específico de períodos anteriores. Una media móvil proporciona información de tendencias que se vería enmascarada por una simple media de todos los datos históricos.

13.1.10 Generación de números aleatorios

La herramienta de análisis Generación de números aleatorios rellena un rango con números aleatorios independientes extraídos de una de varias distribuciones.

13.1.11 Jerarquía y percentil

La herramienta de análisis Jerarquía y percentil crea una tabla que contiene los rangos ordinales y porcentuales de cada valor de un conjunto de datos. Puede analizar la importancia relativa de los valores en un conjunto de datos.

13.1.12 Regresión

La herramienta de análisis Regresión realiza un análisis de regresión lineal utilizando el método de los "mínimos cuadrados" para ajustar una línea a una serie de observaciones. Puede utilizar esta herramienta para analizar la forma en que los valores de una o más variables independientes afectan a una variable dependiente.

13.1.13 Muestreo

La herramienta de análisis Muestreo crea una muestra de población tratando el rango de entrada como una población. Cuando la población sea demasiado grande para procesarla o para presentarla gráficamente, puede utilizarse una muestra representativa. Además, si cree que los datos de entrada son periódicos, puede crear una muestra que contenga únicamente los valores de una parte determinada de un ciclo.

13.1.14 Prueba t

Las herramientas de análisis Prueba t permiten comparar las medias de poblaciones bajo diferentes hipótesis.

- Prueba t para dos muestras suponiendo varianzas iguales Esta herramienta de análisis ejecuta una prueba t de Student en dos muestras. En este tipo de prueba se supone que las varianzas de ambos conjuntos de datos son iguales, y se conoce con el nombre de prueba t homoscedástica.
- Prueba t para dos muestras suponiendo varianzas desiguales Esta herramienta de análisis ejecuta una prueba t de Student en dos muestras. En este tipo de prueba se supone que las varianzas de ambos rangos son desiguales, y se conoce con el nombre de prueba t heteroscedástica.
- Prueba t para medias de dos muestras emparejadas Esta herramienta de análisis y su fórmula ejecutan una prueba t de Student de dos muestras emparejadas para determinar si las medias de la muestra son diferentes. En este tipo de prueba no se supone que las varianzas de ambas poblaciones sean iguales.

13.1.15 Prueba z

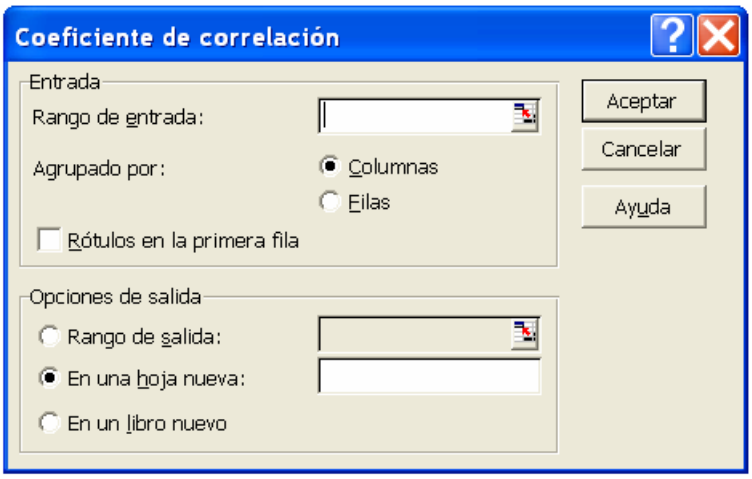
La herramienta de análisis Prueba z para medias de dos muestras realiza una prueba z en las medias de dos muestras con varianzas conocidas. Esta herramienta se utiliza para comprobar las hipótesis acerca de la diferencia entre dos medias de población.

13.2 Análisis de la varianza.

Ver apartado correspondiente.

13.3 Correlación

Devuelve la **matriz de correlaciones** para un conjunto de variables



- Rango de entrada. Introduzca la referencia de celda correspondiente al rango de datos que desee analizar. La referencia deberá contener dos o más rangos adyacentes organizados en columnas o filas.
- Agrupado por. Haga clic en el botón Filas o Columnas para indicar si los datos del rango de entrada están organizados en filas o en columnas.
- Rótulos en la primera fila y rótulos en la primera columna. Si la primera fila del rango de entrada contiene rótulos, active la casilla de verificación Rótulos en la primera fila. Si los rótulos están en la primera columna del rango de entrada, active la casilla de verificación Rótulos en la primera columna. Esta casilla de verificación estará desactivada si el rango de entrada carece de rótulos; Microsoft Excel generará los rótulos de datos correspondientes para la tabla de resultados.
- Rango de salida. Introduzca la referencia correspondiente a la celda superior izquierda de la tabla de resultados. Excel sólo completará media tabla ya que la correlación entre dos rangos de datos es independiente del orden en que se procesen dichos rangos. Las celdas de la tabla de resultados con coordenadas de filas y de columnas iguales contendrán el valor 1, ya que cada conjunto de datos está perfectamente correlacionado consigo mismo.
- En una hoja nueva. Haga clic en esta opción para insertar una hoja nueva en el libro actual y pegar los resultados comenzando por la celda A1 de la nueva hoja de cálculo. Para darle un nombre a la nueva hoja de cálculo, escríbalo en el cuadro.
- En un libro nuevo. Haga clic en esta opción para crear un nuevo libro y pegar los resultados en una hoja nueva del libro creado.

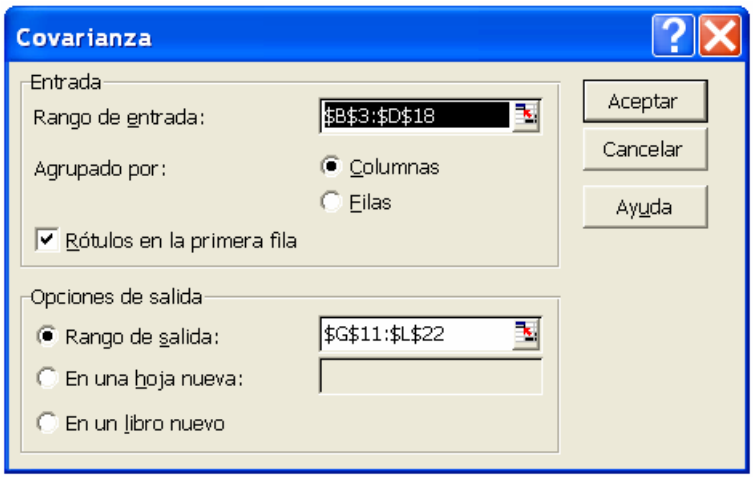
RESULTADO

A	B	C
0,288	0,701	0,057
0,374	0,947	0,313
0,850	0,164	0,594
0,334	0,659	0,521
0,212	0,960	0,087
0,034	0,008	0,835
0,068	0,938	0,529
0,654	0,948	0,105
0,780	0,638	0,585
0,917	0,762	0,221
0,775	0,143	0,783
0,378	0,497	0,484
0,926	0,336	0,784
0,991	0,887	0,533
0,015	0,694	0,623

	A	B	C
A	1,00000		
B	-0,12794	1,00000	
C	0,08240	-0,72442	1,00000

13.4 Covarianza

Calcula la **matriz de varianzas covarianzas** de un conjunto de variables.



- Rango de entrada. Introduzca la referencia de celda correspondiente al rango de datos que desee analizar. La referencia deberá contener dos o más rangos adyacentes organizados en columnas o filas.
- Agrupado por. Haga clic en el botón Filas o Columnas para indicar si los datos del rango de entrada están organizados en filas o en columnas.
- Rótulos en la primera fila y rótulos en la primera columna. Si la primera fila del rango de entrada contiene rótulos, active la casilla de verificación Rótulos en la primera fila. Si los rótulos están en la primera columna del rango de entrada, active la casilla de verificación Rótulos en la primera columna. Esta casilla de verificación estará desactivada si el rango de entrada carece de rótulos; Microsoft Excel generará los rótulos de datos correspondientes para la tabla de resultados.
- Rango de salida. Introduzca la referencia correspondiente a la celda superior izquierda de la tabla de resultados. Excel sólo completará media tabla ya que la covarianza entre dos rangos de datos es independiente del orden en que se

procesen dichos rangos. La diagonal de la tabla contiene la varianza de todos los rangos.

- En una hoja nueva. Haga clic en esta opción para insertar una hoja nueva en el libro actual y pegar los resultados comenzando por la celda A1 de la nueva hoja de cálculo. Para darle un nombre a la nueva hoja de cálculo, escríbalo en el cuadro.
- En un libro nuevo. Haga clic en esta opción para crear un nuevo libro y pegar los resultados en una hoja nueva del libro creado.

RESULTADO

A	B	C
0,288	0,701	0,057
0,374	0,947	0,313
0,850	0,164	0,594
0,334	0,659	0,521
0,212	0,960	0,087
0,034	0,008	0,835
0,068	0,938	0,529
0,654	0,948	0,105
0,780	0,638	0,585
0,917	0,762	0,221
0,775	0,143	0,783
0,378	0,497	0,484
0,926	0,336	0,784
0,991	0,887	0,533
0,015	0,694	0,623

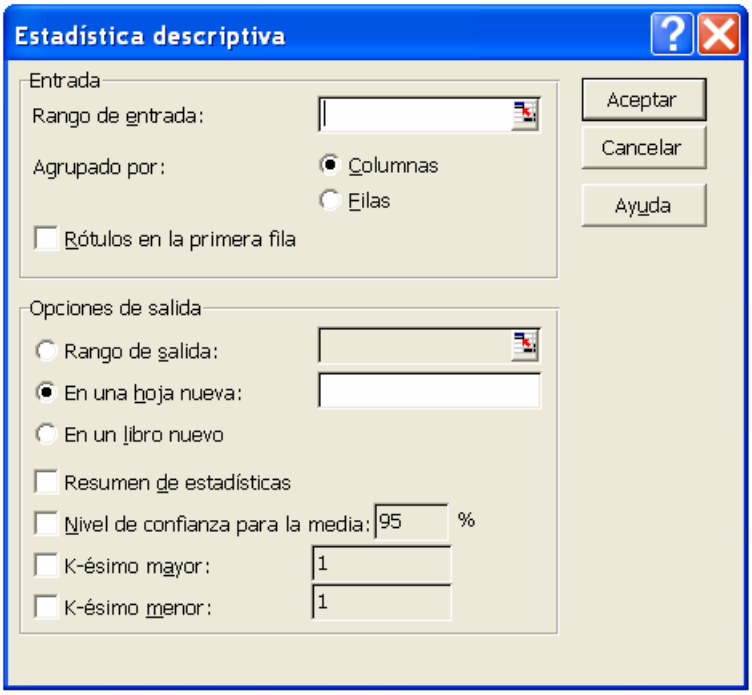
	A	B	C
A	1,00000		
B	-0,12794	1,00000	
C	0,08240	-0,72442	1,00000

	A	B	C
A	0,11478		
B	-0,01344	0,09609	
C	0,00695	-0,05590	0,06197

13.5 Estadística descriptiva

Calcula los estadísticos básicos de un conjunto de datos, para una o varias variables.

- Nivel de confianza para la media Active esta casilla si desea incluir una fila correspondiente al nivel de confianza de la media en la tabla de resultados. En el cuadro, escriba el nivel de confianza que desee utilizar. Por ejemplo, un valor de 95 % calculará el nivel de confianza de la media con un nivel de importancia del 5 %.
- K-ésimo mayor. Active esta casilla si desea incluir una fila correspondiente al valor k-ésimo mayor de cada rango de datos en la tabla de resultados. En el cuadro, escriba el número que va a utilizarse para k. Si escribe 1, esta fila contendrá el máximo del conjunto de datos.
- K-ésimo menor. Active esta casilla si desea incluir una fila correspondiente al valor k-ésimo menor de cada rango de datos en la tabla de resultados. En el cuadro, escriba el número que va a utilizarse para k. Si escribe 1, esta fila contendrá el mínimo del conjunto de datos.
- Resumen de estadísticas. Seleccione esta opción si desea que Excel genere un campo en la tabla de resultados por cada una de las siguientes variables estadísticas: **Media**, **Error típico** (de la media), **Mediana**, **Moda**, **Desviación estándar**, **Varianza**, **Curtosis**, **Coeficiente de asimetría**, **Rango**, **Mínimo**, **Máximo**, **Suma**, **Cuenta**, **Mayor (#)**, **Menor (#)** y **Nivel de confianza**.



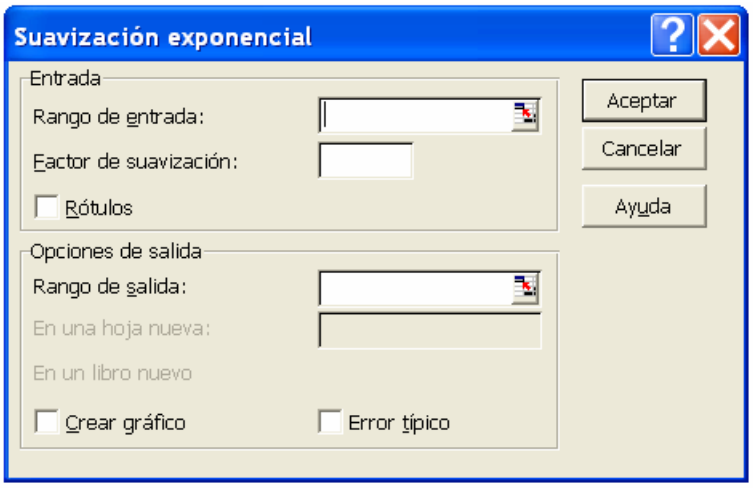
RESULTADO

A	B	C
0,288	0,701	0,057
0,374	0,947	0,313
0,850	0,164	0,594
0,334	0,659	0,521
0,212	0,960	0,087
0,034	0,008	0,835
0,068	0,938	0,529
0,654	0,948	0,105
0,780	0,638	0,585
0,917	0,762	0,221
0,775	0,143	0,783
0,378	0,497	0,484
0,926	0,336	0,784
0,991	0,887	0,533
0,015	0,694	0,623

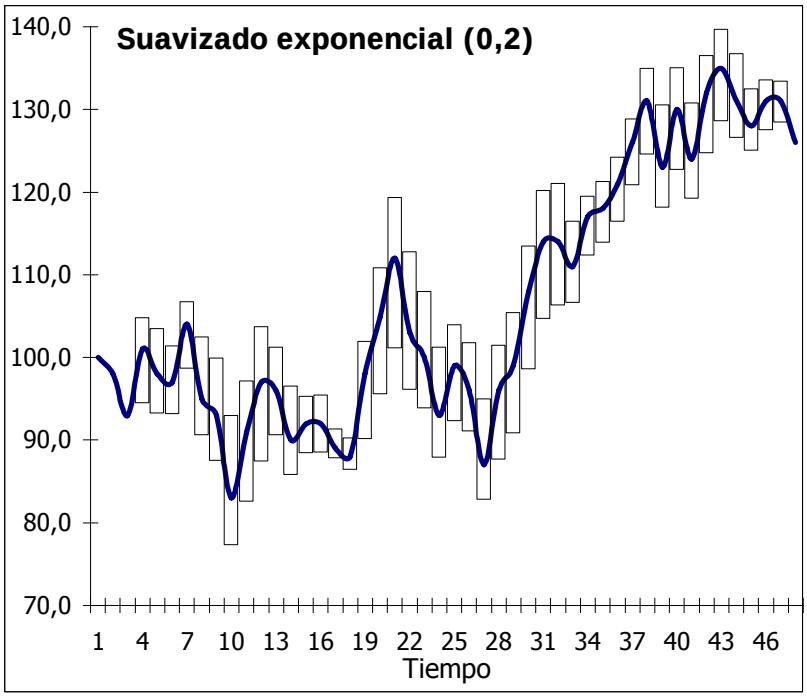
	A	B	C
Media	0,50647	0,61875	0,47019
Error típico	0,09055	0,08285	0,06653
Mediana	0,37794	0,69444	0,52926
Moda	#N/A	#N/A	#N/A
Desviación estándar	0,35068	0,32087	0,25767
Varianza de la muestra	0,12298	0,10296	0,06639
Curtosis	-1,61643	-0,73600	-1,00353
Coeficiente de asimetría	-0,02602	-0,71760	-0,37420
Rango	0,97647	0,95174	0,77754
Mínimo	0,01485	0,00781	0,05723
Máximo	0,99133	0,95956	0,83477
Suma	7,59711	9,28131	7,05286
Cuenta	15	15	15
Mayor (1)	0,99133	0,95956	0,83477
Menor(1)	0,01485	0,00781	0,05723
Nivel de confianza(95,0%)	0,19420	0,17769	0,14269

13.6 Suavización exponencial

Aplica un modelo de **suavizado exponencial** a un conjunto de datos. Es necesario proporcionar el factor de suavización.



- Factor de suavización. Introduzca el factor de suavización que desee utilizar como constante de suavización exponencial. El factor de suavización es un factor correctivo que minimiza la inestabilidad de los datos reunidos entre una población. El factor predeterminado es 0,3. Los valores de 0,2 a 0,3 son constantes de suavización adecuadas. Estos valores indican que el pronóstico actual debe ajustarse entre un 20% y un 30% del error en el pronóstico anterior. Las constantes mayores generan una respuesta más rápida, pero pueden producir proyecciones erróneas. Las constantes más pequeñas pueden dar como resultado retrasos prolongados en los valores pronosticados.
- Crear gráfico. Active esta casilla para generar en la tabla de resultados un gráfico incrustado de los valores reales y los valores pronosticados.
- Error típico. Active esta casilla si desea incluir una columna que contenga valores de error típico en la tabla de resultados. Desactívela si desea una tabla de resultados en una sola columna y sin valores de error típicos.



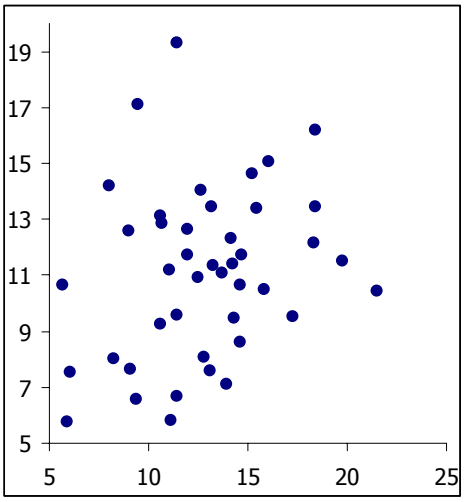
13.7 Prueba t para varianzas de dos muestras

RESULTADO (2 Normales X e Y ambas con media = 14 sigma = 4

X	Y
14,6	10,7
11,4	9,6
11,4	3,07
11,5	6,68
12,6	14
17,3	9,52
9,01	12,6
11,1	11,2
18,3	12,2
11,4	19,3
9,48	17,1
11,9	11,7
18,4	13,5
13,7	11,1
14,7	11,7
9,06	7,61
12,8	8,07
19,8	11,5
5,65	10,7
15,8	10,5
9,38	6,53
13,3	11,3
14,3	9,45
10,6	13,1
10,7	12,8
2,1	11,7
13	20,5
5,9	5,77

Prueba F para varianzas de dos muestras

	X	Y
Media	11,99	11,85
Varianza	5,79	3,72
Observaciones	50	50
Grados de libertad	49	49
F	1,5555	
P(F<=f) una cola	0,0627	
Valor crítico para F (una cola)	1,6073	

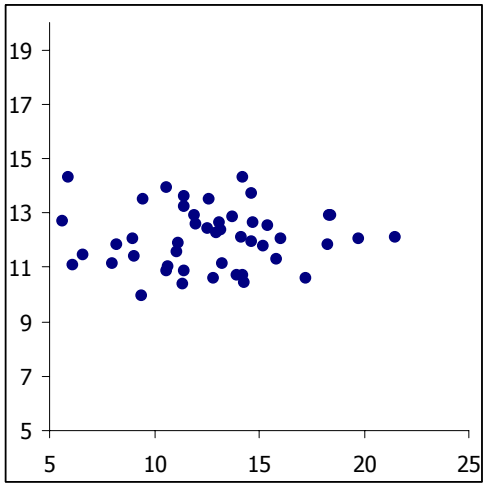


RESULTADO (2 Normales de media=14; X con sigma=4; Y con sigma =1)

X	Y
14,6	11,9
11,4	13,2
11,4	10,4
11,5	13,6
12,6	13,5
17,3	10,6
9,01	12,1
11,1	11,6
18,3	11,8
11,4	10,9
9,48	13,5
11,9	12,9
18,4	12,9
13,7	12,8
14,7	12,6
9,06	11,4
12,8	10,6
19,8	12,1
5,65	12,7
15,8	11,3
9,38	9,93
13,3	11,1
14,3	10,4
10,6	13,9
10,7	11
2,1	13,5
13	12,3
5,9	14,3

Prueba F para varianzas de dos muestras

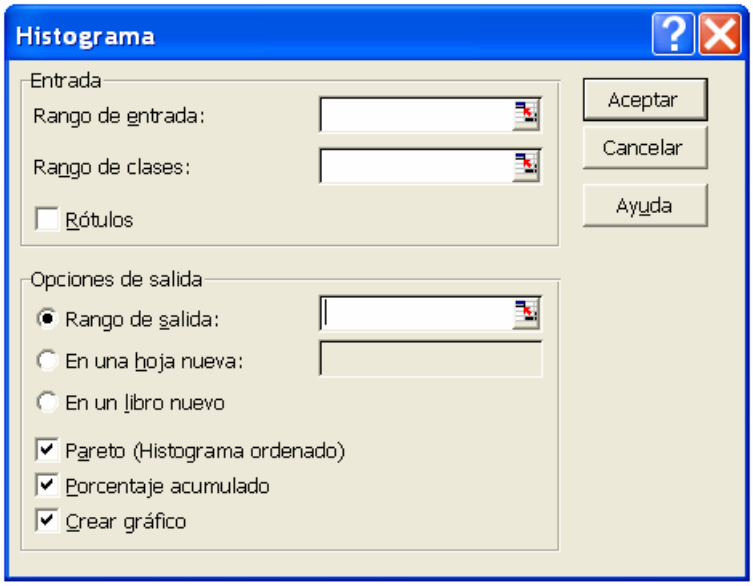
	X	Y
Media	12,12	12,07
Varianza	17,51	1,19
Observaciones	50	50
Grados de libertad	49	49
F	14,6739	
P(F<=f) una cola	0,0000	
Valor crítico para F (una cola)	1,6073	



13.8 Análisis de Fourier

13.9 Histograma

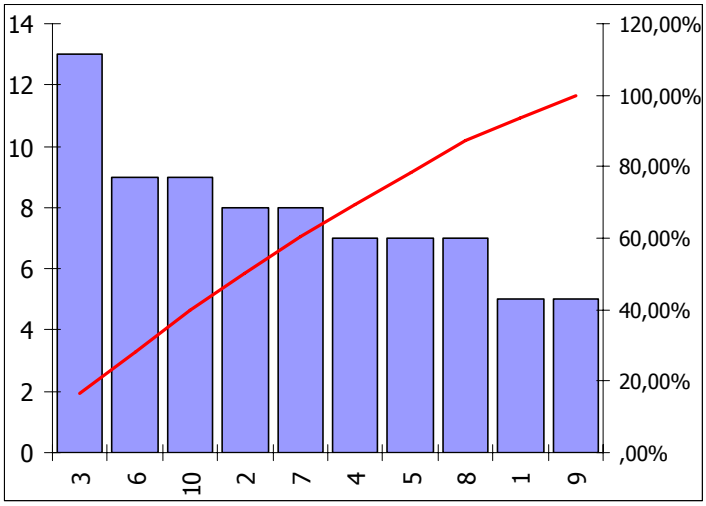
Obtiene la **distribución de frecuencias** de un conjunto de datos. Dibuja un **histograma** y el **diagrama de Pareto**.



