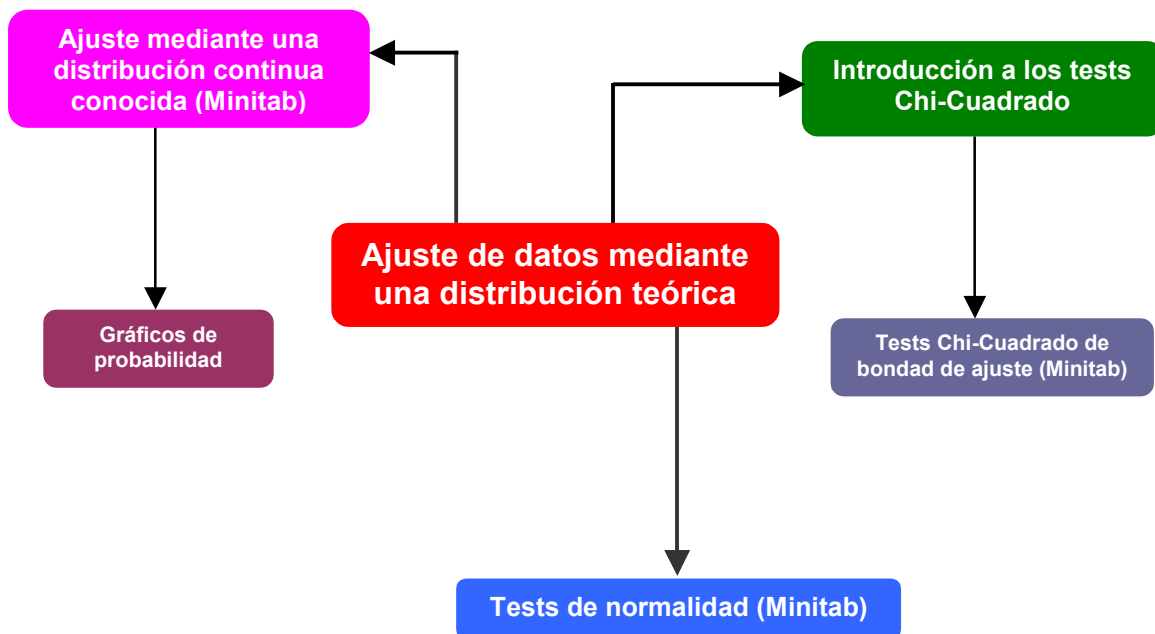


AJUSTE DE DATOS POR UNA DISTRIBUCIÓN TEÓRICA CON MINITAB

Autor: Ángel A. Juan (ajuarp@uoc.edu).

ESQUEMA DE CONTENIDOS



INTRODUCCIÓN

En infinidad de ocasiones nos encontraremos con una serie de datos u observaciones que hemos obtenido al analizar una variable aleatoria de patrón desconocido. Esto ocurrirá, por ejemplo, al registrar los tiempos transcurridos entre llamadas sucesivas a un *call-center*, al registrar los tiempos de fallo de un determinado dispositivo, al contabilizar el número de páginas web distintas que un internauta visita hasta llegar a una que le proporciona la información deseada, etc.

En tales casos, resulta fundamental intentar identificar un patrón conocido (distribución de probabilidad) que nos ayude a explicar el comportamiento de la variable aleatoria. Es lo que se conoce como ajuste de los datos mediante una distribución teórica conocida. Si se logra ajustar los datos por alguna de estas distribuciones, podremos usar las características de ésta para realizar análisis más profundos (inferencia) sobre la población de la cual proviene la muestra o conjunto de observaciones, o incluso para simular algún fenómeno cuyo comportamiento venga descrito por una o varias variables aleatorias (como los mencionados anteriormente).

OBJETIVOS

- Aprender, con ayuda de Minitab, a ajustar observaciones procedentes de una v.a. continua mediante alguna distribución teórica conocida..
- Entender los experimentos multinomiales y los tests Chi-Cuadrado para contrastar la bondad del ajuste.
- Ser capaz de analizar, con ayuda de Minitab, la posible normalidad de un conjunto de datos.

CONOCIMIENTOS PREVIOS

Este math-block supone que el lector está familiarizado con el software estadístico Minitab, así como con conceptos básicos de estadística descriptiva e inferencial (distribuciones de probabilidad, contraste de hipótesis, etc.).

CONCEPTOS FUNDAMENTALES Y CASOS PRÁCTICOS CON SOFTWARE

□ Introducción al ajuste de datos mediante una distribución teórica

Cuando se dispone de un conjunto de observaciones, pertenecientes a una determinada variable aleatoria T de distribución desconocida, lo primero que conviene hacer es tratar de identificar alguna distribución teórica por la cual se puedan ajustar bien dichas observaciones. En otras palabras, se trataría de comprobar si dichas observaciones se distribuyen según un patrón conocido (según una normal, una binomial, etc.), pues ello nos simplificaría el análisis descriptivo de los datos, así como la realización de inferencias sobre la población.

En muchas ocasiones será posible identificar la distribución que mejor se aproxima a las observaciones mediante el uso de **gráficos de probabilidad**. Este tipo de gráficos muestran la función de distribución (f.d.) linealizada de una distribución teórica junto con una nube de puntos que representan estimaciones (no paramétricas) puntuales de la f.d. de T . Evidentemente, cuanto más se aproxime la nube de puntos a la recta que aparece en el gráfico, tanto mejor será el ajuste.

Si se lograra aproximar la distribución de T mediante alguna distribución teórica conocida, sería posible usar esta última para representar gráficamente estimaciones de la función de distribución y/o de la función de densidad (f.d.p.) asociada a las observaciones. En tales casos, se habla de **descripción paramétrica** de la variable T (porque hemos logrado identificar la distribución –y los parámetros asociados- que describen correctamente el comportamiento de la variable aleatoria analizada).

En este capítulo se hará uso del programa estadístico **Minitab** para identificar y describir gráficamente la distribución que mejor se ajuste a un conjunto de observaciones que usaremos como ejemplo. Las posibles distribuciones de ajuste son: la normal, la log-normal (base e), la Weibull, la de valores extremos, la exponencial, la logística y la log-logística.

□ Gráficos de probabilidad

Al representar gráficamente las funciones de distribución (f.d.) de las diferentes distribuciones teóricas, se obtienen curvas muy similares, la mayoría de las cuales resultan difíciles de ser identificadas a simple vista. Es por ello que se utilizan los **gráficos de probabilidad**, los cuales hacen uso de escalas especiales en los ejes, de manera que al representar la f.d. ésta tenga forma lineal.

El primer paso será pues encontrar la transformación adecuada para T y F(T) de modo que al representar T vs. F(T) se obtenga una función lineal.

