



Asignatura:

**Ingeniería Industrial**

## Identificación y descripción gráfica de datos

### Índice de Contenidos

<b>1</b>	<b>Introducción .....</b>	<b>2</b>
<b>2</b>	<b>Gráficos de probabilidad .....</b>	<b>2</b>
2.1	Linealización de una distribución.....	3
2.2	Estimación de la f.d. empírica. ....	9
<b>3</b>	<b>Descripción gráfica de los datos.....</b>	<b>15</b>

# Identificación y descripción gráfica de datos

## 1 Introducción

---

Supóngase que se dispone de una muestra aleatoria de tiempos de fallo asociados a un determinado dispositivo, i.e., se tienen  $n$  observaciones (aleatorias e independientes) de la variable aleatoria  $T =$  “**tiempo transcurrido hasta que se produce el fallo del dispositivo**”. Lo primero que conviene hacer es tratar de identificar alguna distribución conocida a la cual se ajusten bien las observaciones, pues ello simplificaría bastante el análisis descriptivo de los datos, así como la realización de inferencias sobre la población subyacente.

### Descripción paramétrica y descripción no paramétrica

Si se logra aproximar la distribución de  $T$  mediante alguna distribución teórica conocida, será posible realizar una **descripción paramétrica** de la v.a.  $T$ , i.e., se podrá usar la distribución teórica para proporcionar estimaciones de la función de supervivencia,  $R(t)$ , de la función de densidad de probabilidad,  $f(t)$ , y de la tasa de fallos  $h(t)$ .

En caso contrario, será necesario recurrir a una **descripción no paramétrica**, i.e. proporcionar estimaciones puntuales de  $R(t)$ ,  $f(t)$  y  $h(t)$  a partir de las observaciones obtenidas (sin poder suponer que éstas siguen una determinada distribución teórica).

## 2 Gráficos de probabilidad

---

Los gráficos de probabilidad constituyen un método gráfico para tratar de buscar una distribución teórica (exponencial, Weibull, Gamma, etc.) que se ajuste bien a las observaciones. Lo que se pretende con un gráfico de probabilidad es comparar

una función de distribución teórica con la función de distribución empírica que se obtiene a partir de las observaciones.

### Gráfico de probabilidad

Un **gráfico de probabilidad** muestra dos gráficos superpuestos:

- El de la función de distribución  $F(t)$  asociada a una determinada distribución teórica (f.d. teórica)
- El de una nube de puntos superpuesta a la f.d. teórica. Los puntos de esta nube representan estimaciones puntuales (y no paramétricas) de la función de distribución asociada a las observaciones de  $T$  (f.d. empírica)

A fin de facilitar la comparación visual entre ambas funciones de distribución (teórica y empírica), se suelen emplear transformaciones de las variables  $t$  y  $F(t)$  de forma que la f.d. teórica esté "linealizada" (i.e., la representación gráfica de la misma sea una recta).

Evidentemente, cuanto más se aproxime la nube de puntos (f.d. empírica) a la recta (f.d. teórica), tanto mejor será el ajuste. El gráfico de probabilidad permite pues descartar visualmente aquellas distribuciones teóricas que, claramente, no ajustan bien a los datos, así como seleccionar otras distribuciones que, al menos en apariencia, puedan proporcionar buenos ajustes.

### Observación: validación del ajuste

En todo caso, siempre será necesario validar la bondad del ajuste utilizando alguna técnica más objetiva que la simple inspección visual como, por ejemplo, los **contrastos de hipótesis sobre la bondad del ajuste**.

## 2.1 Linealización de una distribución.

Antes de la aparición de *software* especializado, para poder representar un gráfico de probabilidad era necesario utilizar una plantilla especial -que tuviera convenientemente adaptadas las escalas en ambos ejes- a fin de que la f.d. teórica

considerada tuviese forma rectilínea. Para cada distribución teórica, se utilizaba un tipo especial de escalas y, por tanto, una plantilla distinta.