- Rango clases (**opcional**) Introduzca la referencia de celda para un rango que contenga un conjunto opcional de valores límite que definan rangos de clase. Estos valores deberán estar en orden ascendente. Microsoft Excel contará el número de puntos de datos que hay entre el número de clases actual y el número de clases mayor, si lo hay.
- Se contará un número de una clase determinada si es igual o menor que el número de clase situado por debajo de la última clase. Todos los valores por debajo del primer valor de clase se contarán juntos, como los valores por encima del último valor de clase.
- Si omite el rango de clase, Excel creará un conjunto de clases distribuidas uniformemente entre los valores mínimo y máximo de los datos.
- Pareto (Histograma ordenado) Active esta casilla para presentar los datos en orden de frecuencia descendente en la tabla de resultados. Si esta casilla está desactivada, Excel presentará los datos en orden ascendente y omitirá las tres columnas situadas más a la derecha que contienen los datos ordenados.
- Porcentaje acumulado Active esta casilla para generar una columna de tabla de resultados con porcentajes acumulados y para incluir una línea de porcentaje acumulado en el gráfico de histograma. Desactívela para omitir los porcentajes acumulados.
- Crear gráfico Active esta casilla para generar un gráfico de histograma incrustado con la tabla de resultados.

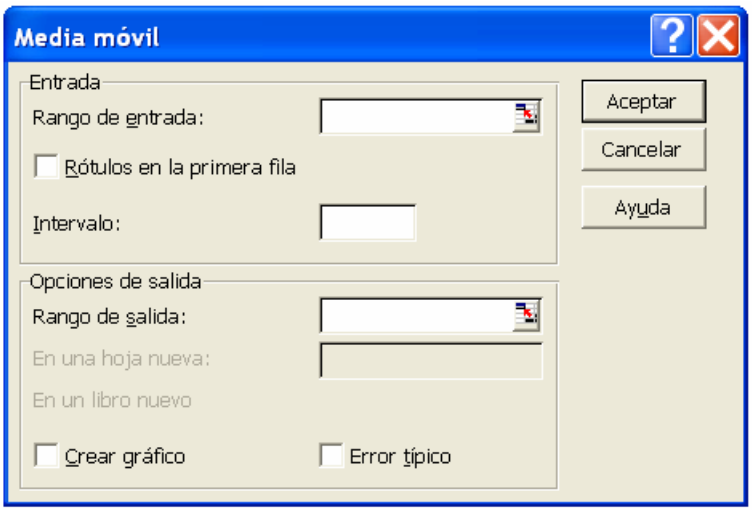
RESULTADO

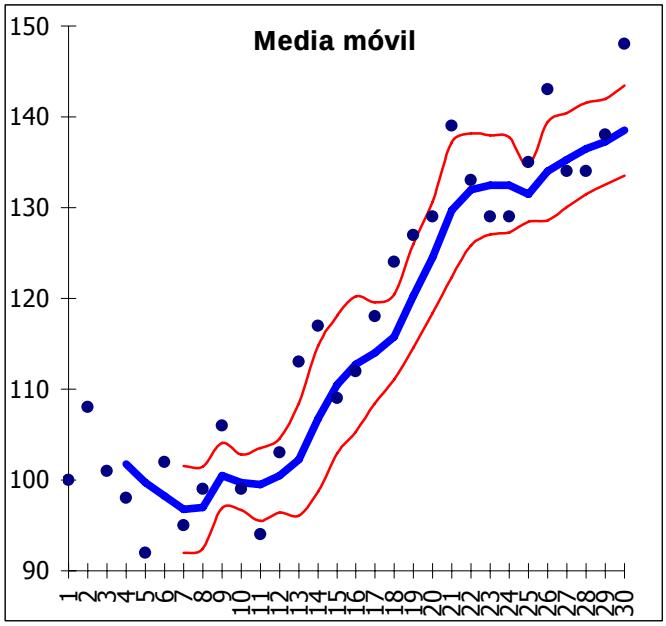
				Clase	Frecuencia	% acumulado	Clase	Frecuencia	% acumulado
8	8	5	**1**	1	5	6,41%	3	13	16,67%
3	2	2	**2**	2	8	16,67%	6	9	28,21%
6	4	5	**3**	3	13	33,33%	10	9	39,74%
9	3	10	**4**	4	7	42,31%	2	8	50,00%
6	2	10	**5**	5	7	51,28%	7	8	60,26%
10	6	10	**6**	6	9	62,82%	4	7	69,23%
2	6	9	**7**	7	8	73,08%	5	7	78,21%
7	5	5	**8**	8	7	82,05%	8	7	87,18%
6	1	5	**9**	9	5	88,46%	1	5	93,59%
5	2	2	**10**	10	9	100,00%	9	5	100,00%
9	7	6		y mayor...	0	100,00%			
9	7	7							
4	3	9							
3	7	3							
1	2	4							
2	5	2							
7	7	8							
8	5	1							
8	2	1							
2	2	1							
5	3	4							
5	6	3							
10	5	8							
2	4	6							
4	5	4							
10	10	10							



13.10 Media móvil

Obtiene la **media móvil** para un intervalo dado.





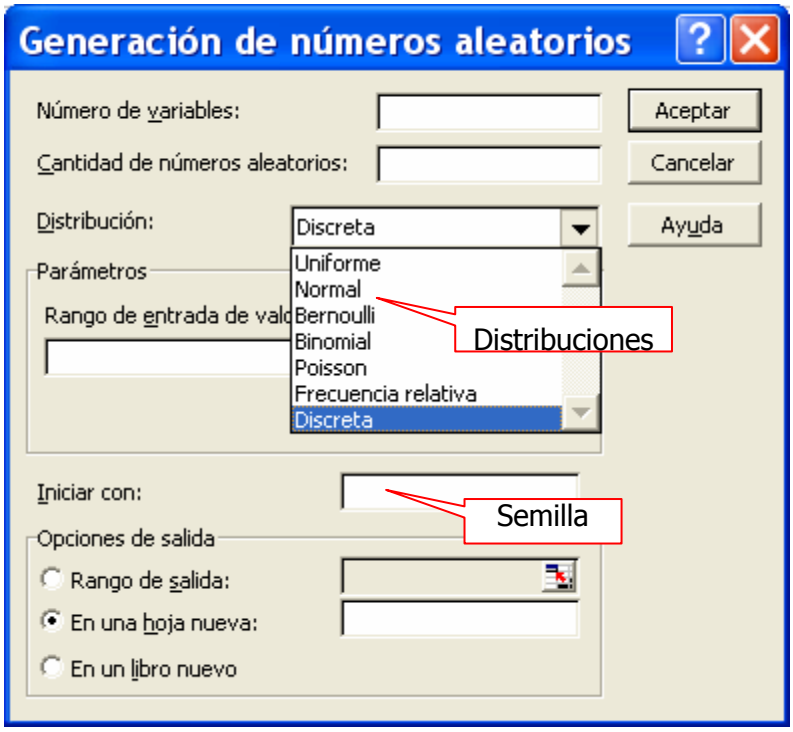
13.11 Generación de números aleatorios

- Número de variables Introduzca el número de columnas de valores que desee incluir en la tabla de resultados. Si no introduce ningún número, Microsoft Excel rellenará todas las columnas del rango de salida que se haya especificado.

- Cantidad de números aleatorios Introduzca el número de puntos de datos que desee ver. Cada punto de datos aparecerá en una fila de la tabla de resultados. Si no introduce ningún número, Microsoft Excel rellenará todas las columnas del rango de salida que se haya especificado.

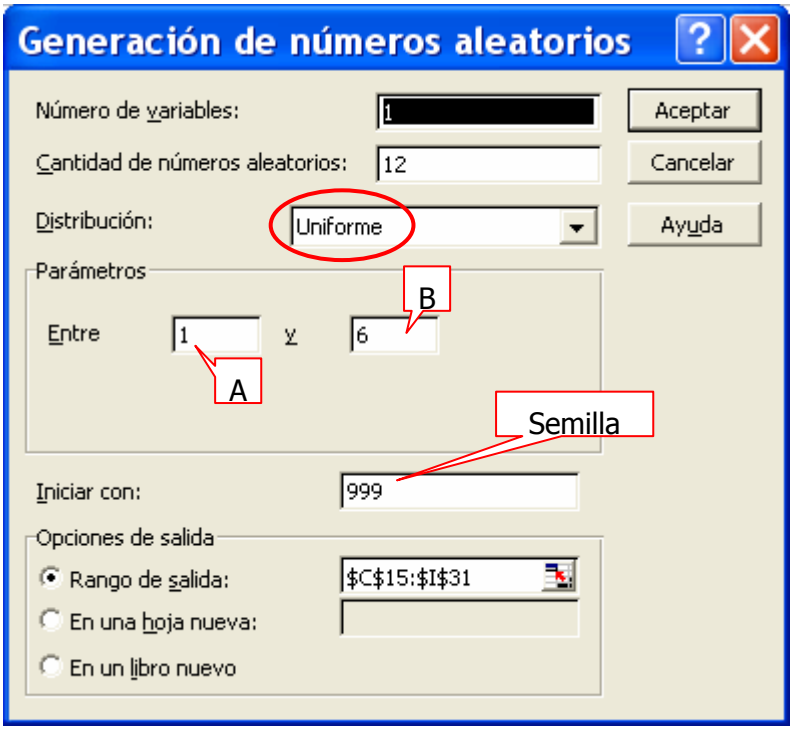
- Distribución Haga clic en el método de distribución que desee utilizar para crear las valores aleatorios.
    - **Uniforme** Caracterizado por los límites inferior y superior. Se extraen las variables con probabilidades iguales de todos los valores del rango.
    - **Normal** Caracterizado por una media y una desviación estándar.
    - **Bernoulli** Caracterizado por la probabilidad de éxito (valor p) en un ensayo dado. La variables aleatorias de Bernoulli tienen el valor 0 o 1.
    - **Binomial** Caracterizado por una probabilidad de éxito (valor p) durante un número de pruebas.
    - **Poisson** Caracterizado por un valor lambda, igual a 1/media.
    - **Frecuencia relativa** Caracterizado por un límite inferior y superior, un incremento, un porcentaje de repetición para valores y un ritmo de repetición de la secuencia.
    - **Discreta** Caracterizado por un valor y el rango de probabilidades asociado. El rango debe contener dos columnas. La columna izquierda deberá contener valores y la derecha probabilidades asociadas con el valor de esa fila. La suma de las probabilidades deberá ser 1.
    - Parámetros Introduzca un valor o valores para caracterizar la distribución seleccionada.
    - Iniciar con Escriba un valor opcional a partir del cual se generarán números aleatorios. Podrá volver a utilizar este valor para generar los mismos números aleatorios más adelante.

Plantilla general



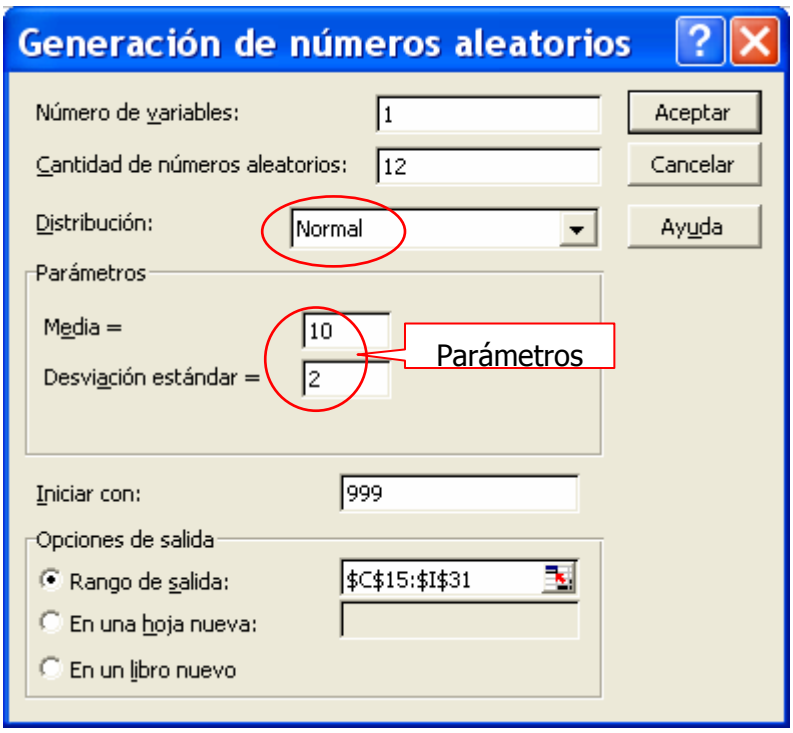
UNIFORME

Genera muestras de una distribución $U_{[A;B]}$



NORMAL

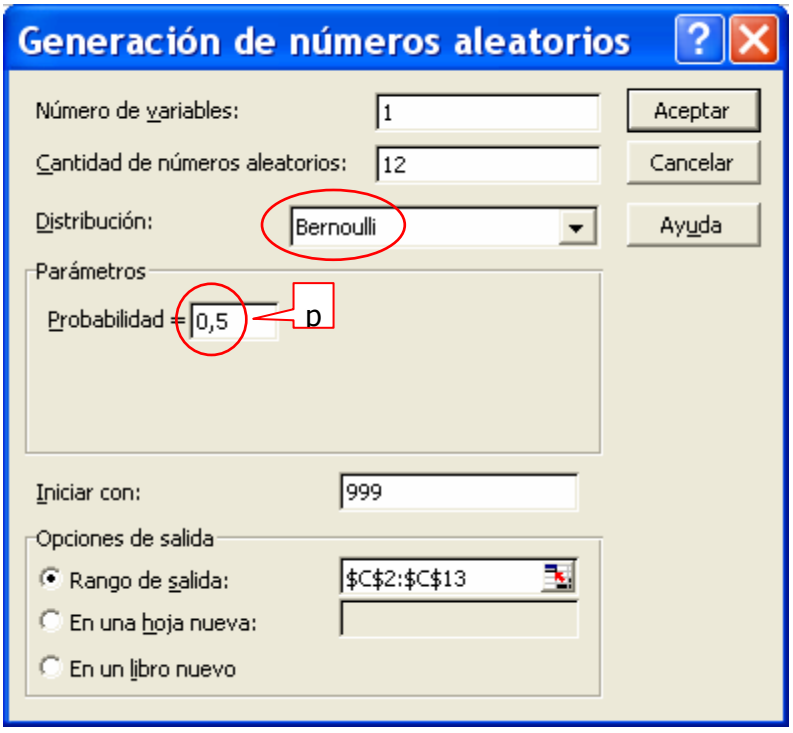
Genera muestras de una distribución $\mathbf{N}_{[\mu;\sigma]}$



7,445
12,850
8,800
11,015
11,234
7,113
7,882
11,295
11,100
9,988
10,877
8,320

BERNOULLI

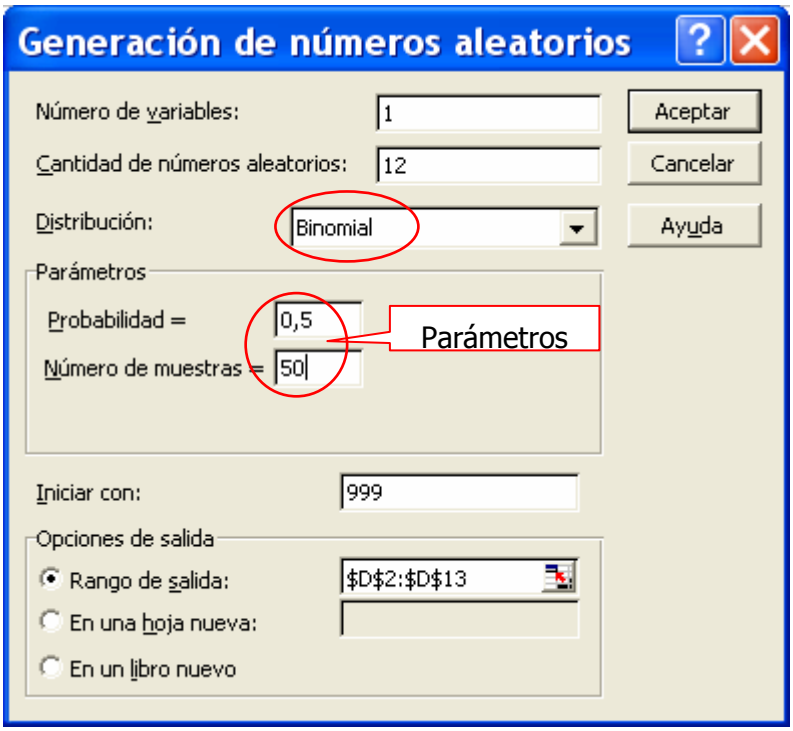
Genera muestras de una distribución de Bernoulli(p)



1
0
1
0
0
1
1
0
0
1
0
1

BINOMIAL

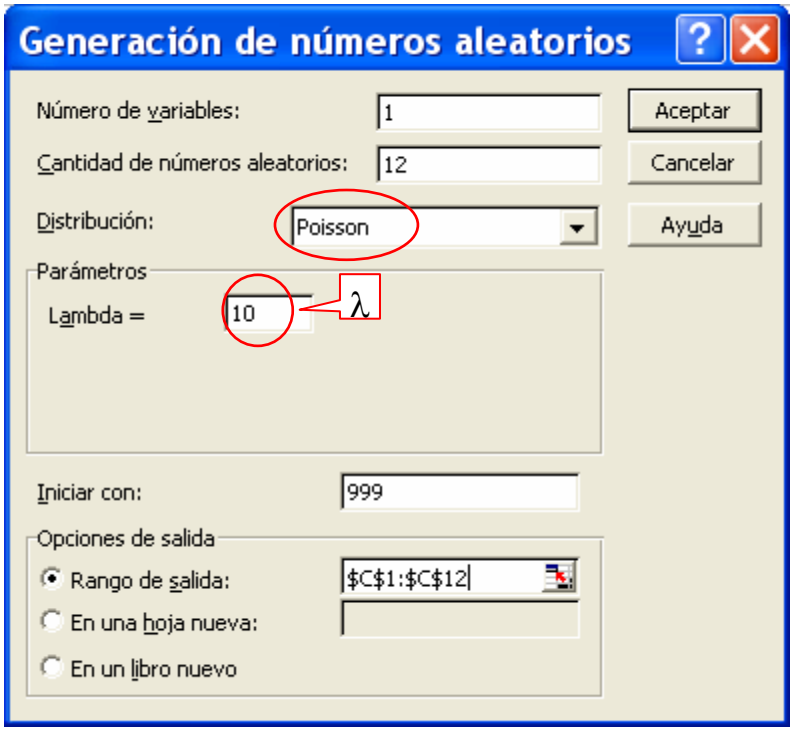
Genera muestras de una distribución de $\mathbf{B}_{(n,p)}$