Ejemplo (linealización de una Weibull): La f.d. asociada a una distribución Weibull de dos parámetros viene dada por la expresión:

$$F(t) = 1 - \exp\{-(t/\alpha)^\beta\} \quad \text{con} \quad \alpha, \beta > 0$$

Esta función puede ser linealizada (i.e., puesta de la forma: $y = a + bx$) como sigue:

$$\begin{aligned} F(t) = 1 - \exp\{-(t/\alpha)^\beta\} &\Rightarrow \ln(1-F(t)) = \ln(\exp\{-(t/\alpha)^\beta\}) \Rightarrow \ln(1-F(t)) = -(t/\alpha)^\beta \Rightarrow \\ &\Rightarrow \ln(-\ln(1-F(t))) = \beta \cdot \ln(t/\alpha) \Rightarrow \ln(\ln(1-F(t))^{-1}) = \beta \cdot \ln(t) - \beta \cdot \ln(\alpha) \end{aligned}$$

Tomando ahora:

$$y = \ln(\ln(1-F(t))^{-1}) \quad \text{y} \quad x = \ln(t)$$

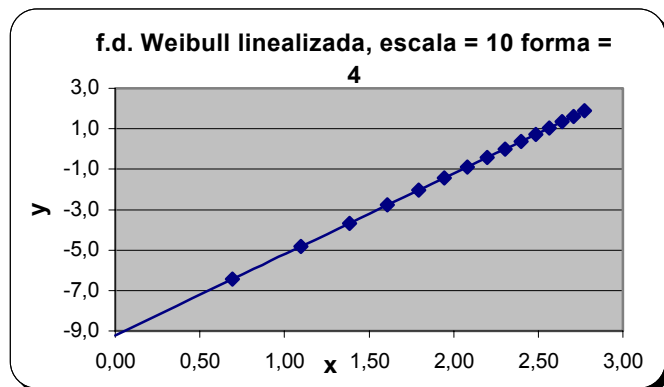
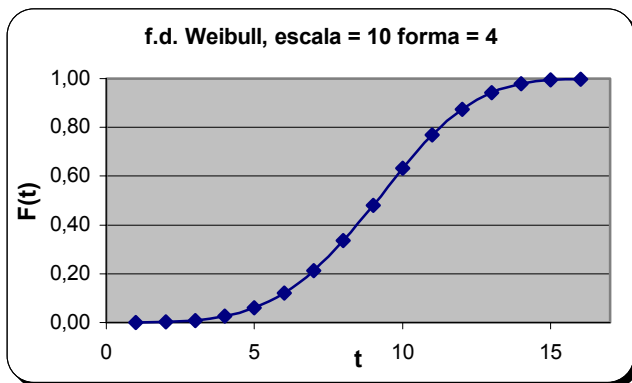
la f.d. puede describirse en forma lineal como:

$$y = \beta \cdot x - \beta \cdot \ln(\alpha)$$

A continuación se representa gráficamente la f.d. de una Weibull (con escala = 10 y forma = 4) y su versión linealizada:

Escala (alpha) = 10 WEIBULL
Forma (beta) = 4

t	F(t)	x = ln(t)	y = ln(ln(1-F(t)) ⁻¹)
1	0,00	0,00	-9,2
2	0,00	0,69	-6,4
3	0,01	1,10	-4,8
4	0,03	1,39	-3,7
5	0,06	1,61	-2,8
6	0,12	1,79	-2,0
7	0,21	1,95	-1,4
8	0,34	2,08	-0,9
9	0,48	2,20	-0,4
10	0,63	2,30	0,0
11	0,77	2,40	0,4
12	0,87	2,48	0,7
13	0,94	2,56	1,0
14	0,98	2,64	1,3
15	0,99	2,71	1,6
16	1,00	2,77	1,9



Una vez conocidas las transformaciones que permiten linealizar la f.d. asociada a una distribución, es posible construir una plantilla especial (con los ejes graduados de forma adecuada) sobre la cual representar una nube de puntos que contenga cada uno de los tiempos de fallo observados (eje x) junto con el valor (estimado) de la f.d. asociado a dicha observación (eje y).

Para cada punto (x_j, y_j) , el valor x_j vendrá dado por la j -ésima observación t_j (instante en que se ha producido el fallo j -ésimo). Más complicado será hallar el valor de la coordenada y_j , la cual representará el valor estimado de $F(t_j)$. Es usual estimar dicho valor mediante los llamados **rangos medianos**, los cuales se pueden calcular, en el caso de la distribución Weibull con observaciones completas (sin censura), mediante la ecuación que se muestra a continuación.

$$F(t_j) \approx \text{rango mediano } j\text{-ésimo} = (1 + F_{(0,5; m, n)} \cdot (n - j + 1) / j)^{-1}$$

donde:

$F_{(0,5; m, n)}$ es la mediana de una F-Snedecor con $m = 2(n - j + 1)$ y $n = 2j$ grados de libertad, j es el orden del fallo, y n es el tamaño muestral.

Como se verá en el apartado siguiente, los programas estadísticos actuales (como **Minitab**) son capaces de realizar los cálculos anteriores, automatizando así el proceso de construcción de estos gráficos de probabilidad.

Cuando se tengan ya representados todos los puntos (x, y) asociados a las observaciones, se deberá hallar la recta de regresión asociada, la cual corresponderá a la f.d. de la distribución elegida cuyos parámetros mejor se ajusten a las observaciones. Para ver si las observaciones pueden aproximarse bien por dicha distribución, habrá que analizar (gráficamente o mediante el estadístico Anderson-Darling) si los puntos representados se encuentran suficientemente próximos a la recta, prestando especial atención a los valores de los extremos.

□ **Ajuste de datos pertenecientes a una v.a. continua con Minitab**

Como hemos comentado, las últimas versiones de **Minitab** incorporan una serie de opciones que permiten intentar ajustar un conjunto de observaciones mediante algunas de las principales distribuciones continuas. Si el proceso tiene éxito (i.e.: si logramos ajustar razonablemente bien alguna de las distribuciones teóricas a los datos), podremos suponer que los datos siguen un patrón caracterizado por una distribución continua conocida, lo cual nos facilitará la obtención de información adicional sobre el comportamiento de la variable aleatoria.

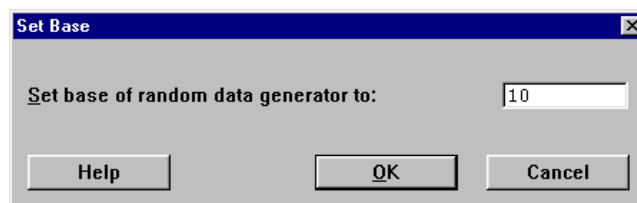
Ejemplo: vamos a usar Minitab para realizar un experimento consistente en tres fases:

Fase 1: generaremos un conjunto de 50 números aleatorios procedentes de una distribución exponencial con parámetro $\lambda = 2$ (media = $1/\lambda = 0,5$).

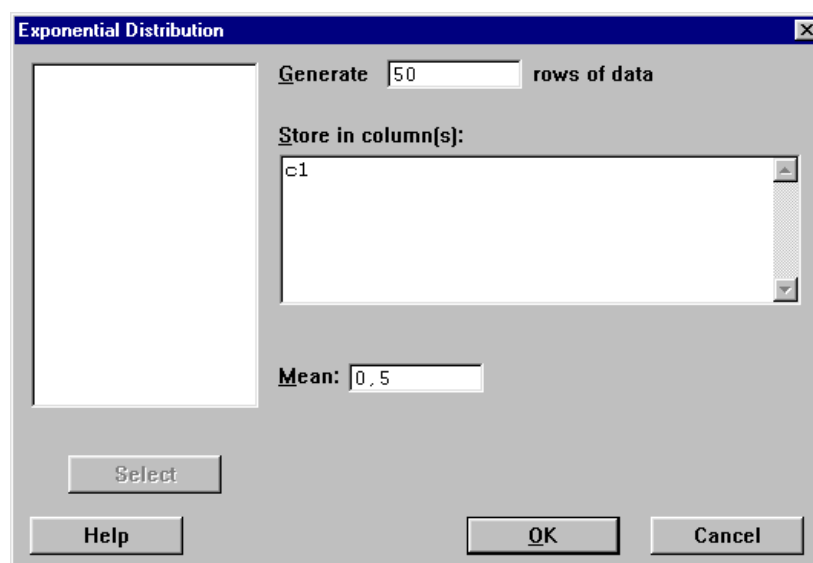
Fase 2: completada la fase anterior, consideraremos los valores obtenidos como observaciones dadas (olvidando que proceden de una distribución exponencial), e intentaremos buscar, con ayuda de Minitab, aquella distribución concreta (con parámetros concretos) que mejor se ajuste a dichas observaciones.

Fase 3: compararemos la f.d. de la distribución verdadera –la exponencial de parámetro λ – con la que hemos obtenido en la fase 2 suponiendo desconocida la procedencia de los datos.