### Proceso de linealización de una f.d.

El **proceso de linealización** de la f.d. asociada a una distribución consiste en encontrar las transformaciones  $g_1$  y  $g_2$  adecuadas para las variables  $t$  y  $F(t)$  de modo que al representar  $y = g_2(F(t))$  vs.  $x = g_1(t)$  se obtenga una recta.

### Linealización de la f.d. asociada a una exponencial

Si  $F(t)$  es la f.d. asociada a una distribución exponencial de parámetro  $\beta$  (*scale*), las transformaciones a emplear para lograr su linealización son:

$$y = -\ln(1 - F(t)) \quad y \quad x = t$$

### Demostración

La f.d. asociada a una distribución exponencial viene dada por la expresión siguiente:

$$F(t) = 1 - \exp\{-t/\beta\}$$

donde  $\beta > 0$  (*scale*) es el parámetro que define la distribución. Esta función puede ser linealizada (i.e., puesta de la forma:  $y = a + bx$ ) como sigue:

$$\begin{aligned} F(t) = 1 - \exp\{-t/\beta\} &\Leftrightarrow \ln(1 - F(t)) = \ln(\exp\{-t/\beta\}) \Leftrightarrow \ln(1 - F(t)) = -(t/\beta) \Leftrightarrow \\ &\Leftrightarrow -\ln(1 - F(t)) = t/\beta \end{aligned}$$

Haciendo el doble cambio de variable siguiente:

$$y = -\ln(1 - F(t)) \quad y \quad x = t$$

Es posible expresar la f.d. anterior como:

$$y = \frac{1}{\beta} x$$

### Linealización de la f.d. asociada a una Weibull

Si  $F(t)$  es la f.d. asociada a una distribución Weibull de parámetros  $\alpha$  (*shape*) y  $\beta$  (*scale*), las transformaciones a emplear para lograr su linealización son:

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

### Demostración

La f.d. asociada a una distribución Weibull viene dada por la expresión siguiente:

$$F(t) = 1 - \exp\left\{-\left(t/\beta\right)^\alpha\right\}$$

donde  $\alpha > 0$  (*shape*) y  $\beta > 0$  (*scale*) son los dos parámetros que definen la distribución. Esta función puede ser linealizada (i.e., puesta de la forma:  $y = a + bx$ ) como sigue:

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

Haciendo el doble cambio de variable siguiente:

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

Es posible expresar la f.d. anterior como:

$$y = \alpha x - \alpha \ln \beta$$

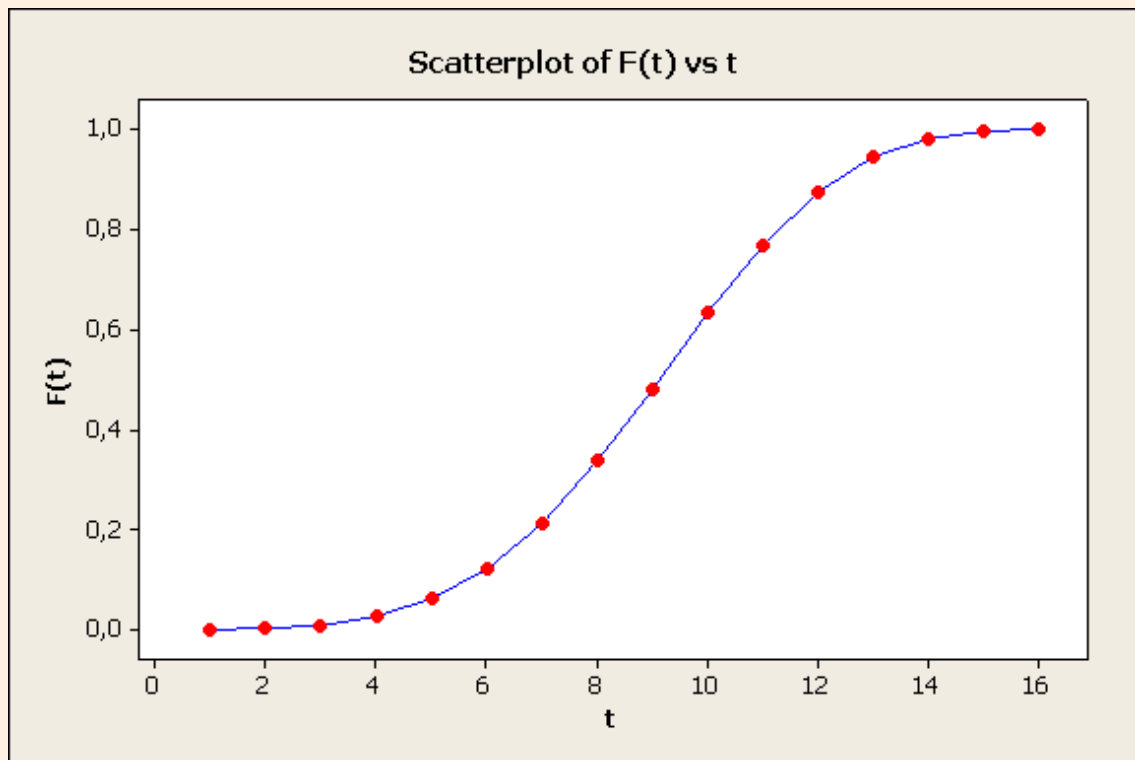
### Ejemplo 1: Linealización de una distribución Weibull