24
30
24
21
22
27
23
27
23
26
27
19

POISSON

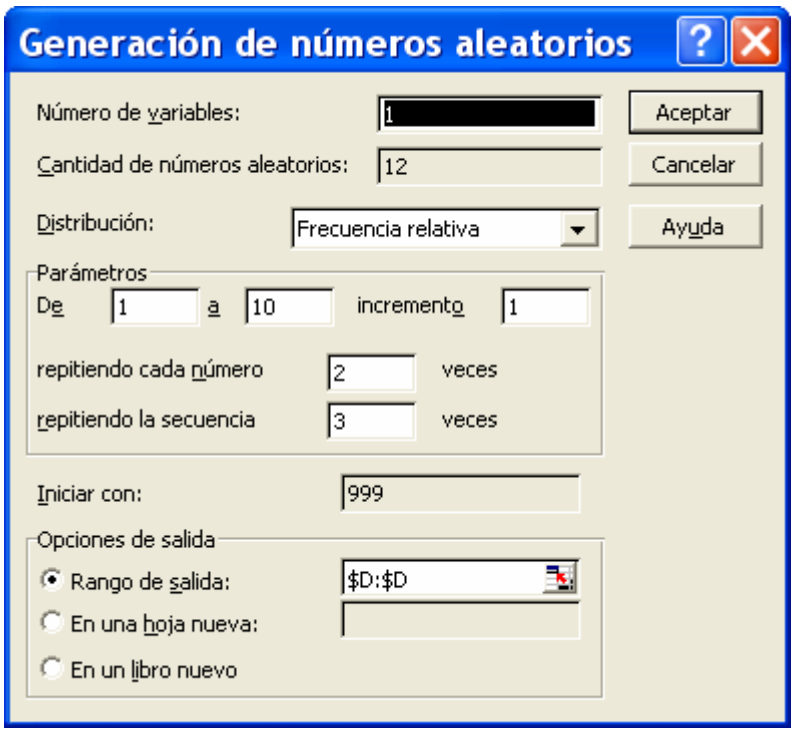
Genera muestras de una distribución de **Poisson**(λ)



12
14
10
8
11
12
9
13
13
6
10
12

FRECUENCIA RELATIVA

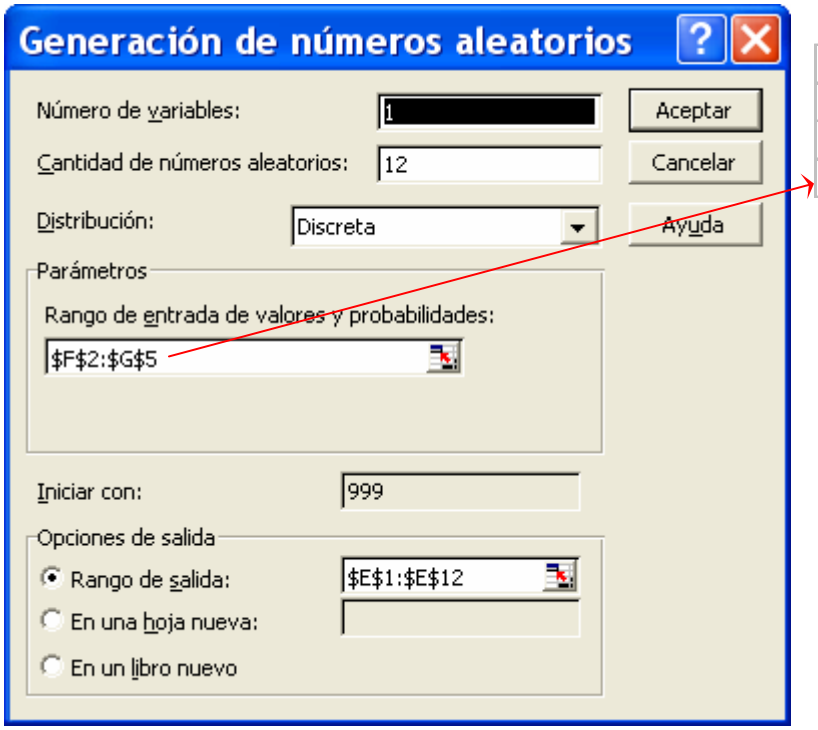
En realidad genera series o secuencias de números



1
1
2
2
3
3
4
4
5
5
6
6
7
7
8
8
9
9
10
10
1
1

DISCRETA

Genera números dada una distribución de frecuencias relativas

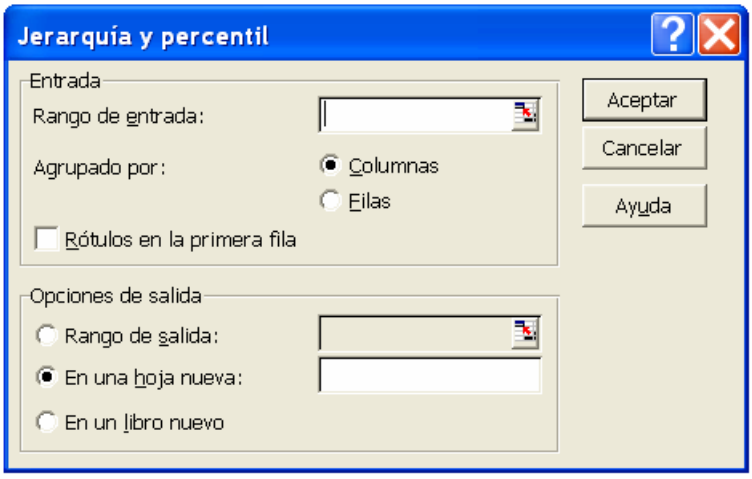


0	0,1875
1	0,1250
2	0,3125
3	0,3750

0
3
1
3
3
0
0
3
3
2
3
1

13.12 Jerarquía y percentil

Realiza el equivalente a las funciones de los mismos nombres.



13.13 Regresión

Ver apartado correspondiente

13.14 Muestreo

- Método de muestreo Haga clic en Periódico o Aleatorio para indicar el intervalo de muestreo que desee.
- Período Introduzca el intervalo periódico en el que desee realizar la muestra. El valor n del período del rango de entrada y cada valor n del período siguiente se copiarán en la columna de resultados. El muestreo terminará cuando se llegue al final del rango de entrada.
- Número de muestras Introduzca el número de valores aleatorios que desee en la columna de resultados. Cada valor se extrae de una posición aleatoria del rango de entrada y puede seleccionarse cualquier número más de una vez.

13.15 Prueba t

13.15.1 Prueba t para dos muestras suponiendo varianzas iguales

13.15.2 Prueba t para dos muestras suponiendo varianzas desiguales

13.15.3 Prueba t para medias de dos muestras emparejadas

13.16 Prueba z

13.17 PROBLEMAS

13.17.1 Simular los resultados del lanzamiento de un dado equilibrado (1000 lanzamientos). Comparar los resultados obtenidos con los esperados.

13.17.2 Simular los resultados de medir un colectivo de 500 personas de las que se sabe que su altura se distribuye según una Normal de media 175 cm. y desviación 8 cm. ¿Qué porcentaje del colectivo tiene una altura superior a 185cm? compara los resultados con los teóricos.

13.17.3 Simular el resultado de un test compuesto por 25 preguntas, cada una de ellas con 4 respuestas de las que sólo una es correcta, contestado por alguien que selecciona la respuesta al azar. Igual pero con dos respuestas posibles de las que sólo una es la correcta.

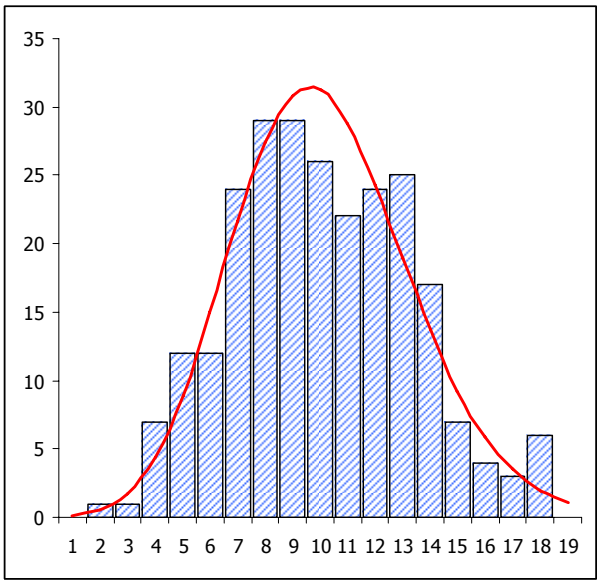
13.17.4 Simular el resultado de una clase de 100 alumnos que se somete a los exámenes descritos en el problema anterior. ¿Que porcentaje aprueba en cada caso? Comparar con los resultados teóricos.

13.17.5 Simular 250 observaciones de una distribución de Poisson de media 10.

a) Obtener la distribución de frecuencias de los datos simulados.
b) Trazar el histograma de los datos y sobreimponer la distribución que cabría esperar se hubiera dado.
c) Utilizar **SOLVER** para estimar el parámetro por mínimos cuadrados

Datos				
7	11	8	13	11
9	8	9	11	8
13	11	7	6	5
9	11	15	17	9
7	14	15	9	13
12	12	16	8	5
9	13	9	9	5
9	12	8	8	7
18	8	7	18	12
11	13	5	10	14
13	10	9	8	14
11	4	11	4	8
8	13	9	10	4
15	6	13	11	11
8	12	6	13	6
12	7	9	10	13
11	13	12	8	13
18	8	8	13	14
10	7	9	7	6
5	9	8	6	8
10	5	16	8	10
12	12	12	7	16
2	12	13	8	12
10	7	4	12	7
9	14	9	10	6

Clase	Obs	Esp	Dife
1	0	0,101	0,01
2	1	0,511	0,24
3	1	1,726	0,53
4	7	4,370	6,91
5	12	8,855	9,89
6	12	14,952	8,71
7	24	21,640	5,57
8	29	27,404	2,55
9	29	30,847	3,41
10	26	31,251	27,57
11	22	28,782	45,99
12	24	24,299	0,09
13	25	18,936	36,77
14	17	13,703	10,87
15	7	9,255	5,08
16	4	5,860	3,46
17	3	3,492	0,24
18	6	1,966	16,28
19	0	1,048	1,10
		Residuos	**185,28**
		Media	**10,13**
		Casos	**250**



13.17.6 Un proceso industrial puede fallar por 6 tipos de motivos (A, B, ..) cuyas probabilidades se dan en la tabla siguiente:

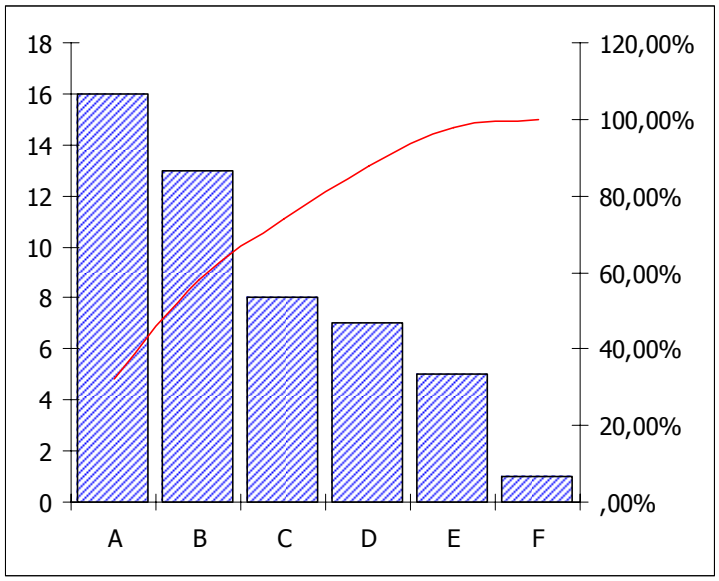
A	B	C	D	E	F
0,12	0,13	0,18	0,23	0,06	0,28

a) Simular 50 fallos del proceso.
b) Obtener la distribución de frecuencias de los fallos.
c) Dibujar un histograma y un diagrama de Pareto.

A	1	0,12
B	2	0,13
C	3	0,18
D	4	0,23
E	5	0,06
F	6	0,28

3	C
4	D
1	A
6	F
6	F
2	B
1	A
4	D
3	C
2	B
2	B
4	D
4	D
4	D
3	C
6	F
4	D
6	F
4	D
4	D
4	D
6	F
4	D
5	E
1	A
4	D
3	C

Clase	Frecuencia	% acumulado	Clase	Frecuencia	% acumulado
A	8	16,00%	D	16	32,00%
B	5	26,00%	F	13	58,00%
C	7	40,00%	A	8	74,00%
D	16	72,00%	C	7	88,00%
E	1	74,00%	B	5	98,00%
F	13	100,00%	E	1	100,00%



14 ACTIVIDADES PROPUESTAS

Prácticas de Excel para la resolución de cuestiones estadísticas

14.1 Actividad 1

Simular 20 puntuaciones al azar de un test donde el valor más bajo sea 15 y el más alto 50. Calcular los percentiles {0%, 25%, 50%, 75% y 100%} para estos datos, y la media y la desviación estándar, y comentar la diferencia entre la media y la mediana como medidas del centro de la distribución. Generalizar la actividad de forma que los valores más bajo (15) y más alto (50) puedan ser modificados por el usuario.

Recordemos que:

La función de Excel ALEATORIO() proporciona un número pseudo-aleatorio, de distribución uniforme en el intervalo 0;1. La generación de puntuaciones aleatorias, no entre cero y la unidad, sino entre dos valores A y B (A<B) se realiza mediante la expresión:

A + ((B-A)*ALEATORIO())

que nos proporcionará una realización de una variable continua (ya que podremos obtener cualquier valor comprendido entre A y B.

También podríamos usar la función de Excel ALEATORIO.ENTRE(A;B) que tendría el mismo efecto que la expresión anterior pero con la importante diferencia de que proporcionaría una variable discreta en vez de continua, es decir sólo obtendríamos puntuaciones comprendidas en el rango {A, A+1, A+2,...,B-1,B}.

Una vez generados los valores deberemos analizarlos para realizar la segunda parte de la actividad. Para obtener los percentiles usamos la función CUARTIL, función que se invoca con dos argumentos, exactamente en la forma CUARTIL(*matriz* ; *cuartil*) siendo *matriz* el rango de celdas de valores numéricos cuyo cuartil desea obtener y *cuartil* un entero en el rango {0,1,2,3,4,5} que le indica a Excel que cuartil deseamos, y que respectivamente serían {0%, 25%, 50%, 75% y 100%}, es decir {mínimo, primer cuartil, mediana, tercer cuartil y máximo}.

Sabido todo esto sólo queda plasmarlo en la hoja de cálculo, añadiendo puesto que hemos decidido generalizar los extremos entre los cuales queremos que se generen las puntuaciones de los tests, controles para poder modificar dichos valores.

Para facilitar la realización de la última parte de la actividad, "observar las diferencias entre la media y la mediana" podemos añadir un gráfico de los valores obtenidos junto con los estadísticos calculados lo que proporcionará más información que la mera observación de los valores numéricos por la del gráfico correspondiente.

La hoja para realizar esta actividad podría tener una aspecto como el del gráfico en la página siguiente:

Los elementos de la hoja son, el enunciado de la actividad, una casilla que contiene un comentario sobre la función CUARTIL, los datos generados, los valores máximo y mínimo junto con los controles para establecer su valor, los estadísticos calculados (los 5 cuartiles más la media y la desviación típica) y finalmente el gráfico de todo lo anterior.

Nótese que al tratarse de valores volátiles, cada vez que pulsemos **F9** obtendremos una muestra diferente y podremos observar la variabilidad de los resultados.

Actividad 1

Simulad 20 puntuaciones al azar de un test donde el valor más bajo sea 15 y el más alto 50. Calculad los cinco percentiles resumen para estos datos, y la media y la desviación estándar, y comentad la diferencia entre la media y la mediana como medidas del centro de la distribución.

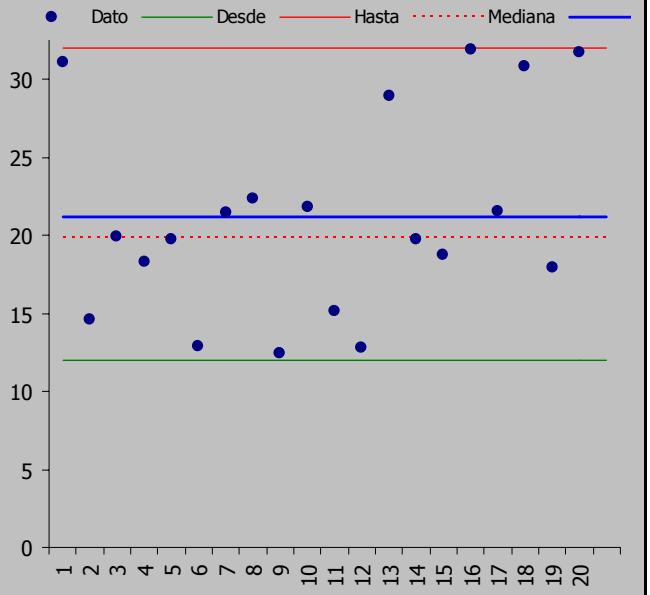
Comentario

Desde 12

Hasta 32

Nº	Dato
1	31,15
2	14,64
3	19,95
4	18,34
5	19,73
6	12,87
7	21,44
8	22,35
9	12,44
10	21,86
11	15,13
12	12,81
13	28,95
14	19,72
15	18,73
16	31,94
17	21,58
18	30,86
19	17,92
20	31,74

Mínimo	12
Q1	17
Mediana	20
Q3	24
Máximo	32
Media	21
Desv Típica	6,5



14.2 Actividad 2

En esta actividad se supone que la distribución de la duración de las llamadas telefónicas hechas a un centro de apoyo psicológico es normal[1], con una media de 157 segundos y una desviación estándar de 52 segundos. Se pide utilizar las tablas para calcular la probabilidad de que una llamada tenga una duración de entre 3 y 4 minutos y la de que una llamada tenga una duración superior a los 4 minutos.

Resolveremos primero la actividad usando la teoría aprendida y las tablas de la normal. El enunciado dice que:

> *la distribución de la duración de las llamadas telefónicas hechas a un centro de apoyo psicológico es normal, con una media de 157 segundos y una desviación estándar de 52 segundos*

si llamamos **D** a la duración de las llamadas, lo que tendremos, por ahora es que:

$$D \approx \mathbf{N}_{(\pi;\sigma)} \text{ con } \pi=157 \text{ y } \sigma=52$$

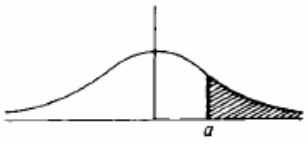
y lo que nos piden es calcular la probabilidad $P(3*60 \le D \le 4*60)$. La teoría nos dice que, para contestar a esta pregunta, debemos primero normalizar y después buscar en la tabla de la distribución Normal, esto, es:

$$\mathbf{P}(180 \le \mathbf{D} \le 240) = \mathbf{P}\left(\frac{180-\pi}{\sigma}\right) \le \mathbf{z} \le \left(\frac{240-\pi}{\sigma}\right) = \mathbf{P}(0{,}4423 \le \mathbf{z} \le 1{,}5961)$$

y usando la tabla llegamos a que:

$$P(3*60 \le D \le 4*60) \approx P(0{,}44 \le Z \le 1{,}6) = 0{,}33\text{-}0{,}0548 = 0{,}2752$$

Tabla 3. Distribución normal (0; 1). $P\ (X \ge a)$



	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,4960	0,4920	0,4880	0,4841	0,4801	0,4761	0,4721	0,4681	0,4641
0,1	0,4602	0,4562	0,4522	0,4483	0,4443	0,4404	0,4364	0,4325	0,4286	0,4247
0,2	0,4207	0,4168	0,4129	0,4091	0,4052	0,4013	0,3974	0,3936	0,3897	0,3859
0,3	0,3821	0,3783	0,3745	0,3707	0,3669	0,3632	0,3594	0,3557	0,3520	0,3483
0,4	0,3446	0,3409	0,3372	0,3336	0,3300	0,3264	0,3228	0,3192	0,3156	0,3121
0,5	0,3085	0,3050	0,3015	0,2981	0,2946	0,2912	0,2877	0,2843	0,2810	0,2776
0,6	0,2743	0,2709	0,2676	0,2644	0,2611	0,2579	0,2546	0,2514	0,2483	0,2451
0,7	0,2420	0,2389	0,2358	0,2327	0,2297	0,2266	0,2236	0,2207	0,2177	0,2148
0,8	0,2119	0,2090	0,2061	0,2033	0,2005	0,1977	0,1949	0,1922	0,1894	0,1867
0,9	0,1841	0,1814	0,1788	0,1762	0,1736	0,1711	0,1685	0,1660	0,1635	0,1611
1,0	0,1587	0,1563	0,1539	0,1515	0,1492	0,1469	0,1446	0,1423	0,1401	0,1379
1,1	0,1357	0,1335	0,1314	0,1292	0,1271	0,1251	0,1230	0,1210	0,1190	0,1170
1,2	0,1151	0,1131	0,1112	0,1094	0,1075	0,1057	0,1038	0,1020	0,1003	0,0985
1,3	0,0968	0,0951	0,0934	0,0918	0,0901	0,0885	0,0869	0,0853	0,0838	0,0823
1,4	0,0808	0,0793	0,0778	0,0764	0,0749	0,0735	0,0721	0,0708	0,0694	0,0681
1,5	0,0668	0,0655	0,0643	0,0630	0,0618	0,0606	0,0594	0,0582	0,0571	0,0559
1,6	0,0548	0,0537	0,0526	0,0516	0,0505	0,0495	0,0485	0,0475	0,0465	0,0455
1,7	0,0446	0,0436	0,0427	0,0418	0,0409	0,0401	0,0392	0,0384	0,0375	0,0367
1,8	0,0359	0,0352	0,0344	0,0336	0,0329	0,0322	0,0314	0,0307	0,0301	0,0294
1,9	0,0287	0,0281	0,0274	0,0268	0,0262	0,0256	0,0250	0,0244	0,0239	0,0233

[1] Formalmente diríamos que la duración de las llamadas se distribuye normalmente o que su función de densidad es normal o gaussiana.