Fase 1: Generamos las 50 observaciones (números aleatorios provenientes de una $\text{exp}(2)$). A fin de obtener los mismos valores aleatorios, podemos utilizar una “semilla” inicial común. En este caso, usaremos el valor 10 como “semilla”. Para indicarle a Minitab tal elección, usaremos la opción **Calc > Set Base**:



Ahora, pasamos ya a la generación de las 50 observaciones mediante la opción **Calc > Random Data > Exponential**:

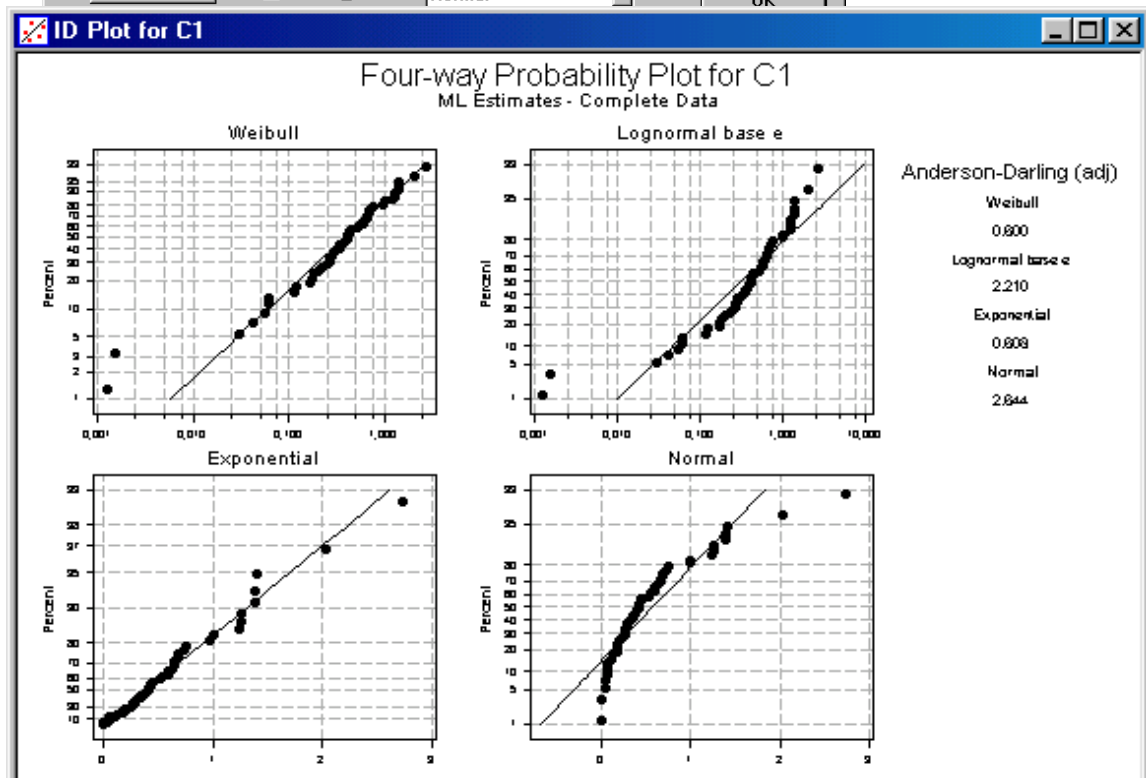
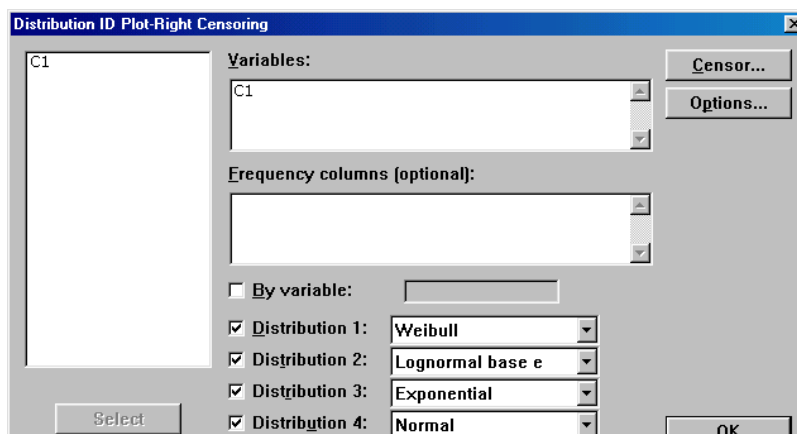


El resultado será algo similar al siguiente:

Worksheet 1 ***	
	C1
↓	
1	0,03030
2	0,28989
3	0,44377
4	0,42514
5	2,73026
6	0,73361
7	0,38022
8	0,39646

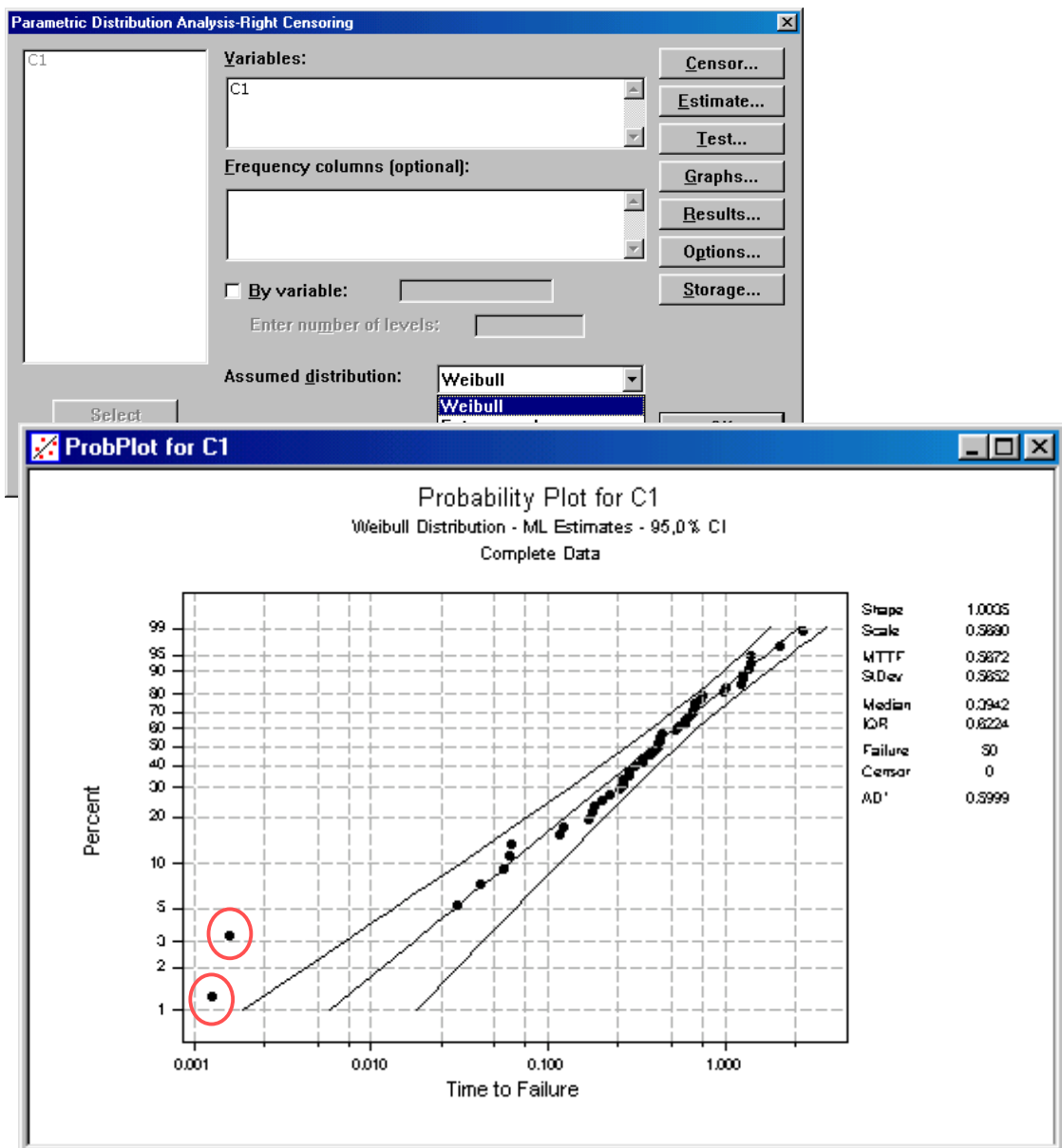
A partir de ahora, consideraremos estos valores como observaciones obtenidas como resultado de una medición (es decir, olvidaremos por un momento que conocemos la distribución de la cual provienen).