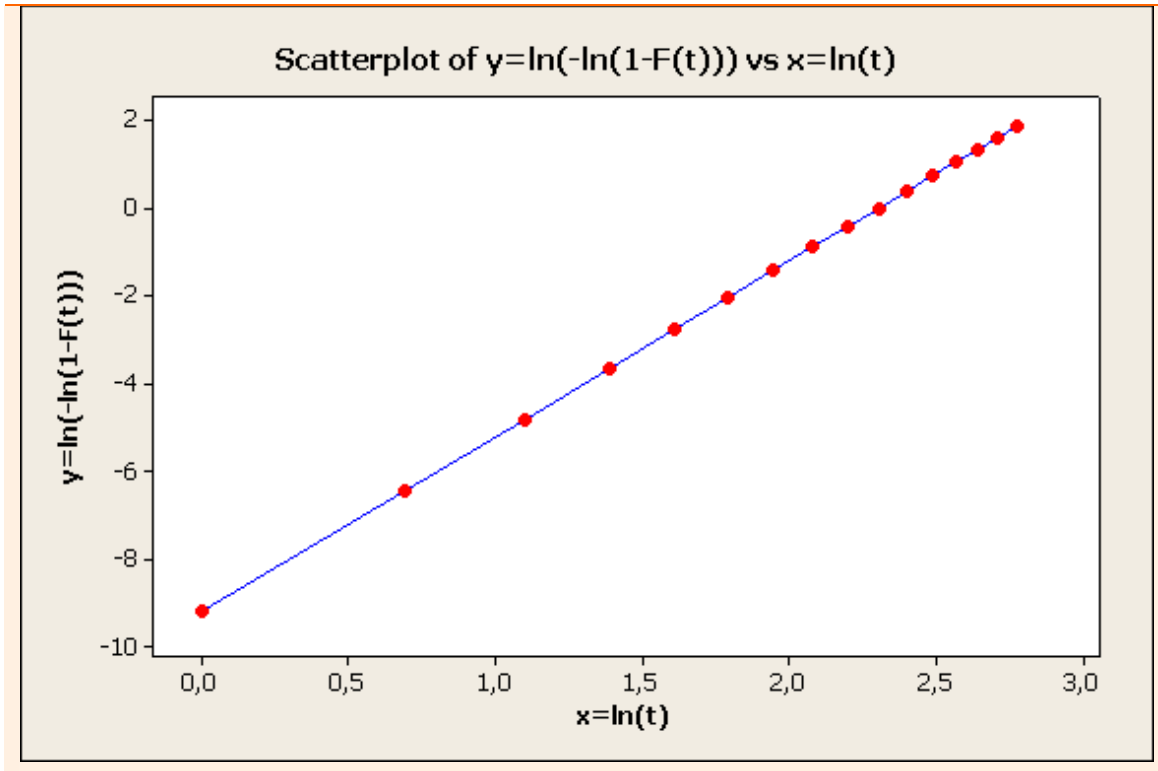
A continuación se representan gráficamente, con ayuda de MINITAB, las dos versiones ("estándar" y "linealizada") de la f.d. asociada a una Weibull con parámetros  $\alpha = 4$  y  $\beta = 10$ :