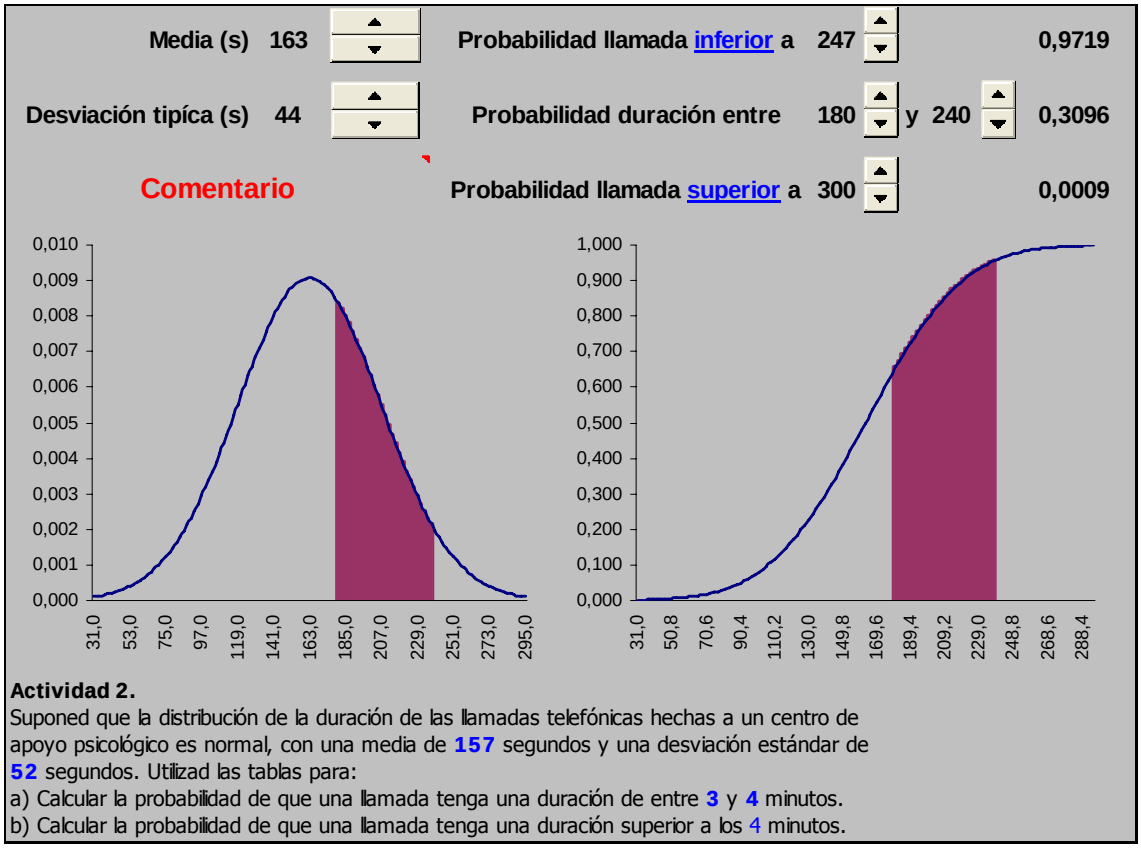
Abordaremos la actividad a través de Excel, para los realizar los cálculos necesarios usaremos la función DISTR.NORM que devuelve la función de distribución de una variable normal de media y desviación típica especificadas.

La forma de invocar esta función es proporcionándola al menos tres argumentos, DISTR.NORM(x; media; desv_estándar; acum), siendo opcional el último. Los argumentos son x, el valor de la variable aleatoria cuya distribución desea obtener; media, que debe ser la media aritmética de la distribución; desv_estándar, la desviación estándar de la distribución y acum que es un valor lógico que determina la forma de la función: si acum es VERDADERO, entonces DISTR.NORM devuelve la función de distribución acumulada; si es FALSO, devuelve la función de densidad.

El cálculo en Excel será tan sencillo como:

DISTR.NORM(A;π;σ;VERDADERO)-DISTR.NORM(B;π;σ;VERDADERO)

sustituyendo los parámetros genéricos por los deseados (A = 240; B = 180; π = 157 ;σ = 52), o por cualesquiera otros si, como es el caso, deseamos generalizar la respuesta. La hoja de la actividad muestra los resultados obtenidos al varia los posibles parámetros del problema.



Actividad 2.
Suponed que la distribución de la duración de las llamadas telefónicas hechas a un centro de apoyo psicológico es normal, con una media de **157** segundos y una desviación estándar de **52** segundos. Utilizad las tablas para:
a) Calcular la probabilidad de que una llamada tenga una duración de entre **3** y **4** minutos.
b) Calcular la probabilidad de que una llamada tenga una duración superior a los 4 minutos.

La hoja tiene también una representación gráfica no sólo de las funciones de densidad y distribución de la variable normal elegida, sino también de las áreas asociadas a las probabilidades pedidas. El detalle de como pueden ser construidas estas gráficas se ha relegado al apéndice 1.

14.3 Actividad 3

Como en la actividad 2, ahora trabajamos con una distribución normal con una media de 157 segundos y una desviación estándar de 52 segundos. La pregunta ahora es ¿cuál es la distribución de la media de 1.000 llamadas telefónicas seleccionadas aleatoriamente?.

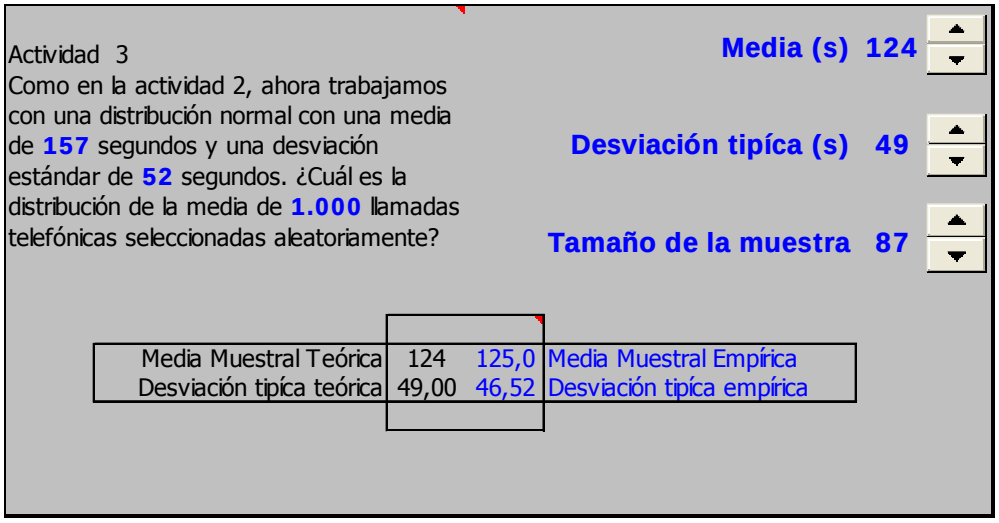
La teoría de la distribución en el muestreo de los parámetros de una distribución normal[2] es sencilla: extraída una muestra de tamaño **n** de una distribución normal $N_{(\pi;\sigma)}$ la media muestral se distribuye:

$$\overline{\mathbf{x}} \cong \mathbf{N}_{(\pi,\sigma/\sqrt{\mathbf{n}})}$$

El cálculo teórico es entonces directo, la media muestral se distribuye con la misma media de la población 157 segundos, y dado que su desviación típica es aproximadamente 1,65, cabe esperar que la mayoría de las ocasiones no sea inferior a 153 ni superior a 161.

Hasta aquí la respuesta, pero podemos desear comprobar por nosotros mismos que la teoría acerca de la distribución en el muestreo es cierta simulando un numero de muestras cada una de tamaño n = 1000, calculando su media y viendo si realmente se adapta a lo predicho por la teoría.

Esto es lo que hace precisamente la hoja de cálculo dedicada a esta actividad, que como vemos está dividida en dos partes, una primera en la que se muestran los resultados de la muestra.



y una segunda (que se mantiene oculta en las columnas KLM) en la que se genera la muestra.

2 En realidad, siguiendo el Teorema Central del Límite, de cualquier distribución si la muestra es suficientemente grande o proviene de una distribución normal.

ind	F	Normal
1	1	119,4
2	1	155,9
3	1	241,6
4	1	179,8
5	1	77,7
6	1	59,2
7	1	209,2
8	1	168,8
9	1	143,0
10	1	181,5
11	1	168,4
12	1	198,7
13	1	90,2
14	1	124,8
15	1	90,0
16	1	96,7
17	1	184,0
18	1	151,3
19	1	95,4
20	1	77,9

Para generar las duraciones de las llamadas que como nos dicen se distribuyen según una distribución normal, usaremos dos funciones de Excel: ALEATORIO() que ya nos es conocida y otra que vemos en este documento por vez primera DISTR.NORM.INV.

Esta última tiene la siguiente sintaxis:

$$\text{DISTR.NORM.INV}(p;\ \pi\ ;\ \sigma)$$

y devuelve el valor crítico de la distribución acumulativa normal de media π y desviación estándar σ. Esto es, dados **p**, π y σ, la función calcula el valor X tal que se verifica que:

$$P(X \approx \mathbf{N}_{(\pi;\sigma)}) = \mathbf{p}$$

Esto nos permite, sin más que sustituir p por un valor aleatorio uniforme obtener realizaciones aleatoria de una distribución normal $\mathbf{N}_{(\pi;\sigma)}$ sin más que usar la fórmula:

$$=\text{DISTR.NORM.INV}(\text{ALEATORIO}();\ \pi\ ;\ \sigma)$$

ésta es, precisamente, la fórmula que figura en la columna cuyo epígrafe es "Normal"[3].

Tras esto sólo queda calcular la media de la muestra y compararla con el valor teórico, puesto que se trata de valores volátiles obtendremos un resultado diferente (extraeremos una muestra diferente) cada vez que recalculemos la hoja (F9).

Media Muestral Teórica	124	111,8	Media Muestral Empírica
Desviación tipíca teórica	49,00	42,19	Desviación tipíca empírica

[3] Las dos columnas anteriores ind y F son un índice y una "bandera" usadas para poder generalizar sobre el tamaño de la muestra y no entraremos en su explicación toda vez que ésta excede el objetivo de la actividad propuesta. El estudiante interesado puede, no obstante, inspeccionar las fórmulas de la hoja y en caso necesario solicitar más información al consultor de la asignatura.

14.4 Actividad 4

En un casino de juego, una máquina de apuestas determinada da al jugador una probabilidad de victoria de 0,4. El resultado de una jugada no tiene ninguna conexión con el resultado de la siguiente. Un jugador juega 200 veces en esta máquina. ¿Cuál es la probabilidad de que el jugador gane 100 veces o más?

La teoría que nos permite responder a la pregunta es sencilla: puesto que se trata de la repetición (en idénticas condiciones hemos de suponer) un número de veces n (200 según el enunciado) de ensayos de Bernouilli independientes cuya probabilidad de éxito es p (0,4 si asociamos el éxito al resultado "ganar"), la variable aleatoria que describe el número de victorias en esas circunstancias es una binomial (n=200; p= 0,4).

La pregunta puede formularse entonces de la forma siguiente ¿qué valor tiene la siguiente probabilidad?:

$$P(X \geq 100) \text{ con } X \approx \mathbf{B}_{(n;p)}$$

Pero existe un inconveniente, para un valor de **n** tan grande no encontraremos tablas de la distribución binomial, y el cálculo de los valores empíricos puede llegar a ser verdaderamente engorroso y obligarnos además a trabajar con números muy pequeños lo que siempre representa un problema.

Afortunadamente la teoría también nos dice que, en estas circunstancias, la distribución binomial queda muy bien representada por una distribución normal de igual media y desviación típica. Esto es, podemos aprovechar el hecho de que:

$$\mathbf{x} \cong \mathbf{B}_{(\mathbf{n};\mathbf{p})} \rightarrow \mathbf{N}_{(\mathbf{n}\cdot\mathbf{p};\sqrt{\mathbf{n}\cdot\mathbf{p}\cdot(1-\mathbf{p})})}$$

Así, con la ayuda de tablas resolveríamos el problema de la forma siguiente: normalizaríamos

$$\mathbf{P}(\mathbf{G} \geq 100) = \mathbf{P}\left(\mathbf{z} \geq \frac{100-\mathbf{np}}{\sqrt{\mathbf{np}(1-\mathbf{p})}}\right) = \mathbf{P}\left(\mathbf{z} \geq \frac{100-200\cdot 0{,}4}{\sqrt{200\cdot 0{,}4\cdot 0{,}6}}\right) = \mathbf{P}(\mathbf{z} \geq 2{,}89)$$

al consultar en la tabla vemos que la probabilidad pedida es:

$$P(X \geq 100) = 0{,}0019$$

a

	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
2,0	0,0228	0,0222	0,0217	0,0212	0,0207	0,0202	0,0197	0,0192	0,0188	0,0183
2,1	0,0179	0,0174	0,0170	0,0166	0,0162	0,0158	0,0154	0,0150	0,0146	0,0143
2,2	0,0139	0,0136	0,0132	0,0129	0,0125	0,0122	0,0119	0,0116	0,0113	0,0110
2,3	0,0107	0,0104	0,0102	0,0099	0,0096	0,0094	0,0091	0,0089	0,0087	0,0084
2,4	0,0082	0,0080	0,0078	0,0076	0,0073	0,0071	0,0070	0,0068	0,0066	0,0064
2,5	0,0062	0,0060	0,0059	0,0057	0,0055	0,0054	0,0052	0,0051	0,0049	0,0048
2,6	0,0047	0,0045	0,0044	0,0043	0,0042	0,0040	0,0039	0,0038	0,0037	0,0036
2,7	0,0035	0,0034	0,0033	0,0032	0,0031	0,0030	0,0029	0,0028	0,0027	0,0026
2,8	0,0026	0,0025	0,0024	0,0023	0,0023	0,0022	0,0021	0,0021	0,0020	0,0019

Pero naturalmente podemos hacer esto mismo gracias al ordenador, y hacerlo además de diferentes formas para ver cuán de próximas está las diferentes aproximaciones a las que la teoría nos tiene acostumbrados.

Excel dispone de una función capaz de calcular probabilidades asociadas a la distribución binomial, se trata de:

DISTR.BINOM

que calcula tanto la función de masa de probabilidad como la función de distribución de una variable aleatoria que se distribuya siguiendo una binomial.

Su sintaxis es **DISTR.BINOM(x; n; p; acum)**, siendo x el número de éxitos en los ensayos; n el número de ensayos independientes; p la probabilidad de éxito en cada ensayo y acum un valor lógico que determina la forma de la función.

Si el argumento acum es VERDADERO, DISTR.BINOM devuelve la función de distribución acumulada, que es la probabilidad de que exista el máximo número de éxitos; si es FALSO, devuelve la función de masa de probabilidad.

Bastará entonces con calcular, en la celda correspondiente, la fórmula:

1 - DISTR.BINOM(100; 200; 0,4;VERDADERO)

para obtener el valor exacto[4] de la probabilidad pedida. Usando el complementario de la función de Excel ya que ésta nos proporcionaría, sin otra modificación, el valor $P_{(X<100)}$, en vez del valor $P_{(X\geq100)}$ pedido.

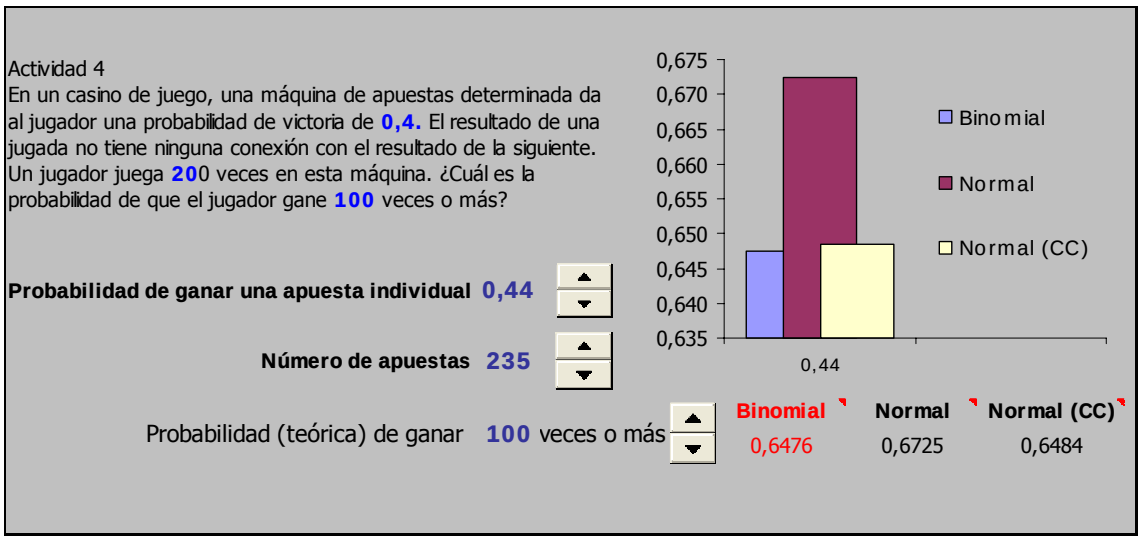
Podemos comprobar como varia este resultado, obtenido directamente sobre el cálculo de la distribución binomial al obtenido cuando aproximamos esta distribución al valor de la normal. Para ello bastará que calculemos, tal como vimos en una actividad anterior, la probabilidad a través de la función DISTR.NORM en la forma siguiente:

1-DISTR.NORM(100;n*p; RAIZ(n*p*(1-p));VERDADERO)

y obtendremos la aproximación usando la distribución Normal.

Aún más, puesto que al aproximar una distribución discreta a través de una distribución continua, como es el caso de la aproximación de la binomial a través de la normal, es habitual llevar a cabo la ***corrección por continuidad***, podemos analizar el impacto que tiene esta corrección en la precisión del resultado anterior.

La hoja dedicada a la resolución de esta actividad se muestra en la página siguiente.



[4] Exacto en el sentido de que no se recurre a la aproximación a la distribución Normal aunque difícilmente creemos que Excel haga el cálculo exacto:

$$P = \sum_{x=0}^{x=100} \binom{200}{x} 0,4^{x} \cdot 0,6^{200-x}$$

14.5 Actividad 5

Preguntamos a una muestra aleatoria de 50 familias ¿qué presupuesto mensual tienen para gastos médicos?. La media de las respuestas es 8.340 euros, y la desviación estándar de la población sabemos que es 2.570 euros. Calculad un intervalo de confianza del 95% para el presupuesto medio mensual para las familias de la población.

La teoría de estimación nos dice que, puesto que la distribución de la media muestral es $N_{(\mu,\sigma/\sqrt{n})}$ podemos deducir un intervalo al (1-α)% de confianza como :

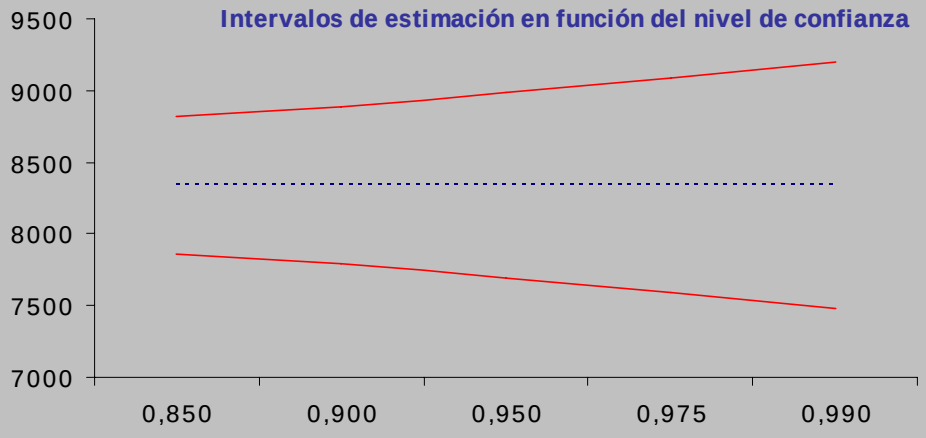
$$\bar{x} \mp z_{(\alpha/2)} \frac{\sigma}{\sqrt{n}}$$

una sencilla aplicación de la instrucción DISTR.NORM nos permitirá hacer todos los cálculos necesarios en Excel. Si queremos observar como incide la elección del nivel de confianza en la amplitud del intervalo de estimación. O cómo varia éste al variar, manteniéndose fijos los demás factores, el tamaño de la muestra, podemos construir fácilmente una hoja como la siguiente:

Tamaño de la muestra 60

Media muestral 8340
Desviación tipíca muestral 2570

1-α	0,850	0,900	0,950	0,975	0,990
Z(α/2)	1,44	1,64	1,96	2,24	2,58
Error estimación	477,62	545,74	650,29	743,66	854,62
Límite inferior	7862,4	7794,3	7689,7	7596,3	7485,4
Valor medio	8340	8340	8340	8340	8340
Limite superior	8817,6	8885,7	8990,3	9083,7	9194,6



Actividad 5
Preguntamos a una muestra aleatoria de **50** familias qué presupuesto mensual tienen para gastos médicos. La media de las respuestas es 8.340 ptas., y la desviación estándar de la población sabemos que es 2.570 ptas. Calculad un intervalo de confianza del **95%** para el presupuesto medio mensual para las familias de la población.