Fase 2: Supongamos pues que disponemos de una serie de datos cuya procedencia nos es desconocida, y que deseamos encontrar –o, al menos intentarlo– una distribución continua que nos sirva para explicar el comportamiento de la v.a. de la cual proceden. Para ello, usaremos como primera aproximación la opción Stat > Reliability/Survival > Distribution ID Plot....:



Como se aprecia en las gráficas anteriores, en esta primera aproximación comienza a quedar claro que, de las cuatro distribuciones usadas (normal, log-normal, exponencial y Weibull), las dos que mejor se ajustan a las observaciones son la exponencial y la Weibull –los puntos se sitúan muy cerca de la línea y el comportamiento de los mismos no sigue un patrón curvilíneo como en el caso de la normal y de la log-normal. Además de las gráficas, el “output” anterior también nos proporciona el **estadístico de Anderson-Darling** ajustado, el cual es un reflejo de cuán lejos se encuentran los puntos respecto de la recta. Por tanto, cuanto menor sea el valor de dicho estadístico, tanto mejor será la bondad del ajuste. De los valores de dicho estadístico, se desprende nuevamente que la Weibull (AD = 0,600) y la exponencial (AD = 0,608) proporcionan un mejor ajuste a las observaciones.

Ahora podemos usar la opción **Stat > Reliability/Survival > Parametric Distribution Analysis...** para afinar algo más en nuestra elección. Como se observa en la siguiente imagen, es posible optar entre un amplio ramillete de distribuciones candidatas. En nuestro caso, optaremos por una Weibull y, posteriormente, repetiremos el proceso con una exponencial:



Distribution Analysis: C1

Variable: C1

Censoring Information	Count
Uncensored value	50

 Estimation Method: Maximum Likelihood
 Distribution: **Weibull**
Parameter Estimates

Parameter	Estimate	Standard Error	95,0% Normal CI	
			Lower	Upper
Shape	1,0035	0,1124	0,8057	1,2499
Scale	0,56802	0,08382	0,42536	0,75853

Log-Likelihood = -21,651

 Goodness-of-Fit
 Anderson-Darling (adjusted) = **0,5999**
Distribution Analysis: C1

Variable: C1

Censoring Information	Count
Uncensored value	50

 Estimation Method: Maximum Likelihood
 Distribution: **Exponential**
Parameter Estimates

Parameter	Estimate	Standard Error	95,0% Normal CI	
			Lower	Upper
Shape	1,00000			
Scale	0,56724	0,08022	0,42992	0,74842

Log-Likelihood = -21,652

 Goodness-of-Fit
 Anderson-Darling (adjusted) = **0,6076**

Como se puede apreciar por los "outputs" anteriores, y dada la gran similitud entre ambos estadísticos AD (0,599 para el ajuste por la Weibull y 0,6076 para el ajuste por la exponencial), las observaciones se podrían ajustar bastante bien tanto por una Weibull con parámetros forma = 1,0035 y escala = 0,56802 como por una exponencial de media = 0,56724. Esto no es de extrañar, ya que la exponencial no es más que una Weibull con parámetro de forma = 1 y parámetro de escala igual a la media.

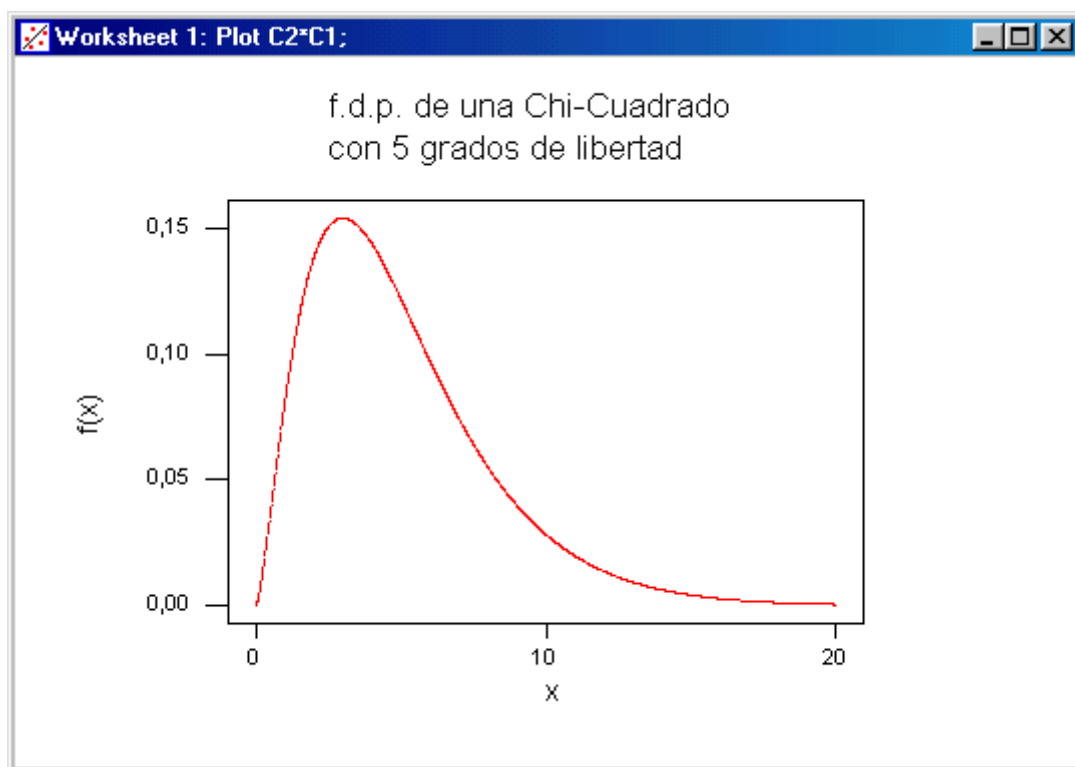
Llegados a este punto, es importante percatarse de la precisión con que hemos sido capaces de ajustar los datos: las observaciones procedían de una exponencial con media 0,5. Pues bien, suponiendo desconocida esta información y partiendo de tan sólo 50 observaciones, hemos logrado casi adivinar el verdadero modelo subyacente a los datos (lógicamente, es de esperar que si dispusiésemos de más observaciones, nuestro ajuste podría ser aún mejor).

□ Introducción a los contrastes Chi-Cuadrado (χ^2)

Una **variable categórica** es una variable que clasifica cada individuo de una población en una de las varias clases -mutuamente excluyentes- en que ésta se divide. Hay muchos problemas en los que los datos están clasificados en categorías, mostrándose los resultados mediante **distribuciones de frecuencias**. Un ejemplo clásico sería la distribución de frecuencias de las notas finales de cualquier asignatura (nº de sobresalientes, nº de notables, etc.).