MINITAB - Untitled - [Worksheet 1 \*\*\*]

File Edit Data Calc Stat Graph Editor Tools Window Help

↓	C1	C2	C3	C4
	t	F(t) Weibull shape=4 scale=10	x=ln(t)	y=ln(-ln(1-F(t)))
1	1	0,000100	0,00000	-9,21034
2	2	0,001599	0,69315	-6,43775
3	3	0,008067	1,09861	-4,81589
4	4	0,025275	1,38629	-3,66516
5	5	0,060587	1,60944	-2,77259
6	6	0,121553	1,79176	-2,04330
7	7	0,213451	1,94591	-1,42670
8	8	0,336084	2,07944	-0,89257
9	9	0,481129	2,19722	-0,42144
10	10	0,632121	2,30259	0,00000
11	11	0,768714	2,39790	0,38124
12	12	0,874268	2,48491	0,72929
13	13	0,942507	2,56495	1,04946
14	14	0,978541	2,63906	1,34589
15	15	0,993670	2,70805	1,62186
16	16	0,998575	2,77259	1,88001
17				





### Observación: Construcción de plantillas para gráficos de probabilidad

Conocidas las transformaciones  $g_1$  y  $g_2$  que linealizan una determinada f.d. teórica, es posible utilizar sus inversas para obtener  $t$  y  $F(t)$  en función de  $x$  e  $y$  respectivamente, i.e.:  $t = g_1^{-1}(x)$  y  $F(t) = g_2^{-1}(y)$ . Con ello, resulta inmediato volver a "etiquetar" los ejes  $x$  e  $y$  a fin de que muestren los correspondientes valores de  $t$  y  $F(t)$ . Ello permitirá representar, sobre estos "ejes transformados", directamente los valores  $t$  y  $F(t)$ . Así es como se construyen las plantillas de gráfico de probabilidad asociadas a una distribución determinada.