14.6 Actividad 6

A partir de una muestra aleatoria de 1.492 adultos, se vio que el 35% estaba a favor de incrementar el precio de la gasolina para subvencionar las autopistas. Calculad el intervalo de confianza del 95% para el verdadero porcentaje de adultos de la población que tengan esta opinión.

Como siempre, resolveremos primero aplicando la teoría conocida. Ésta nos dice que es posible obtener el intervalo de confianza de la proporción pedida gracias a que sabemos que la proporción muestral se distribuye de forma normal y con parámetros conocidos, exactamente sabemos que:

$$\hat{p} \cong N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

de lo cual deducimos que un intervalo del (1-α)% de confianza puede construirse de la forma siguiente:

$$\hat{p} \mp z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}\ (1-\hat{p})}{n}}$$

La construcción de este intervalo en Excel no requiere de instrucciones específicas y puede hacerse directamente a través de las operaciones aritméticas normales; el valor $z_{(\alpha/2)}$ ya sabemos, podemos obtenerlo gracias a la instrucción DISTR.NORM. Lo que si podremos gracias a Excel es generalizar, no sólo sobre los datos iniciales variando el porcentaje inicial de aceptación, sino variando el nivel de confianza para observar como varia la amplitud del intervalo de estimación al aumentar éste.

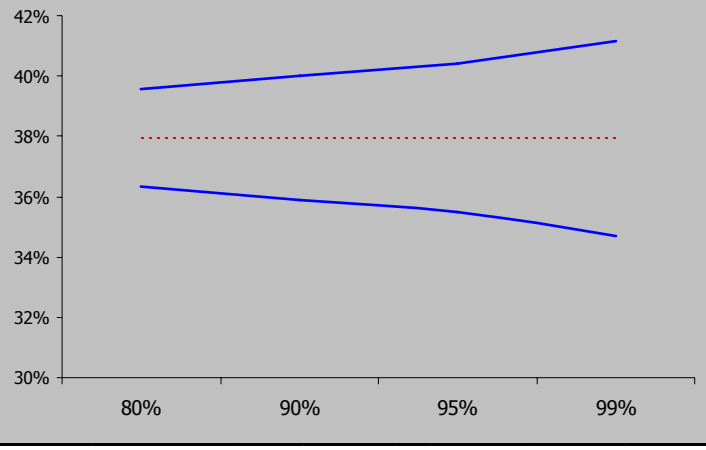
La única dificultad reside en el hecho de hacer el cambio adecuado desde el nivel de confianza dado (1-α)% al valor que es necesario introducir en la fórmula para calcular el intervalo de estimación z(α/2).

Actividad 6

A partir de una muestra aleatoria de 1.492 adultos, se vio que el **35**% estaba a favor de incrementar el precio de la gasolina para subvencionar las autopistas. Calculad el intervalo de confianza del 95% para el verdadero porcentaje de adultos de la población que tengan esta opinión.

Entrevistados	1492	A favor	En contra
		566	926
		37,9%	62,1%

1-α	1-(α/2)	Z	Error	Inferior	Central	Superior
80%	0,900	1,28	0,013	36,3%	37,9%	39,5%
90%	0,950	1,64		35,9%	37,9%	40,0%
95%	0,975	1,96		35,5%	37,9%	40,4%
99%	0,995	2,58		34,7%	37,9%	41,2%



14.7 Actividad 7

Calculad el área en la cola de la distribución t con 24 grados de libertad a la derecha del valor 2,56.

Grados de libertad	0,01	0,005
23	2,4999	2,8073
24	2,4922	2,7969

Notemos primero que la respuesta a esta pregunta, disponiendo únicamente de las tablas, es muy aproximada ya que sólo podríamos deducir que el área pedida estará comprendida entre 0,010 y 0,005

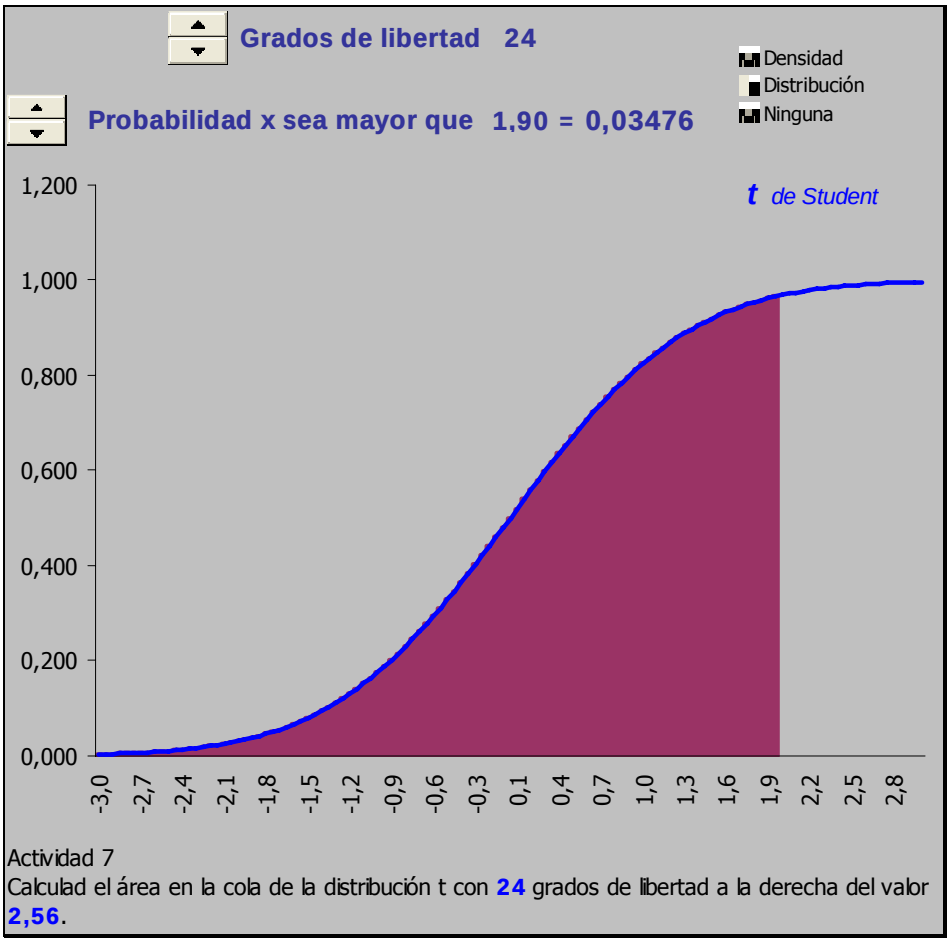
Al disponer de un ordenador y de un software con ciertas capacidades estadísticas podemos contestar a la pregunta de forma exacta. La función Excel que esta relacionada con la variable aleatoria t de Student es DISTR.T, función cuya sintaxis es **DISTR.T(x ;gl; colas)**, siendo x el valor numérico en el que se ha de evaluar la distribución; gl el número de grados de libertad y colas un entero con dos posibles valores (1 y 2) que nos permitirá indicar si nos referimos a (1-α) o a (1-α/2).

La función devuelve la función de distribución de una variable t de Student es decir la probabilidad $P_{(t<x)}$ con t ≈ $t_{gl.}$

Bastará entonces, para obtener la probabilidad pedida, insertar la fórmula siguiente en una celda de la hoja de cálculo:

=DISTR.T(2,56;24;1)

El valor obtenido (0,00859) se encuentra, tal como ya habíamos deducido de las tablas, entre el 1% y el 0,5%.



14.8 Actividad 8

Para una distribución t de Student con 55 grados de libertad ¿cuál es la probabilidad de que la variable aleatoria se encuentre comprendida entre los valores –1,96 y 1,96?.

La actividad es muy parecida a la anterior, se resuelve de nuevo usando la función incluida en la librería de funciones estadísticas

=DISTR.T(x; gl ;colas)

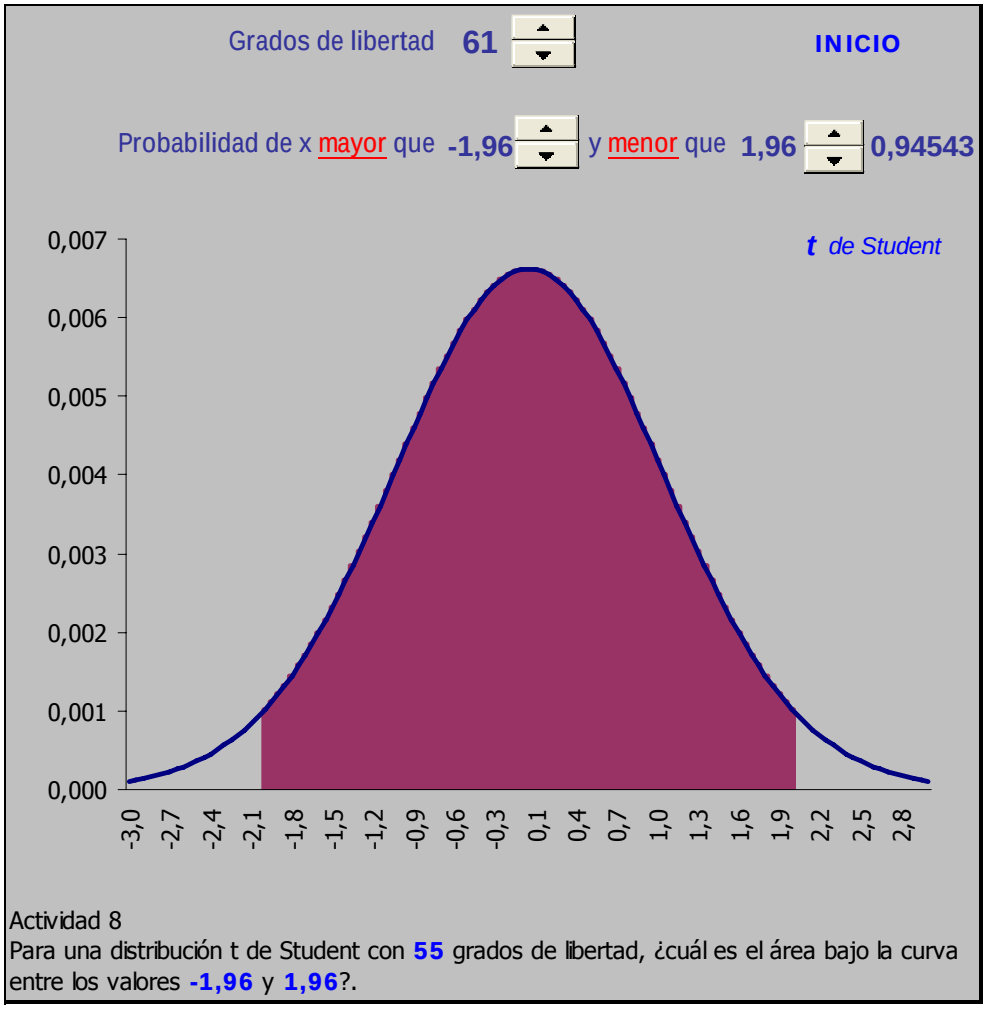
Ahora, puesto que lo que nos piden es:

$$P(\alpha \leqslant t_{55} \leqslant \beta)$$

usaremos una formulación del tipo:

DISTR.T(ABS(x);gl;1)
1-DISTR.T(x;$gl;1)

Para calcular los valores de cada extremo (dependiendo de que x sea negativo o positivo respectivamente) y restaremos los valores obtenidos para calcular la probabilidad pedida. La hoja de calculo que generaliza esta actividad para diversos valores de los grados de libertad y los extremos α y β, tiene el siguiente aspecto:



14.9 Actividad 9

Supongamos que el área entre dos puntos -t y +t, simétricos en torno a cero, es igual a 0,90. Encontrad los valores de t para una distribución t con:

a) 9 grados de libertad;
b) 99 grados de libertad;
b) 999 grados de libertad.

A diferencia de las actividades anteriores en las que lo que buscábamos era la probabilidad de la variable aleatoria t asociada con uno o dos valores, ahora lo que nos piden es encontrar, para una probabilidad dada **p**, el valor α tal que:

$$P(\ \alpha \leqslant t_{gl} \leqslant \alpha) = p$$

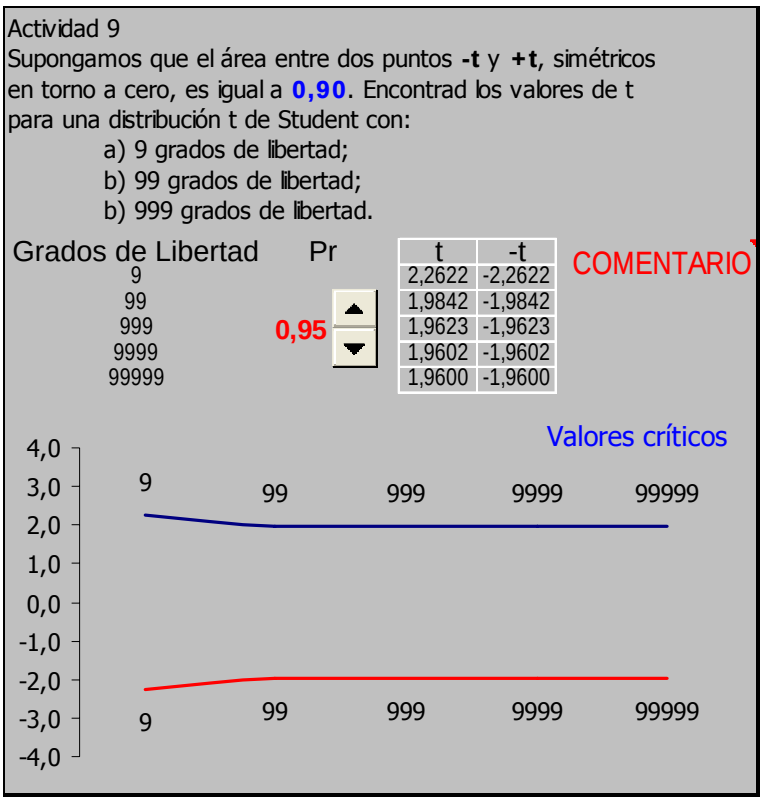
Excel tiene una función que permite encontrar ese valor (conocido como valor crítico), se trata de :

DISTR.T.INV(**p**;gl)

función que devuelve el valor α de la distribución t de Student como función de la probabilidad **p** y los grados de libertad gl. Para ser exactos, el resultado es el valor α, tal que:

$$P(t_{gl} \geqslant \alpha) = p$$

El aspecto de la hoja que generaliza esta actividad para diversos valores de **p** es el siguiente:



Podremos cómo la amplitud del intervalo varia para valores de **gl** menores que 100, pero cómo a partir de ese número no existe prácticamente ninguna variación. También apreciamos cómo al aumentar el valor de **gl** la diferencia entre la ***t*** y la Normal tiende a desaparecer.

14.10 Actividad 10

Calculad el intervalo de confianza del 95% para la media de una población si tenemos una muestra aleatoria de 41 observaciones con media muestral igual a 105,1 y varianza muestral igual a 13,24.

La teoría nos dice que al extraer una muestra de **n** observaciones de una población normal es posible construir, con una confianza dada, un intervalo para la media de la población de la que dicha muestra procede. Basta aplicar el hecho de que la distribución de la media muestral es normal, que su media coincide con la media poblacional y que su desviación está en función de la desviación típica poblacional y del tamaño de la muestra extraída.

Concretamente este intervalo se construye de la forma siguiente:

$$\bar{x} \mp z_{(\alpha/2)} \frac{\sigma_x}{\sqrt{n}}$$

Cuando, como es habitual, la desviación típica de la población es también desconocida y es necesario estimarla a partir de los datos de la muestra, la media muestral no se distribuye de forma normal sino como una distribución ***t*** de Student.

En este caso el intervalo se construye de la forma siguiente:

$$\bar{x} \mp t_{(\alpha/2,n\ 1)} \frac{s_x}{\sqrt{n}}$$

Para construir intervalos de confianza de una forma u otra bastará conocer los valores críticos de las distribuciones implicadas y hacer unos sencillos cálculos. No obstante, Excel tiene una función en su librería de funciones estadísticas que calcula la amplitud del intervalo de confianza para la Normal, el caso menos general, se trata de :

INTERVALO.CONFIANZA(α ; s_x ; n)

El valor que devuelve la función es, como hemos indicado, la anchura del intervalo de confianza, es decir:

$$\mathbf{z}_{(\alpha/2)} \frac{\sigma_{\mathbf{x}}}{\sqrt{\mathbf{n}}}$$

Podemos preguntarnos cómo variará el intervalo de confianza, no sólo al variar el tamaño de la muestra o la variabilidad de ésta, sino al suponer que se verifican las condiciones del Teorema Central de Límite bien porque el tamaño de la muestra es lo suficientemente grande, bien porque sabemos que la población subyacente es normal.

En otras palabras, queremos saber si al estimar la media muestral mediante:

$$\bar{x} \mp z_{(\alpha/2)} \frac{\sigma_x}{\sqrt{n}}$$

existe una diferencia respecto al estimarla mediante:

$$t_{(\alpha/2,n\ 1)} \frac{s_x}{\sqrt{n}}$$

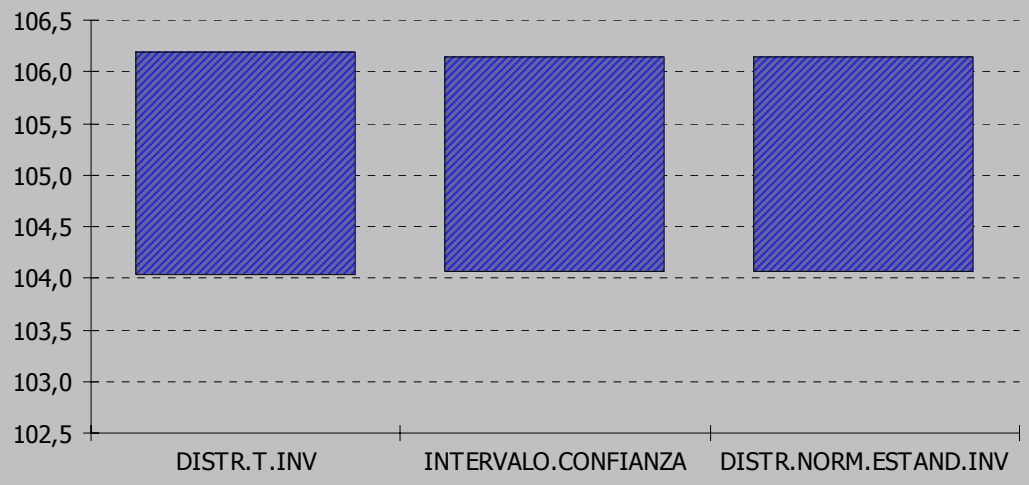
La hoja en la que hemos resuelto esta actividad tiene el aspecto siguiente

Actividad 10

Calculad el intervalo de confianza del **95 %** para la media de una población si tenemos una muestra aleatoria de **41** observaciones con media muestral igual a 105,1 y varianza muestral igual a 13,24.

Nivel de confianza	**0,850**
N° de observaciones	**25**
Media muestral	**105,1**
Varianza muestral	**13,24**

Inf	**Sup**	
104,02	**106,18**	DISTR.T.INV
104,05	**106,15**	INTERVALO.CONFIANZA
104,05	**106,15**	DISTR.NORM.ESTAND.INV



Observamos que para los valores del problema, (α, n, datos de la muestra),

Nivel de confianza	**0,850**
N° de observaciones	**41**
Media muestral	**105,1**
Varianza muestral	**13,24**

obtenemos tres estimaciones de la media muestral:

Inf	**Sup**	
104,02	**106,18**	DISTR.T.INV
104,05	**106,15**	INTERVALO.CONFIANZA
104,05	**106,15**	DISTR.NORM.ESTAND.INV

la primera se corresponde al intervalo calculado a través de la ***t*** de Student, utilizando la inversa de la función de distribución:

$$\mu \pm \text{DISTR.T.INV}(1-\alpha\ ;\ n) * \text{RAIZ}\ (s_x / n)$$

;la segunda utilizando la función INTERVALO.CONFIANZA:

$$\mu \pm \text{INTERVALO.CONFIANZA}\ (1-\alpha\ ;\ \text{RAIZ}\ (s_x)\ ;\ n)$$

;la tercera aplica la inversa de la función de distribución normal:

$$\mu \pm \text{DISTR.NORM.ESTAND.INV}\ ((1-\alpha)/2)*\text{RAIZ}(s_x)/n$$

Como podemos apreciar al alcanzar n un tamaño medio (<30) las diferentes aproximaciones (t y Normal) proporcionan resultados muy próximos entre si. También notamos como las formulaciones segunda y tercera proporcionan idénticos resultados.