La **distribución Chi-cuadrado** puede usarse para realizar contrastes de hipótesis en diferentes situaciones, siendo los principales contrastes los asociados a un experimento multinomial (que veremos en el siguiente apartado) y los asociados a una tabla de contingencia. Estos dos tipos de tests se usan para comparar resultados observados (O) con resultados esperados (E) a fin de determinar alguna de las siguientes propiedades:

- (1) La bondad del ajuste de las observaciones con respecto al modelo teórico seleccionado para explicar su comportamiento (i.e.: contrastar si las frecuencias observadas son coherentes con las que cabría esperar habida cuenta de la distribución teórica que hayamos elegido para explicar el comportamiento de la variable aleatoria),
- (2) La independencia entre clases (i.e.: si los distintas categorías son o no independientes), y
- (3) La homogeneidad de clases (i.e.: si las distintas categorías presentan un comportamiento homogéneo o si, por el contrario, hay claras diferencias entre ellas respecto a algún factor de interés).



Supongamos que tenemos un número k de clases en las cuales se han ido registrado un total de n observaciones (n será, pues, el tamaño muestral). Denotaremos las **frecuencias observadas** en cada clase por O_1, O_2, \dots, O_k . Se cumplirá:

$$O_1 + O_2 + \dots + O_k = n$$

Lo que queremos es comparar las frecuencias observadas con las **frecuencias esperadas** (teóricas), a las que denotaremos por E_1, E_2, \dots, E_k . Se verificará que:

$$E_1 + E_2 + \dots + E_k = n$$

	FRECUENCIA OBSERVADA	FRECUENCIA ESPERADA
CLASE 1	O_1	E_1
CLASE 2	O_2	E_2
...
CLASE K	O_K	E_K
Total	N	N

Llegados a este punto, el problema es determinar si las frecuencias observadas están o no en concordancia con las frecuencias esperadas (es decir, si el número de resultados observados en cada clase corresponde aproximadamente al número esperado). Para comprobarlo, haremos uso de un contraste de hipótesis usando la distribución Chi-cuadrado:

El estadístico de contraste será $\chi^{2*} = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i}$

Observar que este valor será la suma de k números no negativos. El numerador de cada término es la diferencia entre la frecuencia observada y la frecuencia esperada. Por tanto, cuanto más cerca estén entre sí ambos valores más pequeño será el numerador, y viceversa. El denominador permite relativizar el tamaño del numerador.

Las ideas anteriores sugieren que, cuanto menor sean el valor del estadístico χ^{2*} , más coherentes serán las observaciones obtenidas con los valores esperados. Por el contrario, valores grandes de este estadístico indicarán falta de concordancia entre las observaciones y lo esperado. En este tipo de contraste se suele rechazar la hipótesis nula (los valores observados son coherentes con los esperados) cuando el estadístico es mayor que un determinado valor crítico.

Notas:

1. El valor del estadístico χ^{2*} se podrá aproximar por una distribución Chi-cuadrado cuando el tamaño muestral n sea grande (normalmente es suficiente con $n > 30$), y todas las frecuencias esperadas sean iguales o mayores a 5 (en ocasiones deberemos agrupar varias categorías a fin de que se cumpla este requisito).
2. Se supone que las observaciones son obtenidas mediante muestreo aleatorio a partir de una población que previamente ha sido dividida en categorías.

□ Test χ^2 de bondad de ajuste

Un **experimento multinomial** es la generalización de un experimento binomial:

1. Consiste en n pruebas idénticas e independientes.
2. Para cada prueba, hay un número k de resultados posibles.
3. Cada uno de los k posibles resultados tiene una probabilidad de ocurrencia p_i asociada ($p_1 + p_2 + \dots + p_k = 1$), la cual permanece constante durante el desarrollo del experimento.
4. El experimento dará lugar a un conjunto de frecuencias observadas (O_1, O_2, \dots, O_k) para cada resultado. Obviamente, $O_1 + O_2 + \dots + O_k = n$.

En ocasiones estaremos interesados en comparar los resultados obtenidos al realizar un experimento multinomial con los resultados esperados (teóricos). Ello nos permitirá saber si nuestro modelo teórico se ajusta bien o no a las observaciones. Para ello, recurriremos a la distribución Chi-cuadrado, la cual nos permitirá realizar un **contraste sobre la bondad del ajuste**.

Concretamente, usaremos el estadístico $\chi^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i}$ con $k - 1$ grados de libertad.

Podemos calcular cada frecuencia esperada (teórica) multiplicando el número total de pruebas n por la probabilidad de ocurrencia asociada, es decir:

$$E_i = n \cdot p_i \quad i = 1, \dots, k$$

Ejemplo (test χ^2 bondad ajuste): Un artículo aparecido en El Mundo describía la experiencia llevada a cabo por un estudiante de Bachillerato, el cual había lanzado en 3.590 ocasiones cinco monedas iguales al aire (lo que hace un total de 17.950 lanzamientos), obteniendo 464 más caras que cruces.

¿Es este resultado estadísticamente significativo? ¿Podemos concluir que las monedas no eran simétricas?

En la siguiente tabla se muestran los resultados registrados por el estudiante:

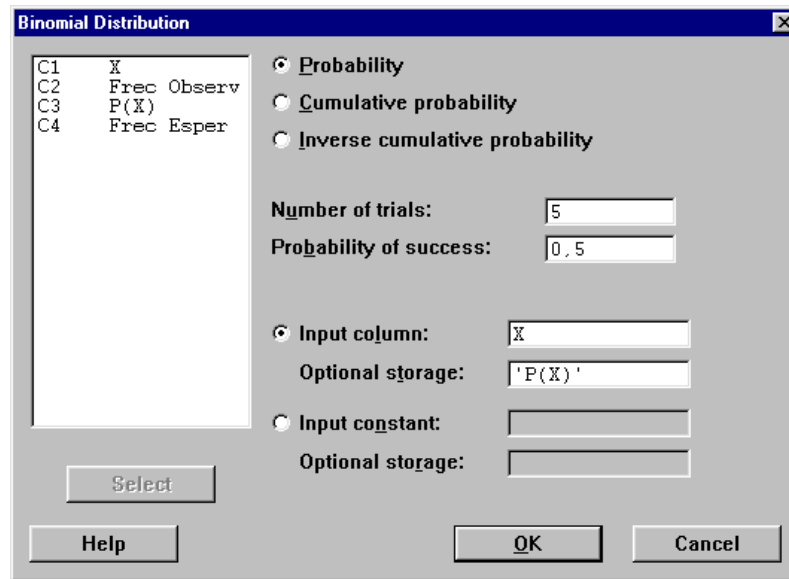
Número de caras en los cinco lanzamientos	Frecuencia Observada
0	100
1	524
2	1.080
3	1.126
4	655
5	105
<i>Total de lanzamientos de 5 monedas</i>	<i>3.590</i>

Si las monedas fuesen simétricas, la probabilidad de obtener cara en cada lanzamiento sería de 0,5. Por tanto, al lanzar cinco monedas simultáneamente, el número de caras obtenidas, que denotaremos por X , seguiría una distribución binomial con $n = 5$ y $p = 0,5$. Esta será nuestra hipótesis nula:

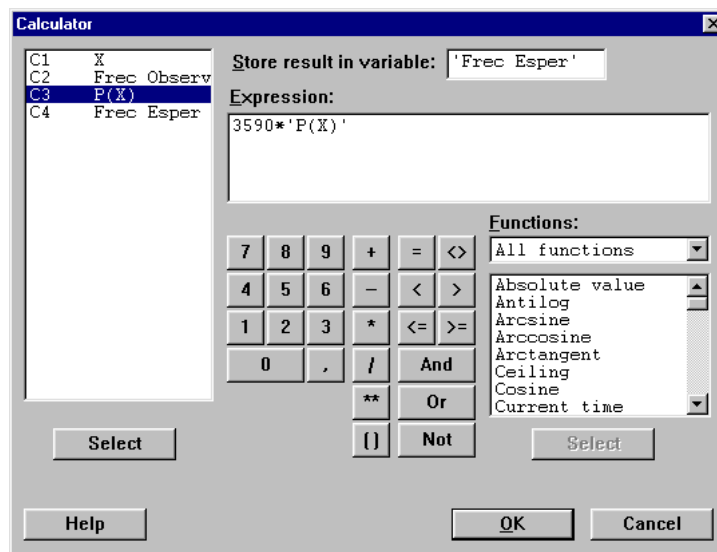
$$H_0 : X \text{ sigue una distribución } B(5;0.5)$$