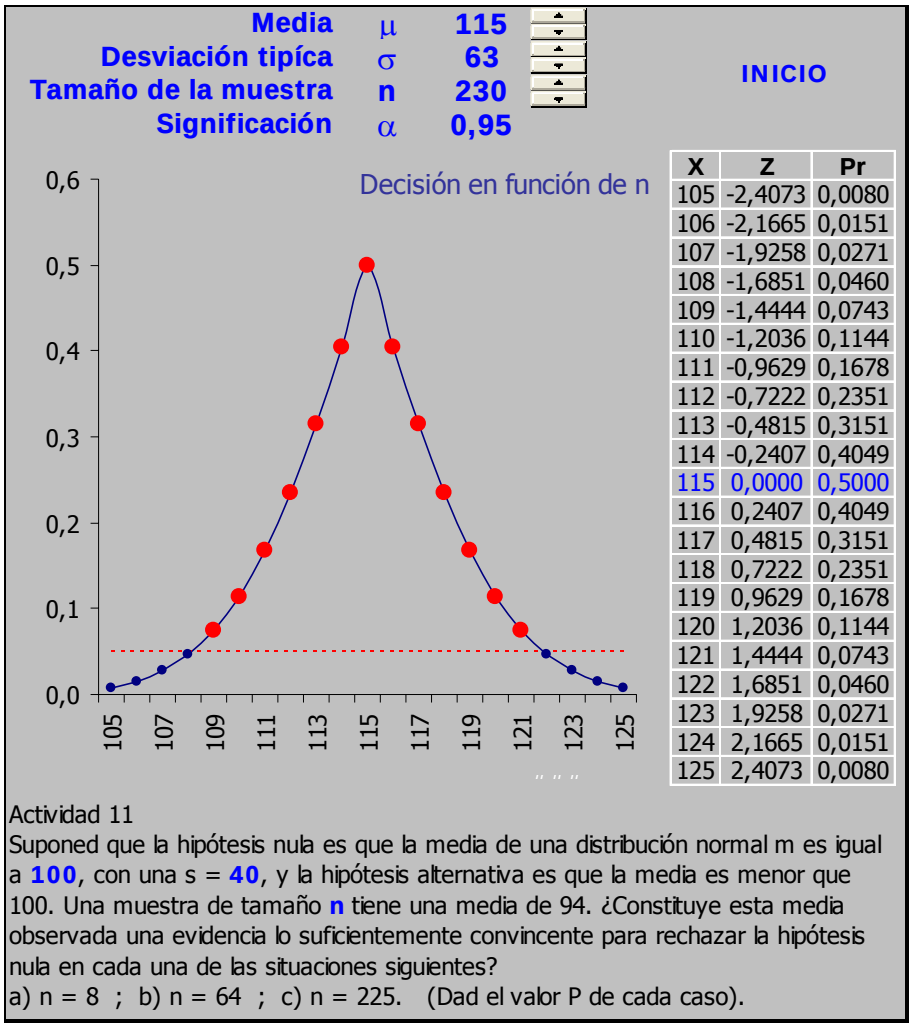
14.11 Actividad 11

Suponed que la hipótesis nula es que la media de una distribución normal m es igual a 100, con una s = 40, y la hipótesis alternativa es que la media es menor que 100. Una muestra de tamaño n tiene una media de 94. ¿Constituye esta media observada una evidencia lo suficientemente convincente para rechazar la hipótesis nula en cada una de las situaciones siguientes?:

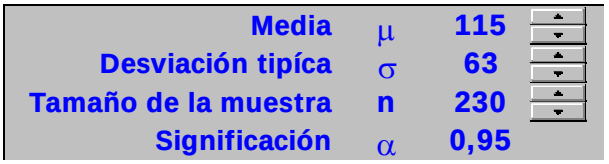
a) n = 8
b) n = 64
c) n = 225.

(Dad el valor P de cada caso).

La hoja de cálculo mediante la que resolvemos la actividad es la siguiente:

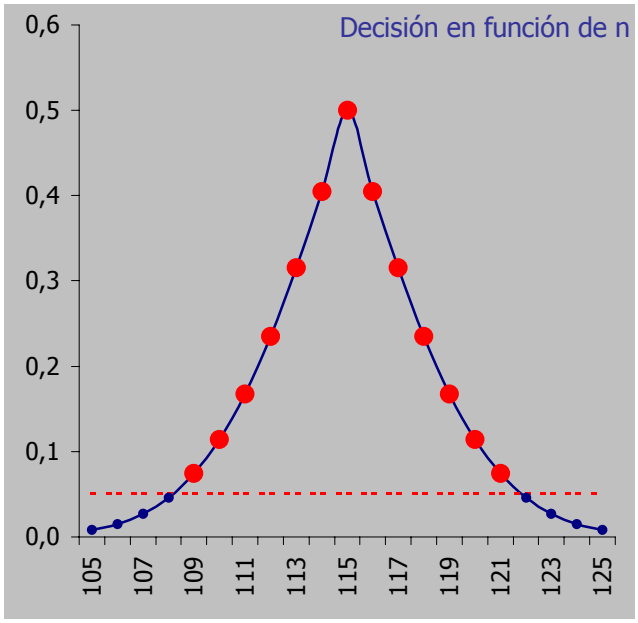


Como siempre dispone de una zona en la que el usuario puede introducir otros valores diferentes a los propuestos



de una tabla y de un gráfico (con ciertas capacidades añadidas) de los valores de la tabla anterior:

X	Z	Pr
105	-2,4073	0,0080
106	-2,1665	0,0151
107	-1,9258	0,0271
108	-1,6851	0,0460
109	-1,4444	0,0743
110	-1,2036	0,1144
111	-0,9629	0,1678
112	-0,7222	0,2351
113	-0,4815	0,3151
114	-0,2407	0,4049
115	0,0000	0,5000
116	0,2407	0,4049
117	0,4815	0,3151
118	0,7222	0,2351
119	0,9629	0,1678
120	1,2036	0,1144
121	1,4444	0,0743
122	1,6851	0,0460
123	1,9258	0,0271
124	2,1665	0,0151
125	2,4073	0,0080



El proceso es muy sencillo, explicaremos en primer lugar los valores de la tabla, la columna **X** contiene valores de la variable aleatoria en la cercanía del valor propuesto por el usuario como media (115 en el ejemplo); la columna **Z** calcula el estadístico para el contraste que estamos realizando, es decir:

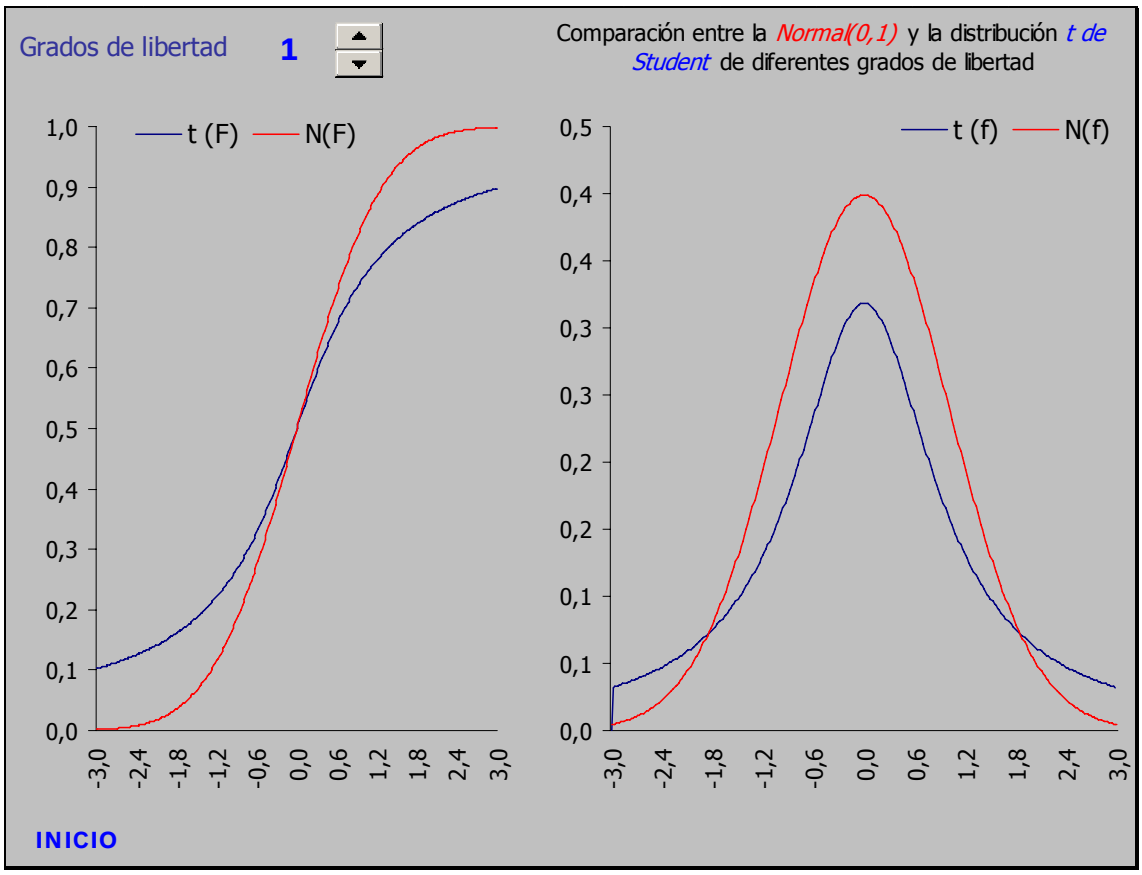
$$Z = \frac{\bar{x} - \mu_o}{\left(S_x / \sqrt{n} \right)}$$

Finalmente, en la columna **Pr,** se calcula la probabilidad asociada a una discrepancia como la recién calculada, bajo la hipótesis nula.

La lógica del contraste nos dice que cuando esta probabilidad sea superior a α podremos mantener la hipótesis nula sosteniendo que la diferencia entre la media muestral observada y la media teórica es, únicamente, producto del azar; si por el contrario esta probabilidad es inferior a α, deberemos rechazar a hipótesis nula ya que la evidencia en su contra es demasiado patente.

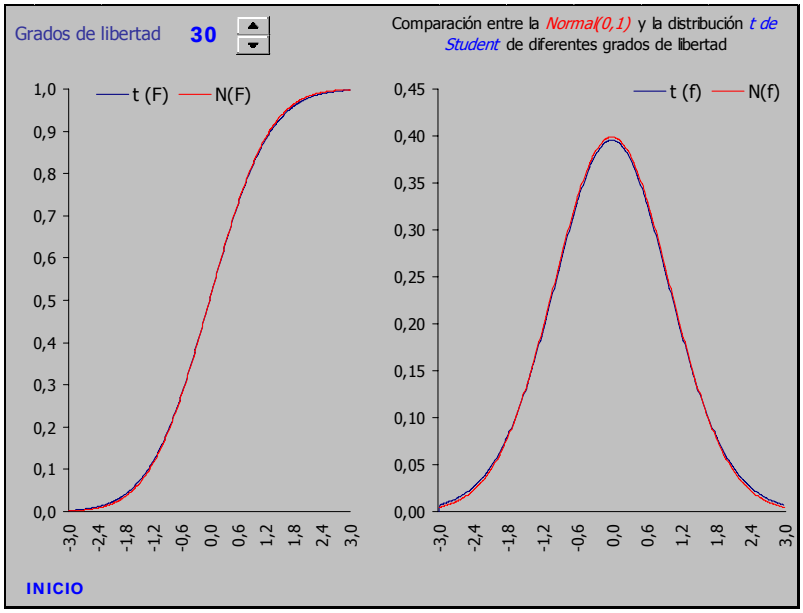
El gráfico representa los valores de **Pr**, señalando aquellos que son inferiores a α y que dan lugar a la región de rechazo que, puesto que la hipótesis nula es bilateral, está repartido simétricamente a ambos lados de valor teórico.

Finalmente hemos incluido una última hoja en la que llevamos a cabo una comparación a través de la observación de las gráficas correspondientes a las funciones de densidad y distribución, de las variable aleatorias Normal y ***t*** de Student. Esta hoja tiene el aspecto siguiente:



El único valor que el usuario puede cambiar es el de los grados de libertad de la distribución t de Student, ya que la comparación se realiza siempre respecto de la Normal(0;1).

Apreciamos que las diferencias son notables para grados de libertad reducidos, pero que, como ya hemos podido apreciar en las actividades anteriores, estas diferencias desaparecen al aumentar este valor.



14.12 Actividad 12

Un club de esquí organiza un curso de buena forma física de dos semanas para ejecutivos. Hace que se pesen cinco de los participantes seleccionados al azar antes del curso y después del curso. Contrastad si ha habido una reducción de peso significativa (contrastadlo al nivel del 5% y suponed que hay una distribución normal para los datos).

Lo primero que debemos notar es que, por las circunstancias del problema, los datos son "emparejados": una misma persona es pesada antes y después de manera que cada par de datos de los que forman las muestras a comparar están referidos a un mismo objeto estadístico. Como es lógico, en estos casos los tamaños muestrales son idénticos.

La teoría nos dice que el problema, tal como ha sido planteado, consiste en la contrastación de una hipótesis de la forma siguiente:

$$\begin{cases} \mathbf{H_o} : \mathbf{P_{antes}} \geq \mathbf{P_{después}} \\ \mathbf{H_1} : \mathbf{P_{antes}} < \mathbf{P_{después}} \end{cases}$$

que, alternativamente, podemos plantear también como:

$$\begin{cases} \mathbf{H_o} : \left(\mathbf{P_{antes}} - \mathbf{P_{después}}\right) \geq 0 \\ \mathbf{H_1} : \left(\mathbf{P_{antes}} - \mathbf{P_{después}}\right) < 0 \end{cases}$$

Para lo cual debemos calcular un estadístico de contraste de la forma:

$$\mathbf{dis} = \frac{\overline{\mathbf{X}}_{\mathbf{dif}}}{\mathbf{s_x} / \sqrt{\mathbf{n}}}$$

que se distribuye con arreglo a una distribución t de Student, es decir:

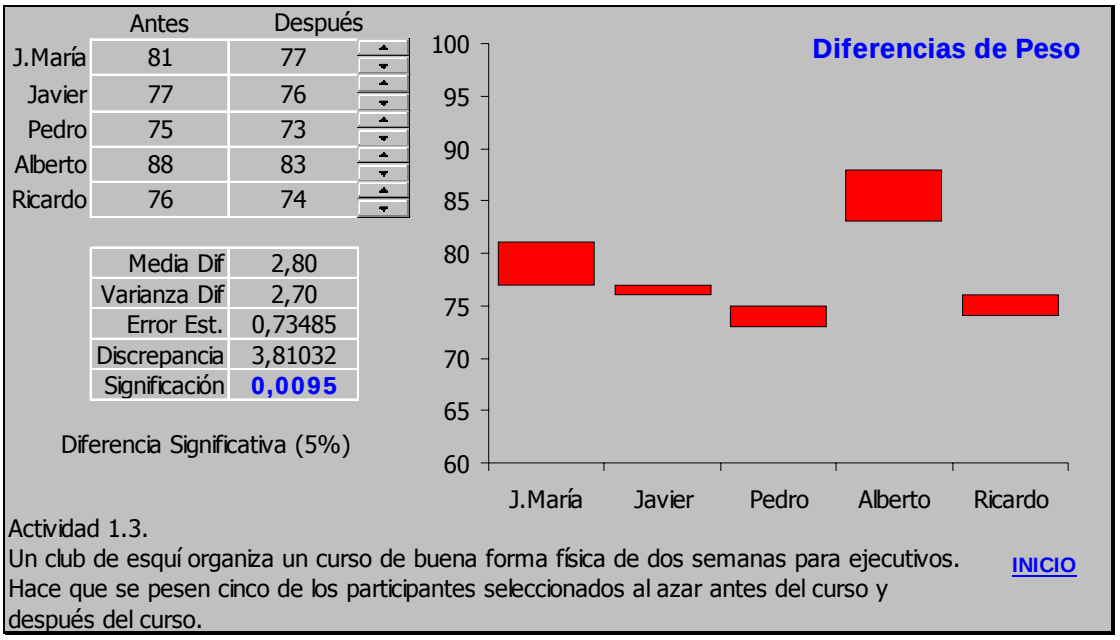
$$\mathbf{dif_i} = \left(\mathbf{P_{antes}} - \mathbf{P_{después}}\right)_\mathbf{i} \Rightarrow \mathbf{dis} \approx \mathbf{t_{n-1}}$$

Sabido esto, la resolución de la actividad es sencilla: calculamos la media de las diferencias de peso, su cuasi-desviación típica, calculamos la discrepancia **dis** y finalmente aplicamos la función, ya conocida, DISTR.T para obtener el p.valor de la prueba. Exactamente eso es lo que hace la hoja de cálculo que resuelve la actividad, básicamente se reduce a realizar los cálculos siguientes[5]:

a	Media Dif	={PROMEDIO(Despues-Antes)}
b	Varianza Dif	={VAR(Despues - Antes)}
c	Error Est.	=(**b**/5)^0,5
d	Discrepancia	=(**a**/**c**)
e	Significación	=DISTR.T(**d**;4;1)

El aspecto de la hoja es el que aparece en la página siguiente:

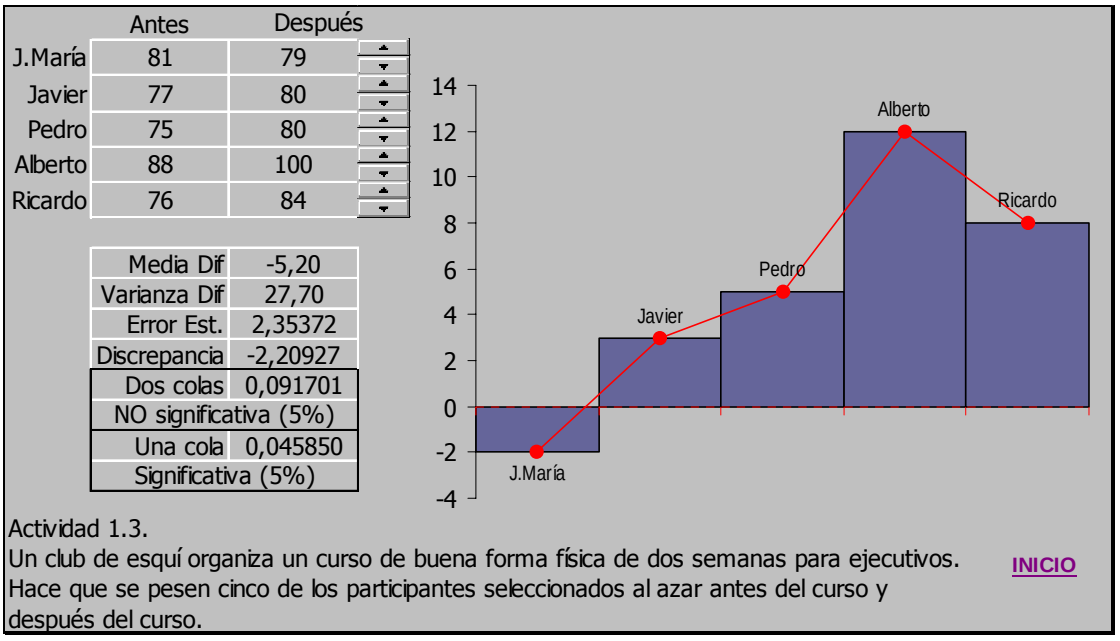
[5] Nótese el empleo en las dos primeros cálculos de formulas matriciales, que aparecen escritas entre llaves {} cuando el usuario las introduce de la forma habitual, es decir usando la combinación de teclas **Ctrl Shift Enter** .



El usuario puede modificar los valores de los pesos después de la dieta para observar cómo estas variaciones afectan al sostenimiento de la hipótesis. Una forma alternativa a esta es a través del empleo de una función de librería específicamente destinada a contrastar hipótesis de medias, nos referimos a la función

PRUEBA.T (matriz1 ; matriz2; colas ; tipo)

siendo D1 el rango que contiene el primer conjunto de datos; D2 el rango del segundo conjunto de datos; colas (1 ó 2) especifica si la hipótesis es unilateral o bilateral; y Tipo es un entero (1,2 o 3) que indica el tipo de prueba t que se realiza. (1 para muestras emparejadas; 2 para el caso general e iguales varianzas; 3 para el caso general y varianzas diferentes). El empleo de esta función nos permite ampliar el experimento, para incluir el contraste de la hipótesis nula "los pesos antes y después son diferentes" (2 colas). El aspecto de la resolución alternativa es el siguiente:



14.13 Actividad 13

Hasta el momento se sabía que el porcentaje de a favor de una determinada opción era del 52%. Repetida la encuesta entre 1500 personas, el porcentaje ha bajado al 46%. ¿Es compatible el nuevo resultado con lo aceptado anteriormente?.

El procedimiento adecuado para contestar a la pregunta implícita en la actividad es el de contraste de hipótesis, concretamente el de una proporción muestral respecto de una proporción de referencia, esto es:

$$H_o : \hat{p} = \pi_0$$
$$H_1 : \hat{p} \neq \pi_0$$

la teoría nos dice que al calcular una discrepancia de la forma:

$$dis = \frac{\hat{p} - \pi_0}{\sqrt{\frac{\pi_0(1 - \pi_0)}{n}}}$$

ésta se distribuye de forma normal, exactamente tendremos que **dis** $\approx N_{(0;1)}$.

Llevar a cabo el contraste es entonces muy simple, basta con calcular el error estándar y a continuación el valor de la discrepancia. Una vez calculado éste bastará con aplica la función de librería incluida en Excel que calcula la probabilidad asociada a una valor de x cuando éste valor proviene de una variable aleatoria normal estándar

DISTR.NORM.ESTAND(dis)

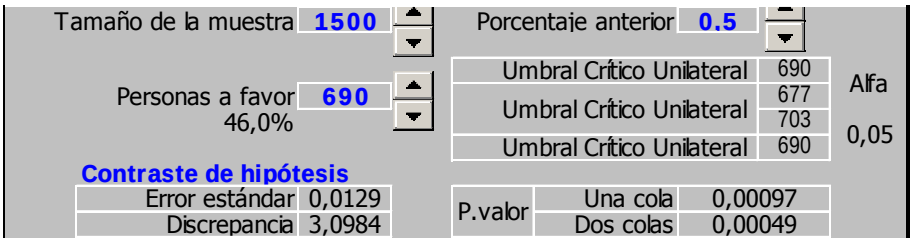
Si queremos calcular el p.valor de la prueba también para el caso de hipótesis unilateral deberemos utilizar

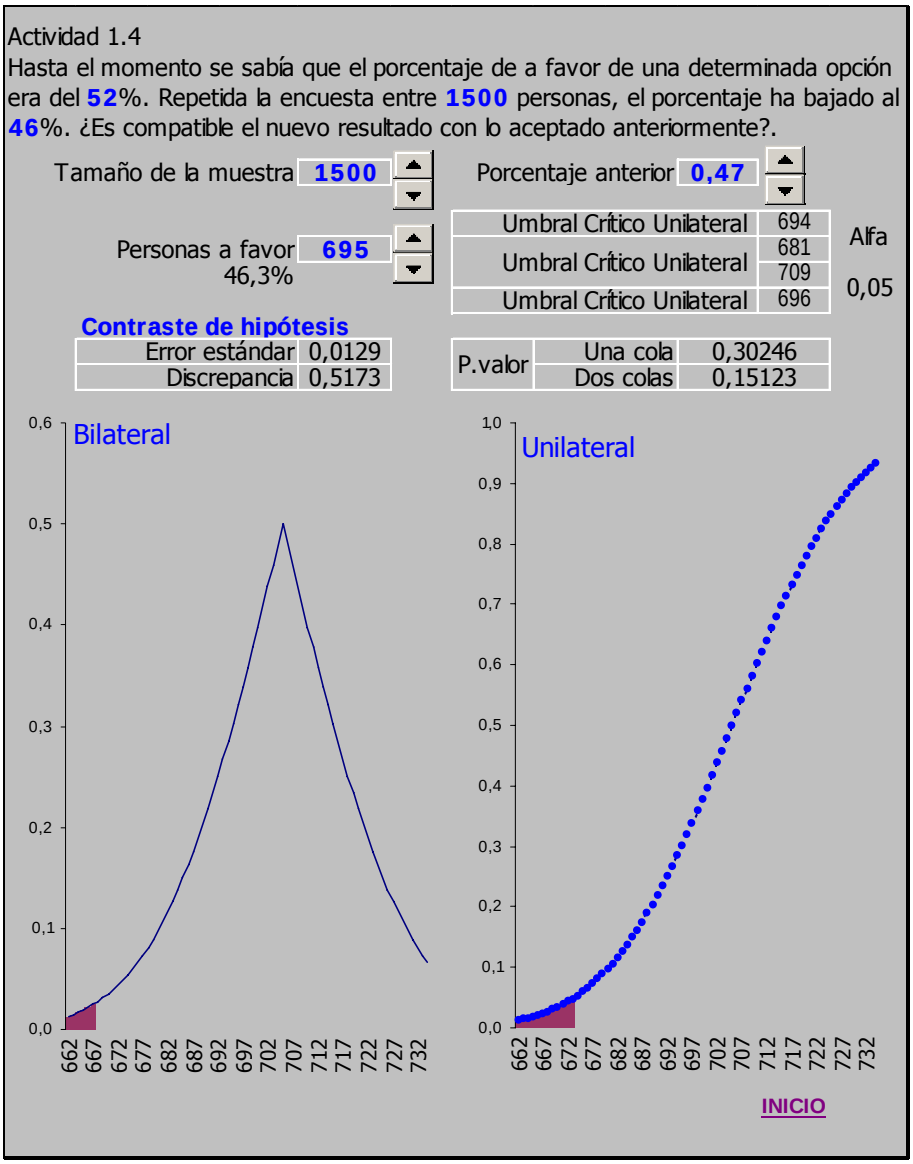
DISTR.NORM.ESTAND(dis/2)

Y para evitar problemas con el sentido de la discrepancia (téngase en cuenta que el orden de los sumandos es arbitrario) deberemos usar una fórmula del tipo:

=SI(dis>0;1-DISTR.NORM.ESTAND(dis);DISTR.NORM.ESTAND(dis))

El aspecto de la hoja que resuelve esta actividad es el que aparece en la página siguiente. En ella se ha generalizado el problema para permitir que el usuario modifique los valores que intervienen en él.





Se muestran dos tipos de resultados:

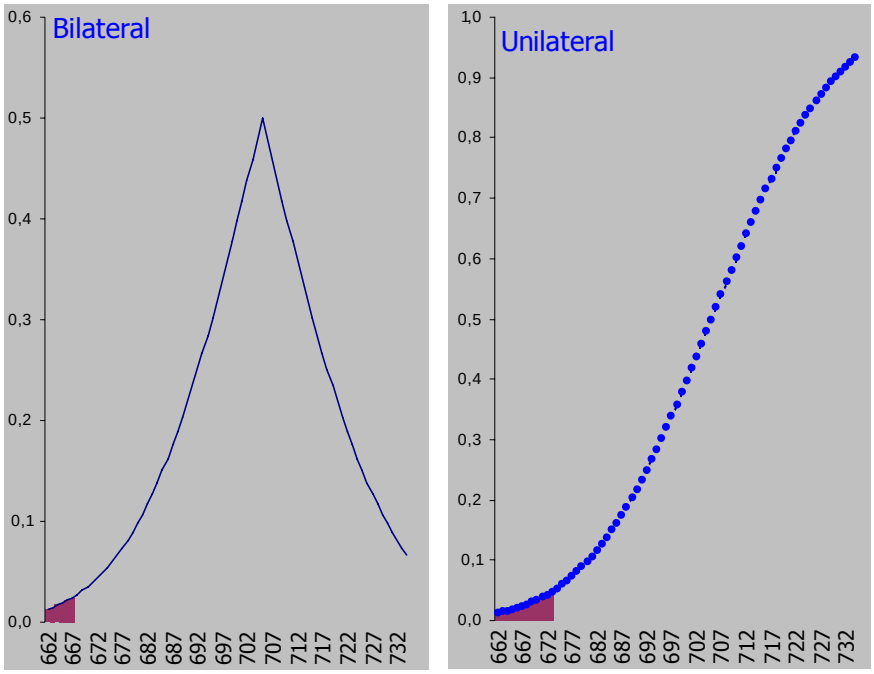
a) Los correspondientes al contraste de hipótesis:

Contraste de hipótesis				
Error estándar	0,0129	P.valor	Una cola	0,30246
Discrepancia	0,5173		Dos colas	0,15123

b) y los relacionados con el cálculo de los valores críticos, es decir el número de personas que habrían de estar a favor para mantener/rechazar la hipótesis nula frente a las tres posibles alternativas

Umbral Crítico Unilateral	694
Umbral Crítico Unilateral	681
	709
Umbral Crítico Unilateral	696

Los gráficos son las representaciones de los p.valores y las regiones de rechazo de los diferentes contrastes al variar el valor de la discrepancia en un entorno del valor propuesto.



calculados gracias a la tabla de valores que también figura en la hoja:

N	p(obs)	tst	P(tst)	Umbral	P(tst)	Umbral
662	0,44133	-2,227	0,01299	0,0130	0,01299	0,0130
663	0,44200	-2,175	0,01483	0,0148	0,01483	0,0148
664	0,44267	-2,123	0,01688	0,0169	0,01688	0,0169
665	0,44333	-2,071	0,01917	0,0192	0,01917	0,0192
666	0,44400	-2,019	0,02172	0,0217	0,02172	0,0217
667	0,44467	-1,968	0,02456	0,0246	0,02456	0,0246
668	0,44533	-1,916	0,02769	0,0277	0,02769	
669	0,44600	-1,864	0,03116	0,0312	0,03116	
670	0,44667	-1,812	0,03497	0,0350	0,03497	
671	0,44733	-1,760	0,03916	0,0392	0,03916	
672	0,44800	-1,709	0,04375	0,0438	0,04375	
673	0,44867	-1,657	0,04877	0,0488	0,04877	
674	0,44933	-1,605	0,05423		0,05423	
675	0,45000	-1,553	0,06017		0,06017	
676	0,45067	-1,502	0,06660		0,06660	
677	0,45133	-1,450	0,07355		0,07355	
678	0,45200	-1,398	0,08105		0,08105	
679	0,45267	-1,346	0,08911		0,08911	
680	0,45333	-1,294	0,09775		0,09775	
681	0,45400	-1,243	0,10699		0,10699	
682	0,45467	-1,191	0,11684		0,11684	
683	0,45533	-1,139	0,12732		0,12732	
684	0,45600	-1,087	0,13844		0,13844	
685	0,45667	-1,036	0,15020		0,15020	
686	0,45733	-0,984	0,16261		0,16261	

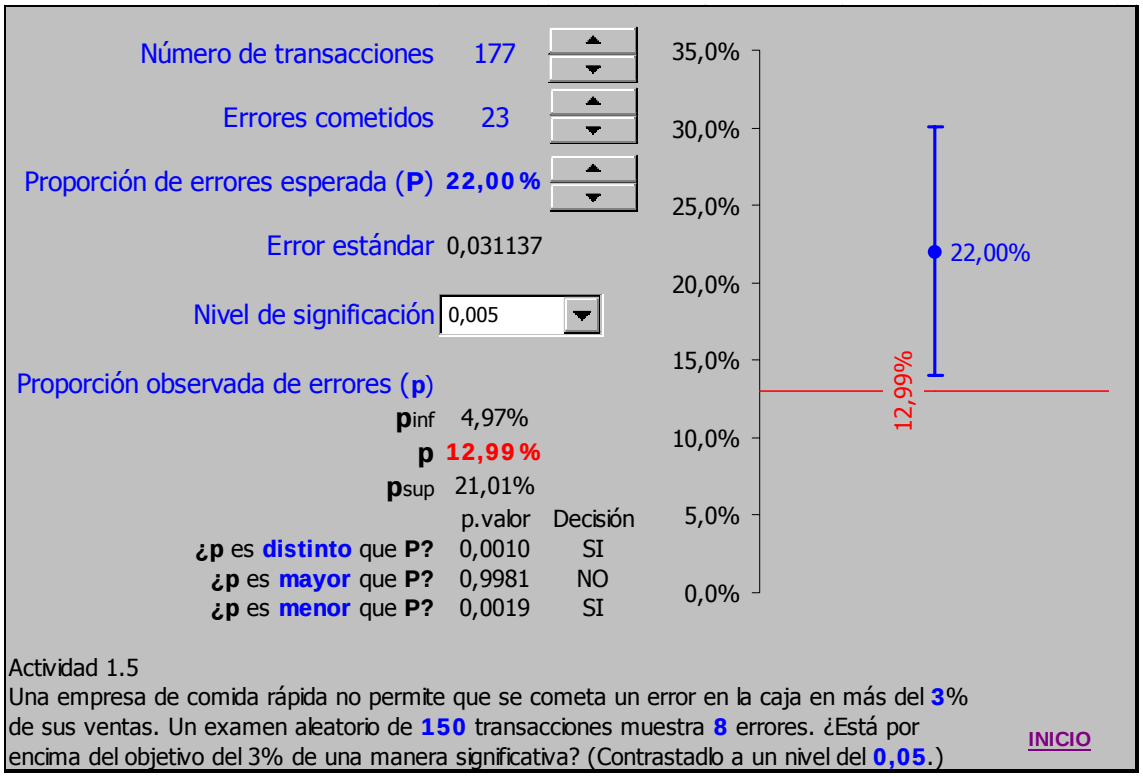
14.14 Actividad 14

Una empresa de comida rápida no permite que se cometa un error en la caja en más del 3% de sus ventas. Un examen aleatorio de 150 transacciones muestra 8 errores. ¿Está por encima del objetivo del 3% de una manera significativa? (Contrastadlo a un nivel del 0,05.).

La actividad es análoga a la anterior, se ha generalizado para permitir modificar los valores que definen el problema (transacciones, errores cometidos y proporción esperada), ahora se permite también al usuario modificar el nivel de confianza.

Los resultados son el error estándar y la decisión tomada (rechazar o mantener) según el valor obtenido para la discrepancia y el nivel de significación adecuado.

El gráfico incluido en la hoja de cálculo permite la representación gráfica de las principales magnitudes involucradas en el proceso de decisión.



14.15 Actividad 15

Una encuesta hecha a 1.000 españoles en el año 1994 reveló que 431 de las personas encuestadas piensan que la economía empeora. De estas 431 personas, 201 son hombres y 230 mujeres, mientras que en la muestra total hay 496 hombres y 504 mujeres. ¿Hay una diferencia significativa (a un nivel del 5%) entre la proporción de hombres y la de mujeres?.

El procedimiento adecuado para contestar a la pregunta implícita en la actividad es el de contraste de hipótesis, concretamente el de dos proporciones, esto es:

$$H_o : \hat{p}_A = \hat{p}_B$$
$$H_1 : \hat{p}_A \neq \hat{p}_B$$

la teoría nos dice que al calcular una discrepancia de la forma:

$$dis = \frac{\hat{p}_A - \hat{p}_B}{\sqrt{\hat{p}_C(1 - \hat{p}_C)\ \frac{1}{n_1} + \frac{1}{n_2}}}$$

siendo p_C la proporción conjunta (total de éxitos, esto es personas a favor, entre total de encuestados)

$$\hat{p}_C = \frac{e_1 + e_2}{n_1 + n_2}$$

ésta se distribuye de forma normal N(0;1).

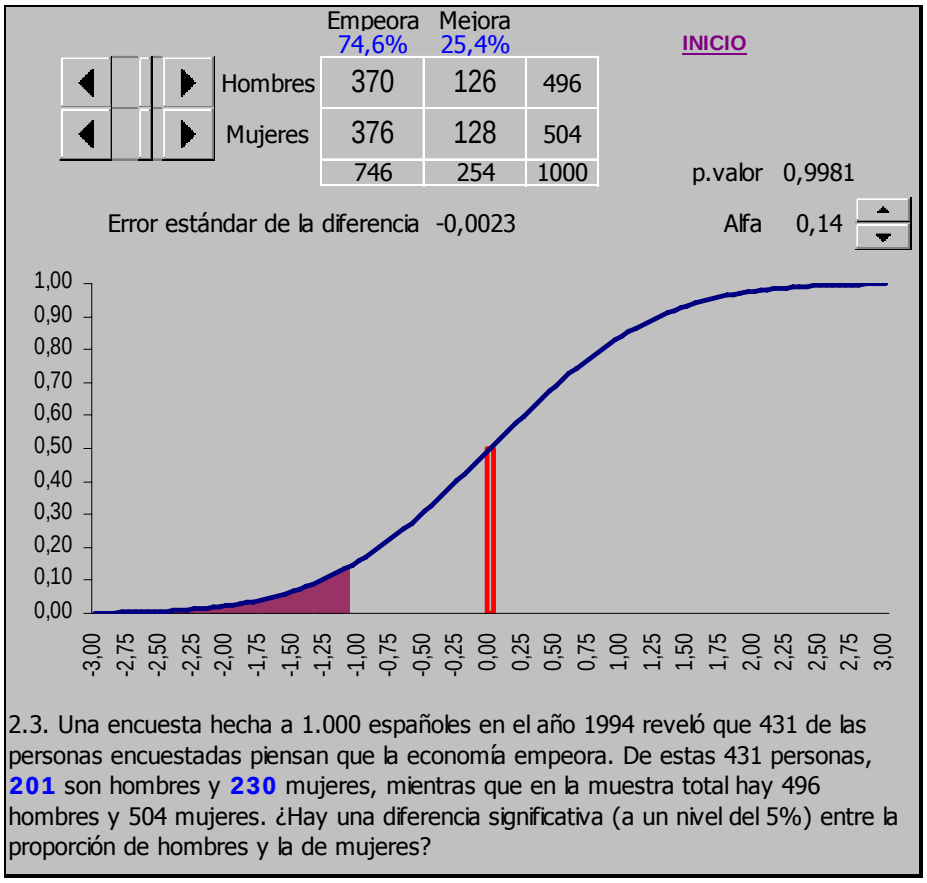
Tampoco ahora, como en el caso de las dos actividades anteriores en las que el contraste era de una proporción muestral respecto de una valor de referencia, existe en Excel una función especifica para realizar la prueba.

Sin embargo, la aritmética implicada en el contraste es, como hemos visto, extraordinariamente sencilla y no tendremos ningún problema al trasladas las operaciones necesaria para llevarla a cabo a la hija de cálculo.

La parte que no podremos hacer nosotros, el calculo de la probabilidad asociada a la discrepancia obtenida bajo la hipótesis nula, la podremos obtener con la función, ya expuesta anteriormente, que proporciona el valor de la función de distribución de la normal para un valor de x:

DISTR.NORM.ESTAND(dis)

El aspecto de la hoja que resuelve la actividad es el siguiente:



El usuario puede modificar los valores del número de hombres y mujeres que están a favor.

También, a diferencia de las hojas anteriores en las que el valor de nivel de significación estaba restringido a unos cuantos valores normalmente utilizados, el usuario dispone de absoluta libertad para elegir α.

La representación gráfica es la de la región de rechazo de la hipótesis nula (área sombreada bajo la curva), el valor dis, el estadístico obtenido (el segmento de color rojo) y la función de distribución de la normal estandarizada (línea azul continua)

14.16 Actividad 16

Basándonos en una muestra aleatoria de tamaño 10 de las mediciones de control de la contaminación del aire, calculamos que la varianza de la muestra es 14,2. Contrastad la hipótesis de que la varianza de la población es igual a 10 contra la alternativa de que haya aumentado (presuponed una distribución normal para las mediciones).

Tal como señala el propio enunciado de la actividad, ésta consiste en llevar a cabo un contraste sobre la varianza de una muestra (que se supone proviene de una población normal) respecto a un valor teórico o de referencia, es decir

$$H_o : s_x = \sigma_0$$
$$H_1 : s_x \neq \sigma_0$$

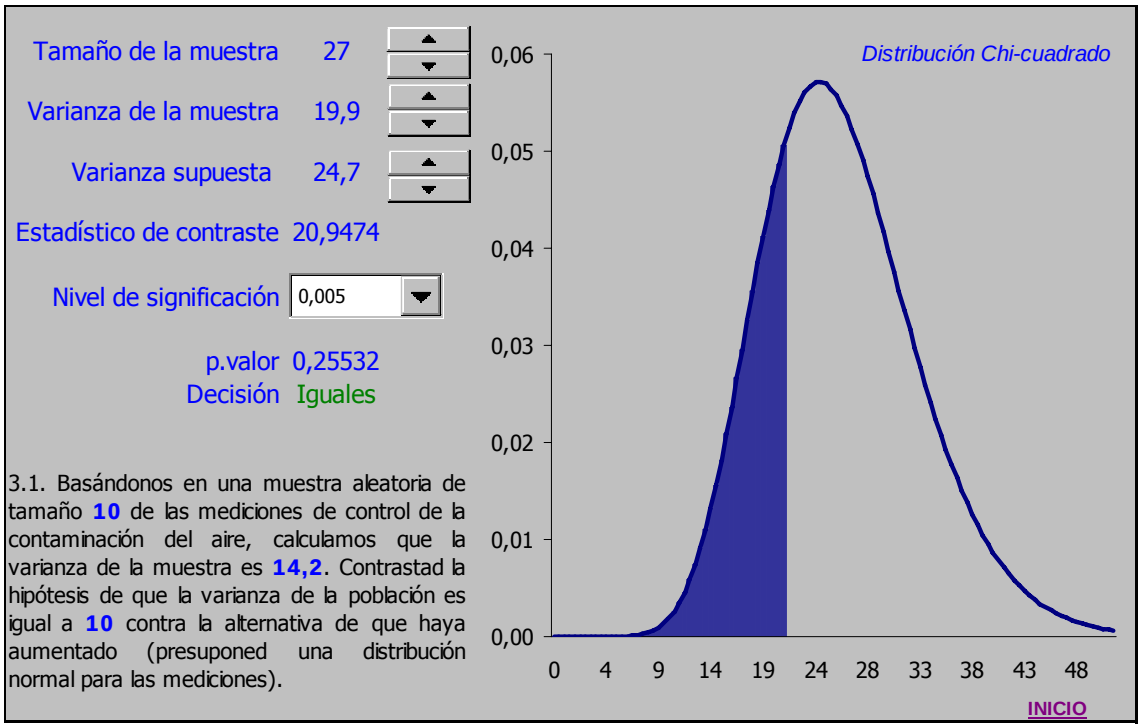
la teoría (repasar la página 22 del material y siguiente) nos dice que la discrepancia a usar en este tipo de contrastes es de la forma:

$$dis = (n-1)\frac{s_x^2}{\sigma^2}$$

Bastará entonces con realizar estos cálculos en la hoja hasta llegar a la obtención del valor de **dis**. Una vez obtenido éste utilizaremos la función que permite obtener probabilidades asociadas a variables aleatorias que se distribuyen según una χ^2

DISTR.CHI(x ; gl)

El aspecto de la hoja que resuelve la actividad es el siguiente



14.17 Actividad 17

Realizar una plantilla para llevar a cabo contrastes de medias a partir de los datos de la muestra.