Bajo la hipótesis anterior, podemos calcular el valor esperado para la probabilidad de obtener 0 caras, 1 cara, 2 caras, etc.:

Calc > Probability Distributions > Binomial :



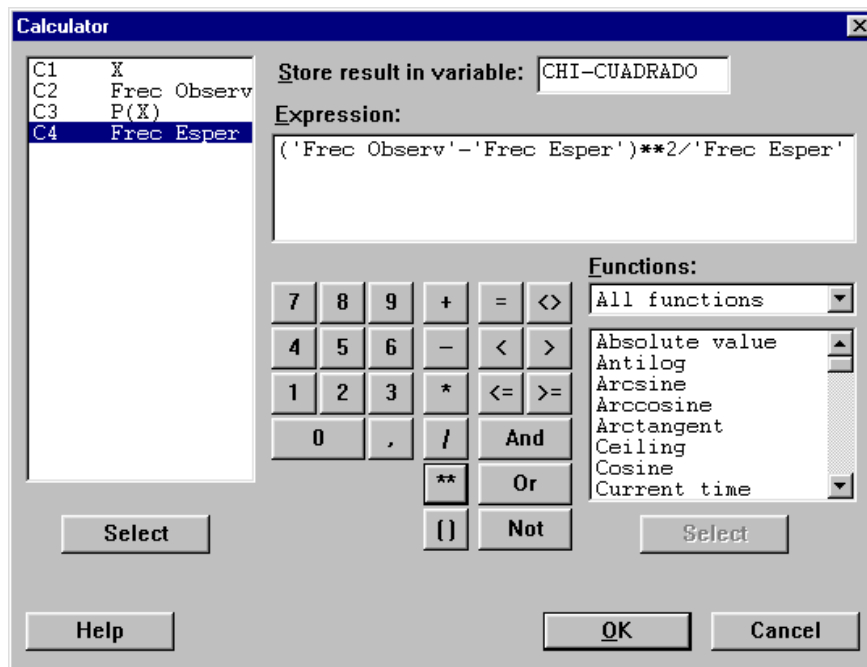
Calc > Calculator:



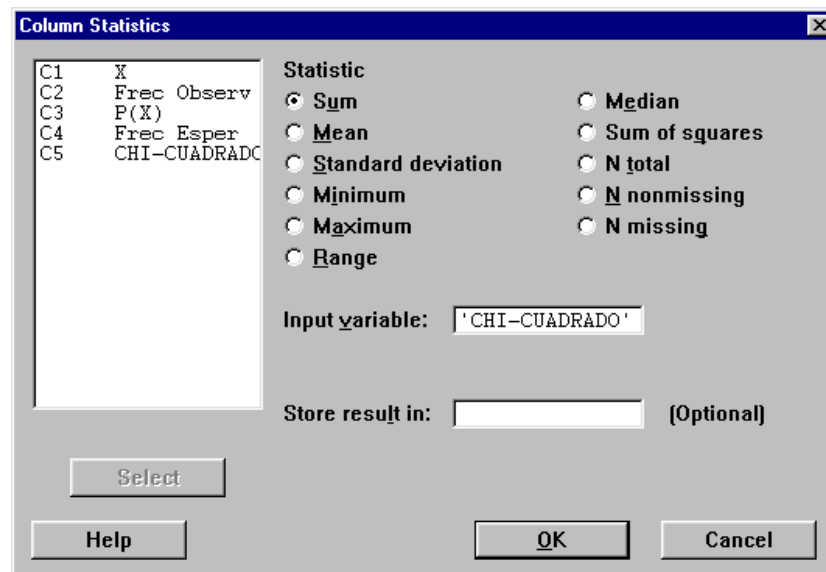
	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
	X	Frec Observ	P(X)	Frec Esper						
1	0	100	0,03125	112,19						
2	1	524	0,15625	560,94						
3	2	1080	0,31250	1121,88						
4	3	1126	0,31250	1121,88						
5	4	655	0,15625	560,94						
6	5	105	0,03125	112,19						
7										
8										

Los valores observados y los esperados no parecen coincidir. Observar que, incluso en el caso de que nuestra hipótesis nula fuese cierta, ambos valores no serían exactamente iguales -ya que siempre habrá cierto margen de variación. La dificultad está en determinar si las diferencias entre ambos valores son o no significativas. Calculemos el estadístico χ^{2*} :

Calc > Calculator :



Calc > Column Statistics :

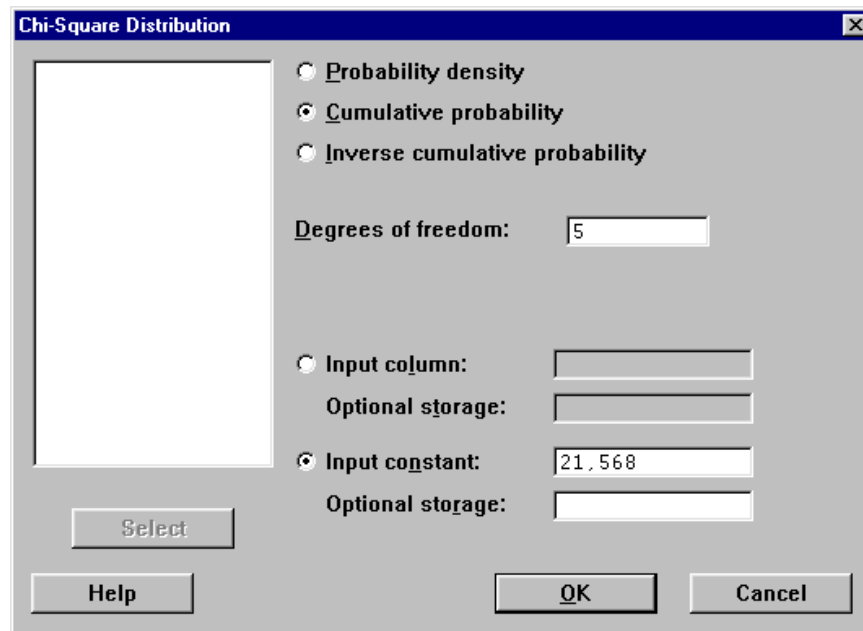


Column Sum

Sum of CHI-CUADRADO = 21,568

Calculamos finalmente el p-valor asociado a este estadístico. En este caso, como trabajamos con un contraste unilateral, $p\text{-valor} = P(\chi^2 > 21,568) = 1 - P(\chi^2 \leq 21,568)$ donde χ^2 sigue una distribución Chi-cuadrado con $k - 1 = 5$ grados de libertad. Por tanto:

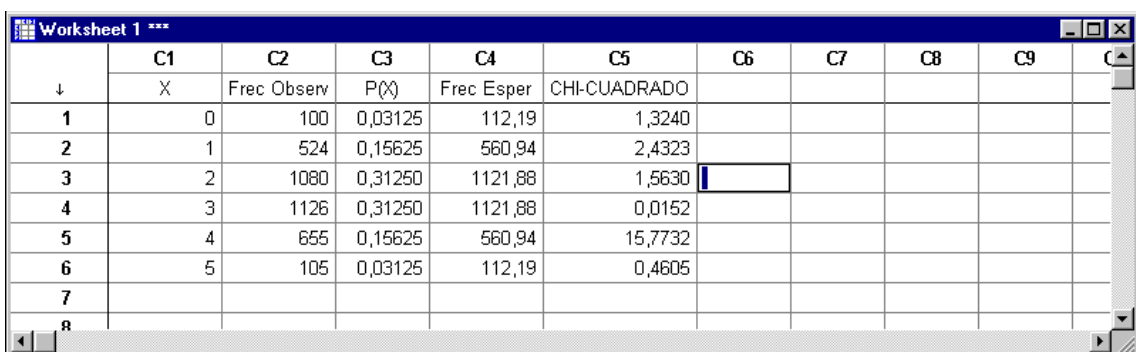
Calc > Probability Distributions > Chi-Square :



Cumulative Distribution Function	
Chi-Square with 5 DF	
x	P(X <= x)
21,5680	0,9994

Así pues, $p\text{-valor} = 1 - 0,9994 = 0,0006 < 0,05$. Por tanto, podemos considerar que el p-valor es significativo (al menos para $\alpha = 0,05$), motivo por el cual rechazaremos la hipótesis nula, i.e.: las monedas no parecen ser simétricas (i.e.: no siguen una distribución binomial con parámetro $p = 0,5$).