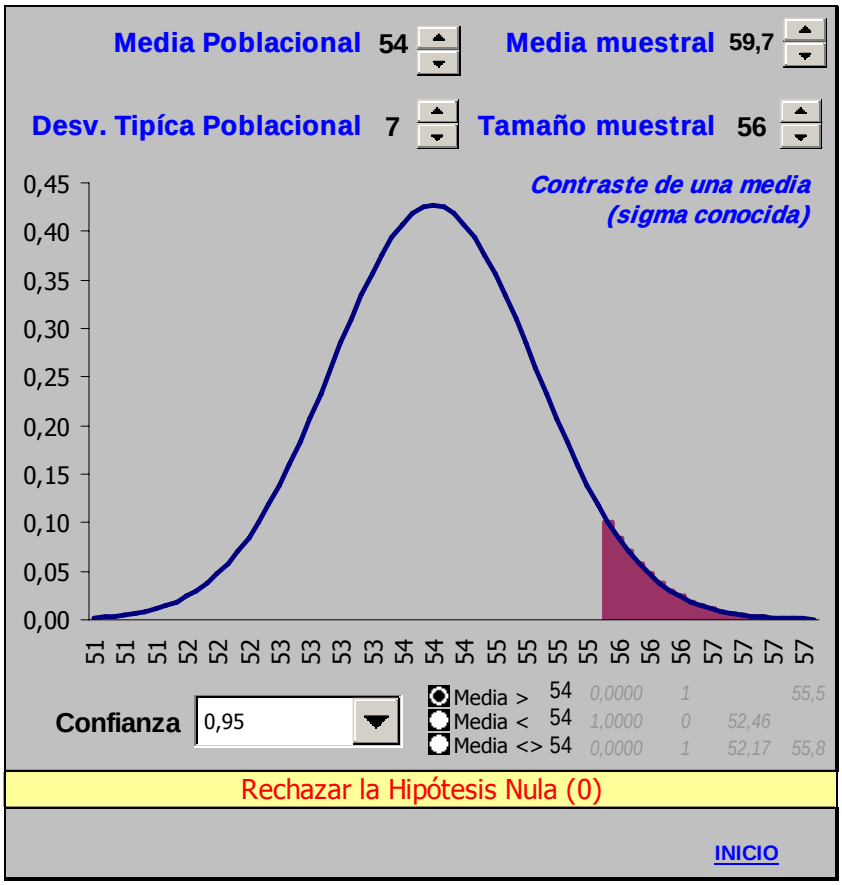
Para ello bastará, una vez introducidos todos aquellas mecanismos que permitan modificar los valores que intervienen en el proceso, que calculemos el error estándar, la discrepancia y el p.valor del contraste, operaciones todas ellas extraordinariamente sencillas.

Hecho esto, bastara aplicar la función de Excel asociada con la probabilidad de la distribución normal

DISTR.NORM(x ;Media; Desviación; VERDADERO)

Y si deseamos algún tipo de representación gráfica (la elegida por nosotros corresponde a la región de rechazo, será necesario crear una tabla de p.valores calculados en la proximidad de los valores propuestos para volcadlos en la gráfica correspondiente.

Esta hoja podría tener un aspecto como el siguiente. Apreciamos en ella los controles para la elección de los valores involucrados en el contraste, la función de densidad normal junto con la región de rechazo en función del tipo de contraste, los p.valores asociados a cada una de las posibles hipótesis alternativas y la decisión a tomar en función del valor de α elegido.



14.18 Actividad 18

Realizar una plantilla para representar la distribución F de Snedecor.

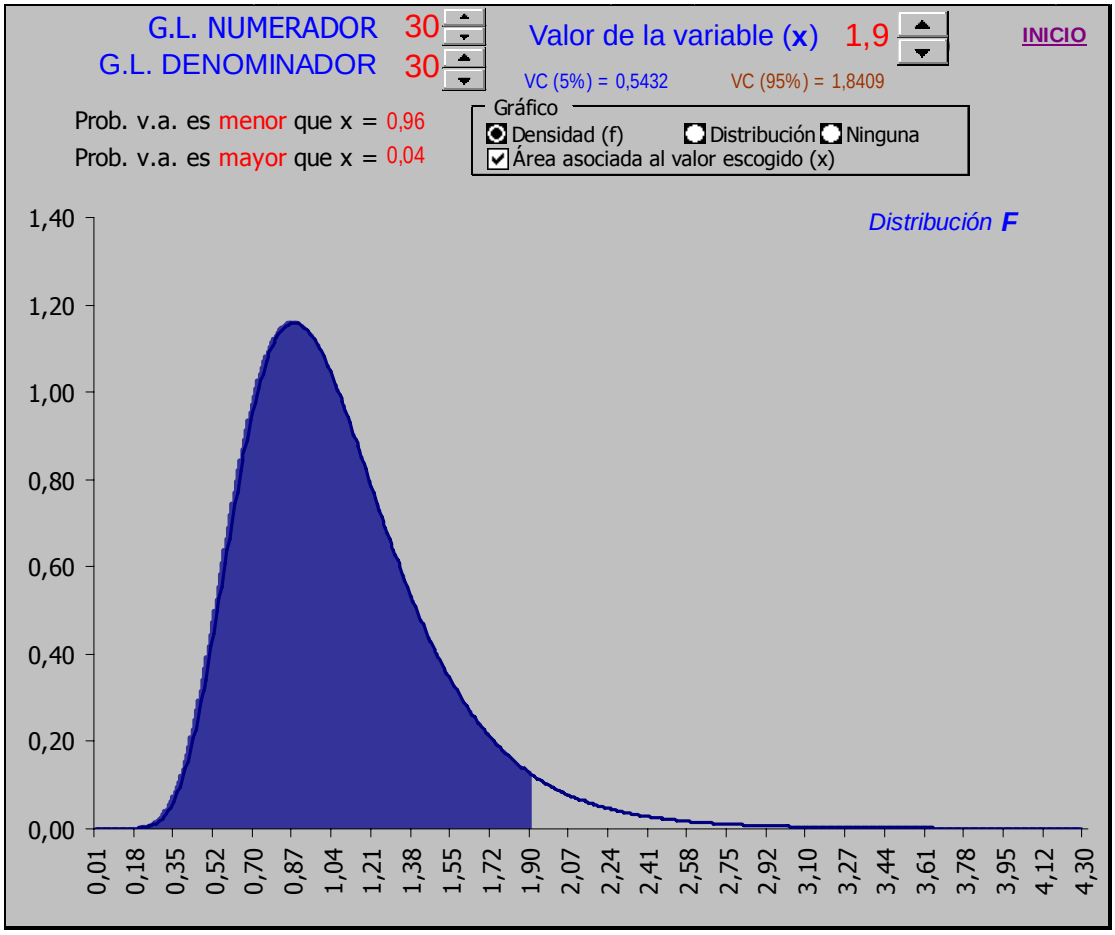
Para esto deberemos conocer la función que Excel incluye en la librería estadística que permite el cálculo de las probabilidades asociadas a la distribución **F**, éstas son:

DISTR.F (x, gl1, gl2,)

DISTR.F.INV (p, gl1, gl2)

Que devuelven, respectivamente, la probabilidad asociada a una valor de **x** distribuido según una $\mathbf{F}_{(gl1,gl2)}$ y el valor crítico de la distribución, es decir, el valor de **x** tal que la probabilidad obtenida coincida con la probabilidad **p** pedida.

De nuevo, si deseamos mejorar la presentación con el gráfico de la distribución deberemos construir una tabla que volcaremos al gráfico correspondiente, el aspecto de la hoja una vez construida podría ser como el siguiente:



14.19 Actividad 19

Realizar una plantilla para representar la distribución χ2.

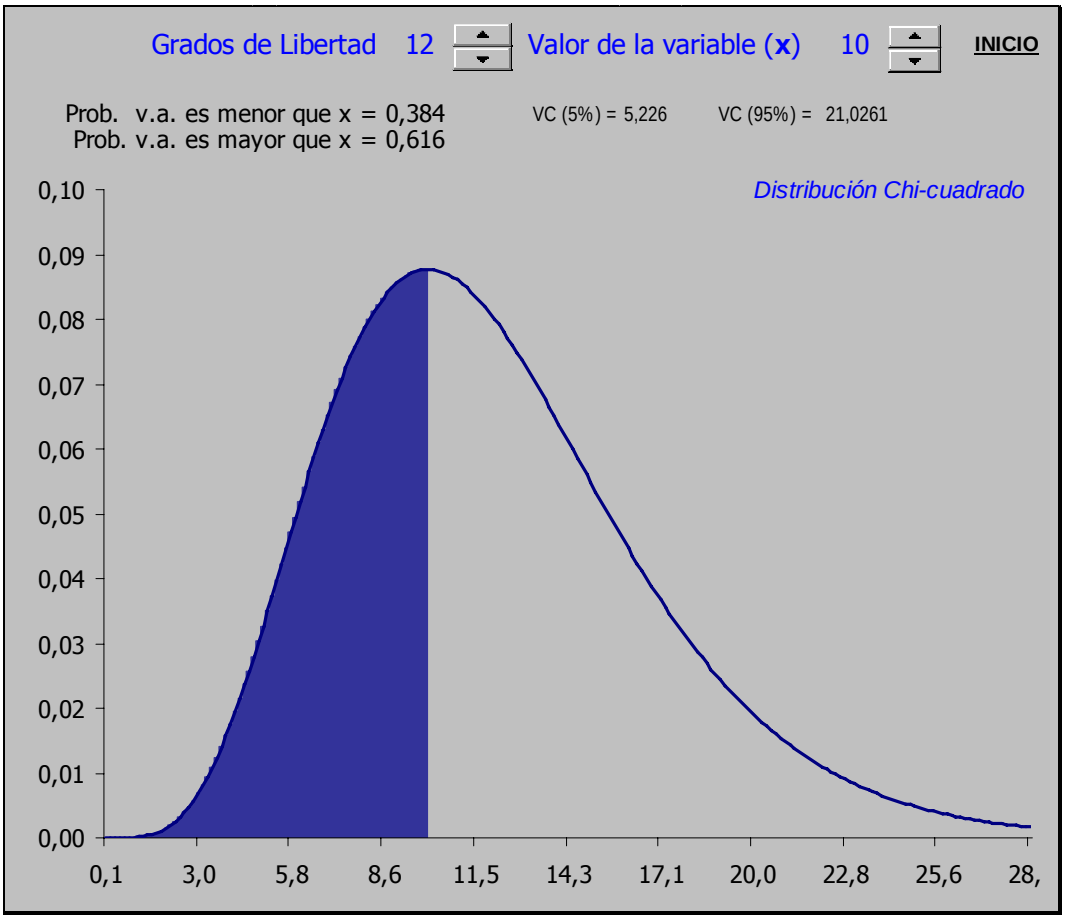
Como en la actividad anterior, para esto deberemos conocer la función que Excel incluye en la librería estadística que permite el cálculo de las probabilidades asociadas a la distribución χ^2, éstas son:

DISTR.CHI (x, gl,)

DISTR.CHI.INV (p, gl)

Que devuelven, respectivamente, la probabilidad asociada a una valor de **x** distribuido según una χ^2_{gl} y el valor crítico de la distribución, es decir, el valor de **x** tal que la probabilidad obtenida coincida con la probabilidad **p** pedida.

De nuevo, si deseamos mejorar la presentación con el gráfico de la distribución deberemos construir una tabla que volcaremos al gráfico correspondiente, el aspecto de la hoja una vez construida podría ser como el siguiente:



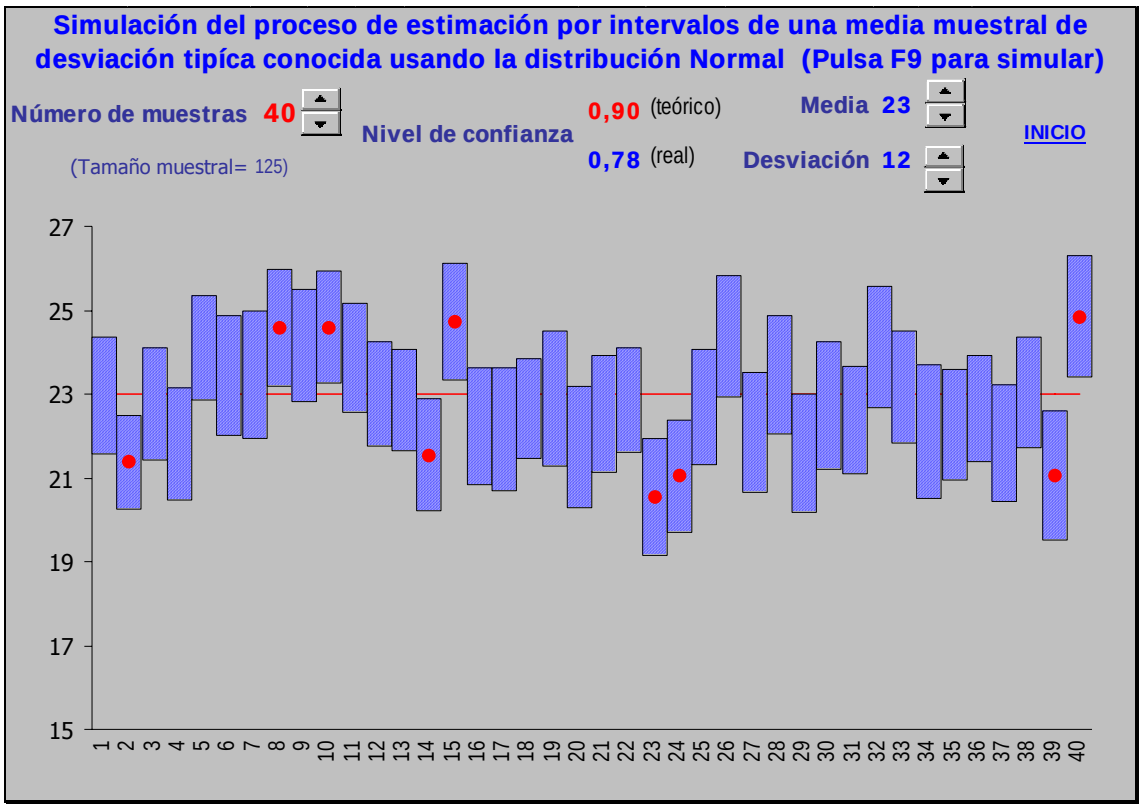
14.20 Actividad 20

Simular el proceso de extracción de una muestra y comprobar empíricamente la distribución muestral de la media.

Para llevar a cabo a actividad propuesta son necesarios conocimientos de la hoja de cálculo que superan a los exigibles a los estudiantes de la asignatura.

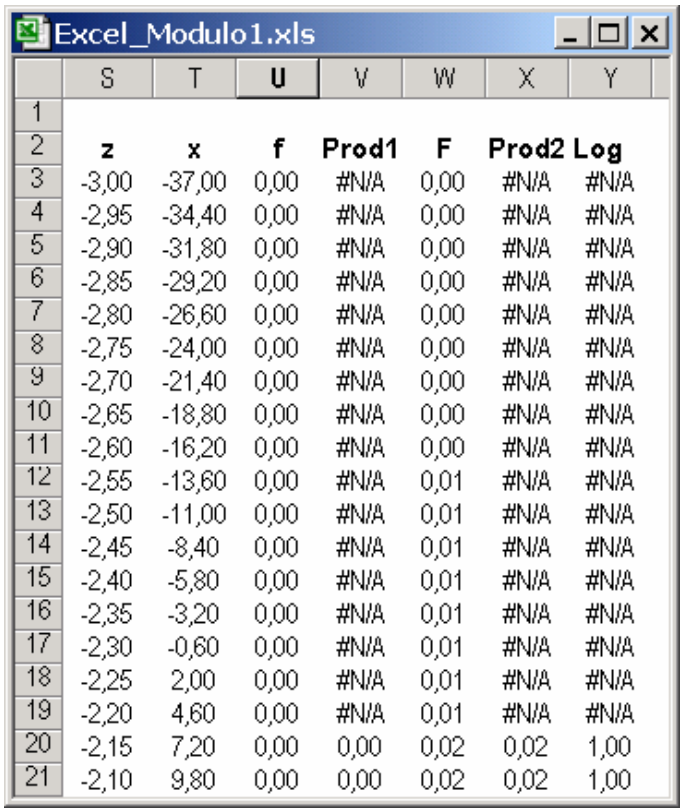
El libro ***ExMod2a.xls*** contiene una única hoja en la que se lleva a cabo el muestreo de una distribución normal y cuyo aspecto es el que aparece en esta página.

Los estudiantes interesados en su desarrollo (que no está directamente relacionado con el objetivo docente actual) pueden solicitar una explicación detallada de como se ha realizado dicha hoja al consultor de la asignatura.



14.21 Anexo :1 Gráficos en la hoja de la actividad 2

Ocultas, gracias a la elección de un color de tinta "blanco", las columnas S, T,..,Y de la hoja de cálculo presentarían - de ser visibles - el siguiente aspecto:



Excel_Modulo1.xls

	S	T	U	V	W	X	Y
1							
2	**z**	**x**	**f**	**Prod1**	**F**	**Prod2**	**Log**
3	-3,00	-37,00	0,00	#N/A	0,00	#N/A	#N/A
4	-2,95	-34,40	0,00	#N/A	0,00	#N/A	#N/A
5	-2,90	-31,80	0,00	#N/A	0,00	#N/A	#N/A
6	-2,85	-29,20	0,00	#N/A	0,00	#N/A	#N/A
7	-2,80	-26,60	0,00	#N/A	0,00	#N/A	#N/A
8	-2,75	-24,00	0,00	#N/A	0,00	#N/A	#N/A
9	-2,70	-21,40	0,00	#N/A	0,00	#N/A	#N/A
10	-2,65	-18,80	0,00	#N/A	0,00	#N/A	#N/A
11	-2,60	-16,20	0,00	#N/A	0,00	#N/A	#N/A
12	-2,55	-13,60	0,00	#N/A	0,01	#N/A	#N/A
13	-2,50	-11,00	0,00	#N/A	0,01	#N/A	#N/A
14	-2,45	-8,40	0,00	#N/A	0,01	#N/A	#N/A
15	-2,40	-5,80	0,00	#N/A	0,01	#N/A	#N/A
16	-2,35	-3,20	0,00	#N/A	0,01	#N/A	#N/A
17	-2,30	-0,60	0,00	#N/A	0,01	#N/A	#N/A
18	-2,25	2,00	0,00	#N/A	0,01	#N/A	#N/A
19	-2,20	4,60	0,00	#N/A	0,01	#N/A	#N/A
20	-2,15	7,20	0,00	0,00	0,02	0,02	1,00
21	-2,10	9,80	0,00	0,00	0,02	0,02	1,00

Extendiéndose desde la fila 3 hasta la 123 (se ha presentado sólo un fragmento en la descripción anterior), aparecen 7 columnas cuyas fórmulas son las siguientes:

Columna de la hoja de cálculo	Valor obtenido
S	El argumento x de la función DISTR.NORM que forzaremos a valores fijos desde –3 hasta 3 creando así el rango suficiente de variación que comentábamos anteriormente. Llamaremos, siguiendo la convención habitual en estadística z a estos valores.
T	La expansión de los valores anteriores en el rango de la distribución normal que variarán así desde menos tres desviaciones típicas a la izquierda de la media, hasta tres desviaciones típicas a la derecha de la media. Valores que denominaremos x y que calcularemos mediante la fórmula "x = media + (z* desv_estándar)"
U	La función de densidad (f) del valor anterior que obtenemos mediante la fórmula: DISTR.NORM(x; media ;desv_estándar ;FALSO)
V	El producto del valor anterior (f) por la variable Log, calculada en la última columna. Este cálculo se hace con la intención de que el valor f calculado anteriormente se anule cuando x no esté entre los límites señalados por el usuario.
W	La función de distribución (F) de la variable aleatoria que obtenemos mediante la fórmula DISTR.NORM(x; media ;desv_estándar ;VERDADERO)
X	El producto del valor anterior (F) por la variable Log, calculada en la última columna. Este cálculo se hace con la intención de que el valor F calculado anteriormente se anule cuando x no esté entre los límites señalados por el usuario.
Y	Una variable lógica (Log) que vale 0 o 1 si x se encuentra o no entre los límites para los que se pide calcular la probabilidad.

Una vez calculadas estas columnas bastará con utilizar los gráficos que Excel pone a nuestra disposición.