Profundicemos un poco más en nuestro análisis: observar que, en la columna CHI-CUADRADO, aparece un valor (enorme) de 15,7732 asociado a la obtención de 4 caras:



	C1	C2	C3	C4	C5	C6	C7	C8	C9
↓	X	Frec Observ	P(X)	Frec Esper	CHI-CUADRADO				
1	0	100	0,03125	112,19	1,3240				
2	1	524	0,15625	560,94	2,4323				
3	2	1080	0,31250	1121,88	1,5630				
4	3	1126	0,31250	1121,88	0,0152				
5	4	655	0,15625	560,94	15,7732				
6	5	105	0,03125	112,19	0,4605				
7									
8									

Este valor es un reflejo de que hay una discrepancia “anormal” entre los valores observados y los esperados para esta categoría. Es posible que haya habido un error en los registros, contabilizándose algunos resultados de 4 cruces como resultados de 4 caras.

□ Test de normalidad con Minitab

Muchas de las variables que nos podemos encontrar en la vida real siguen una distribución normal o aproximadamente normal. Ésta es una de las razones por las cuales dicha distribución es tan importante en estadística. El hecho de que la normal sea una distribución tan frecuente implica que en muchas ocasiones tendremos la sospecha –infundada bien por el tipo de variable con que estemos trabajando, o bien por el análisis visual del histograma de los datos- de que las observaciones obtenidas provienen de una distribución normal. Si, en efecto, los datos siguiesen una distribución normal, ello nos simplificaría notablemente la realización de análisis posteriores y de inferencia sobre la población a partir de la cual hemos obtenido las observaciones (suponiendo que éstas constituyen una muestra aleatoria de la misma).

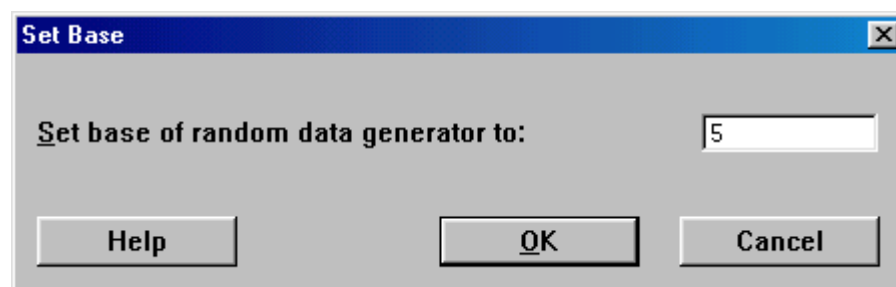
Minitab nos puede ayudar a contrastar si un determinado conjunto de observaciones se comporta o no según una distribución normal. Para ello, el programa proporciona una serie de tests de normalidad que complementan el análisis gráfico de las observaciones.

Ejemplo: En este ejemplo vamos a realizar un experimento consistente en dos fases:

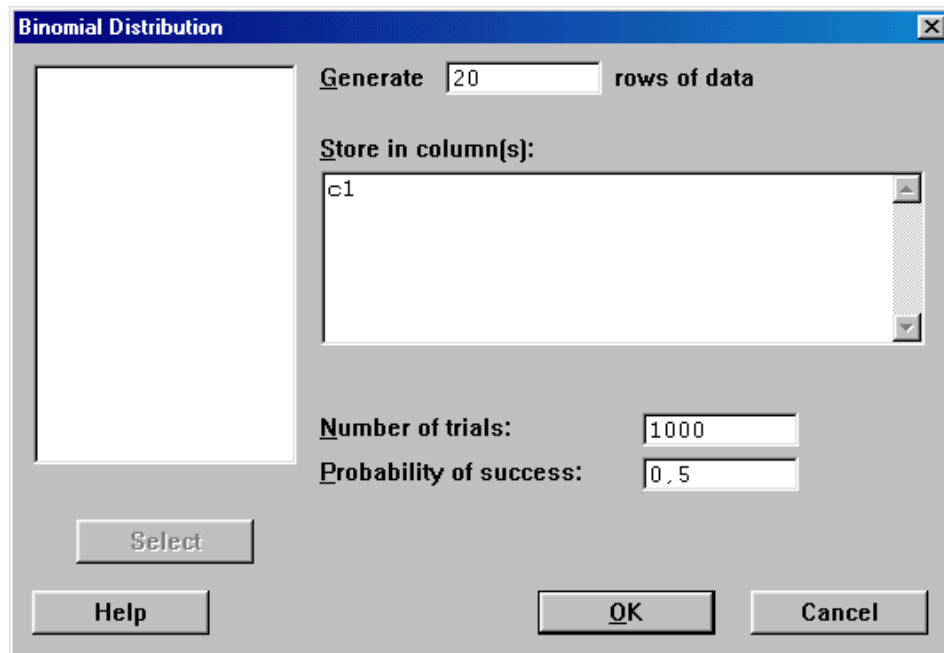
Fase 1: Generaremos 20 observaciones aleatorias procedentes de una distribución binomial con parámetros $n = 1.000$ y $p = 0,5$.

Fase 2: Completada la fase 1, olvidaremos la procedencia de los datos generados, y supondremos que han sido obtenidos al medir alguna variable cuya distribución es desconocida, pero de la cual se sospecha que sigue un comportamiento aproximadamente normal. Nos interesará pues realizar un test para contrastar si la hipótesis anterior es sostenible desde un punto de vista estadístico (i.e.: si, en efecto, tiene sentido suponer que los datos se distribuyen según una normal).

Fase 1: A fin de obtener exactamente los mismos resultados que se muestran a continuación, se recomienda usar una “semilla” que inicialice la generación de números aleatorios. Para ello usaremos la opción `Calc > Set Base...`



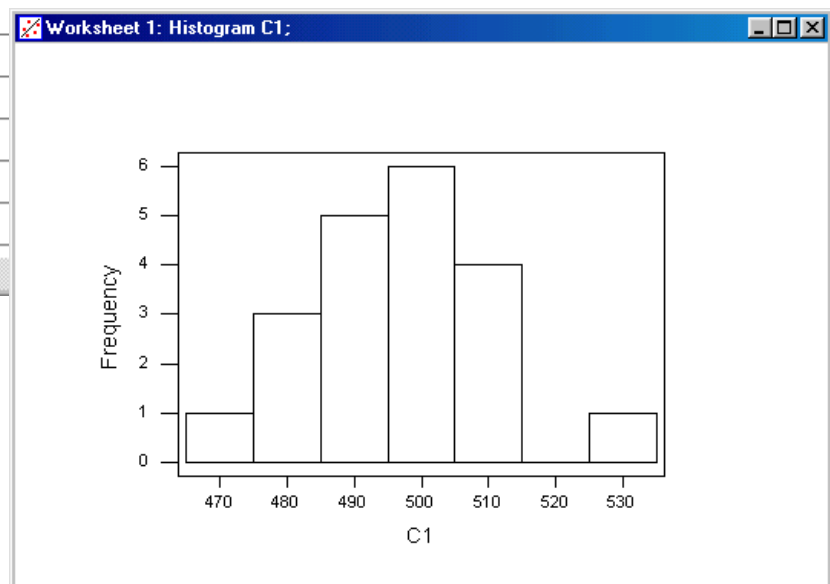
Establecida la “semilla”, podemos generar los números aleatorios usando la opción `Calc > Random Data > Binomial...` :



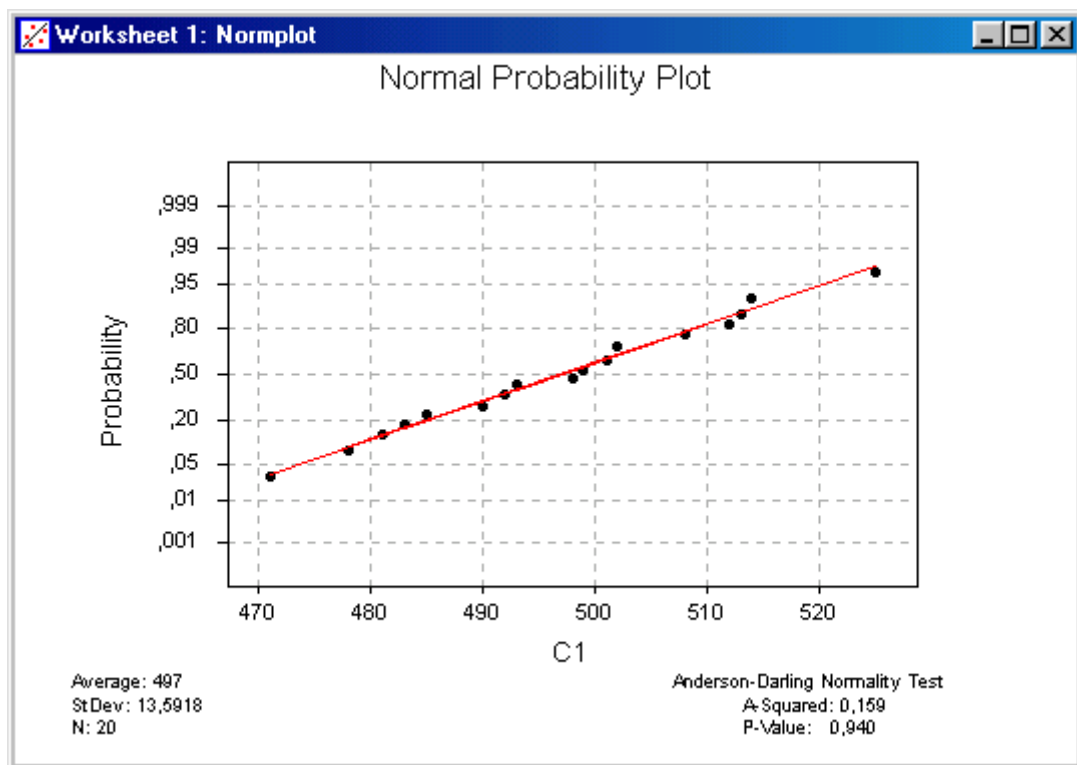
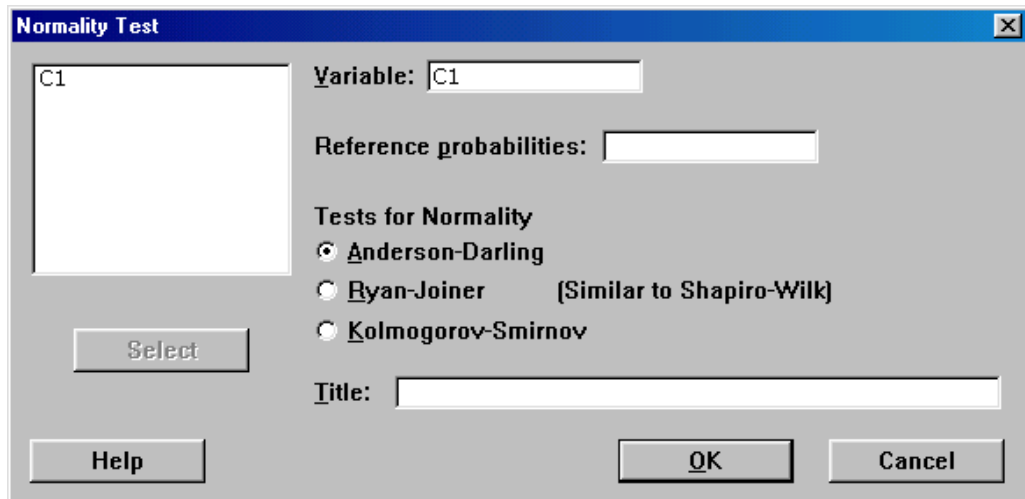
Obtendremos el siguiente listado de números aleatorios:

Worksheet 1 ***	
	C1
↓	
1	490
2	485
3	508
4	493
5	501
6	501
7	525
8	513

Podemos representar los datos anteriores mediante un histograma para comprobar que, en efecto, la forma en que estos se distribuyen nos recuerda a la de la distribución normal:



Fase 2: Ahora, usaremos el test de normalidad incorporado en **Stat > Basic Statistics > Normality Test...** para contrastar la hipótesis nula de que los datos anteriores siguen una distribución normal:



Como se aprecia en el gráfico anterior, los puntos se acercan bastante a la recta –lo cual es un claro indicio de que siguen una distribución aproximadamente normal. Además, el p-valor asociado al contraste de Anderson-Darling es $p\text{-value} = 0,940$ (mucho mayor que 0,05), por lo que estamos muy lejos de rechazar la hipótesis nula de que los datos se distribuyen de forma normal.

Este experimento nos ha permitido, además, comprobar empíricamente un resultado teórico de sumo interés: cuando una variable aleatoria se distribuye según una binomial con parámetros n y p , bajo determinadas condiciones (n suficientemente grande y p cercano a 0,5) es posible aproximar el comportamiento de dicha variable mediante una distribución normal de media $n \cdot p$ y varianza $n \cdot p \cdot (1-p)$.

BIBLIOGRAFÍA

- [1] Ross, S. M. (2001): "Simulation". Academic Press. ISBN 0125980531.
- [2] Matloff, N.S. (1997): "Probability Modeling and Computer Simulation". PWS Publishing Co.
- [3] Yakowitz, R. (1994): "Computational Probability and Simulation". Addison-Wesley Pub. Co.
- [4] Thompson, J.R. (2000): "Simulation: a modeler's approach". John Wiley & Sons. ISBN: 0471251844.
- [5] Rubinstein, R. (1998): "Modern simulation and modeling" John Wiley & Sons. ISBN: 0471170771.

ENLACES

- ❑ <http://www.itl.nist.gov/div898/handbook/index.htm>
Libro on-line "Engineering Statistics Handbook" (ver apartado *goodness-of-fit*)
- ❑ <http://www.palisade.com/html/bestfit.html>
Página web de @Risk.xls dedicada al ajuste de datos"
- ❑ http://isgwww.cs.uni-magdeburg.de/~graham/its_01/lectures/06-Inputmodeling-4.pdf
PDF con diapositivas en las que se explica cómo ajustar observaciones mediante una distribución de probabilidad conocida.
- ❑ <http://www.cse.msu.edu/~cse808/note/lecture9.ppt>
PowerPoint en el que se explica cómo llevar a cabo el ajuste de datos.
- ❑ <http://www.dal.ca/~jblake/ieng3432/Slides/6.1%20Input%20Analysis.ppt>
PowerPoint que explica la importancia de las distribuciones de probabilidad en la simulación.
- ❑ <http://www.informs-cs.org/wsc00papers/038.PDF>
Artículo de Averill M. Law en el que se comentan aspectos interesantes sobre el ajuste de datos mediante distribuciones teóricas dentro del ámbito de la simulación.