### Ejemplo 2: Plantilla para gráfico de probabilidad de una Weibull

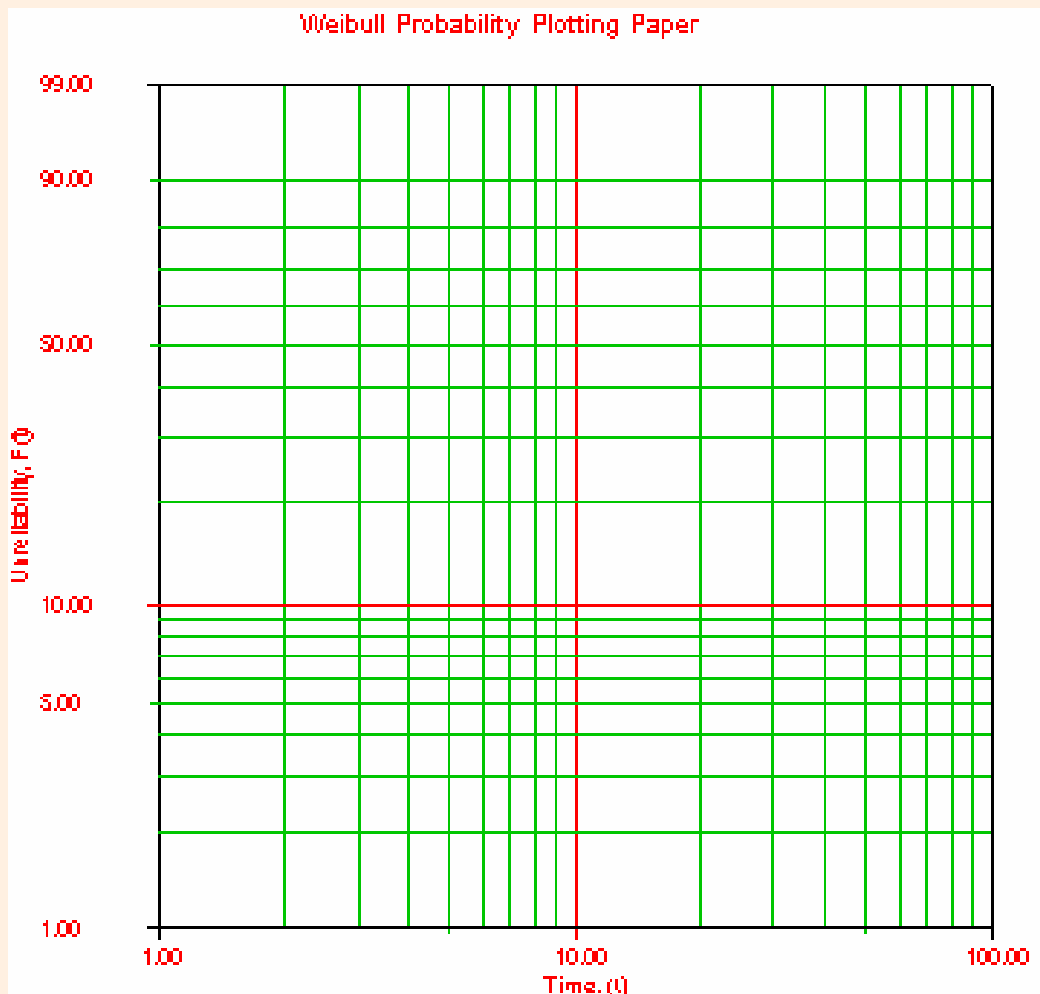
En el caso de una distribución Weibull, vimos que las transformaciones a aplicar eran:

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

Deshaciendo dichas transformaciones se obtiene que:

$$F(t) = 1 - \left( \exp\{\exp\{y\}\} \right)^{-1} \quad y \quad t = \exp\{x\}$$

Por tanto, bastaría con utilizar dichas “transformaciones inversas” para volver a etiquetar los ejes del último de los gráficos del ejemplo anterior. El resultado sería una plantilla como la siguiente, sobre la cual se representará directamente (sin transformaciones adicionales)  $F(t)$  vs.  $t$ :



## 2.2 Estimación de la f.d. empírica.

Una vez seleccionada la distribución teórica que se usará para tratar de ajustar las observaciones, se procederá a representar la nube de puntos sobre la correspondiente plantilla de gráfico de probabilidad. Cada uno de los puntos  $(t_i, \hat{F}(t_i))$ ,  $i=1,2,\dots,n$ , representará a cada uno de los  $n$  tiempos de fallo observados,  $t_i$ , junto con el correspondiente valor estimado de la f.d. experimental,  $\hat{F}(t_i)$ . El diagrama de probabilidad permitirá apreciar visualmente si la nube de puntos sigue un patrón aproximadamente lineal (lo cual favorecería la hipótesis de que la distribución teórica seleccionada se ajusta bien a las observaciones) o, por el contrario, si ésta sigue un patrón muy distinto del lineal (lo cual permitiría descartar la distribución teórica seleccionada a efectos de explicar el comportamiento de las observaciones). A continuación se explica un método (no el único, aunque sí uno de los más utilizados) que permite obtener estimaciones puntuales de la f.d. empírica,  $\hat{F}(t_i)$ , a partir de las observaciones,  $t_i$ .

### Proposición: Estimación puntual (no paramétrica) de la f.d. empírica

Sean  $t_1 < t_2 < \dots < t_n$  son los  $n$  tiempos de fallo observados. Para  $i=1,2,\dots,n$ , se puede obtener la segunda componente del punto  $(t_i, \hat{F}(t_i))$  tomando:

$$\hat{F}(t_i) = \frac{i-0.3}{n+0.4}$$

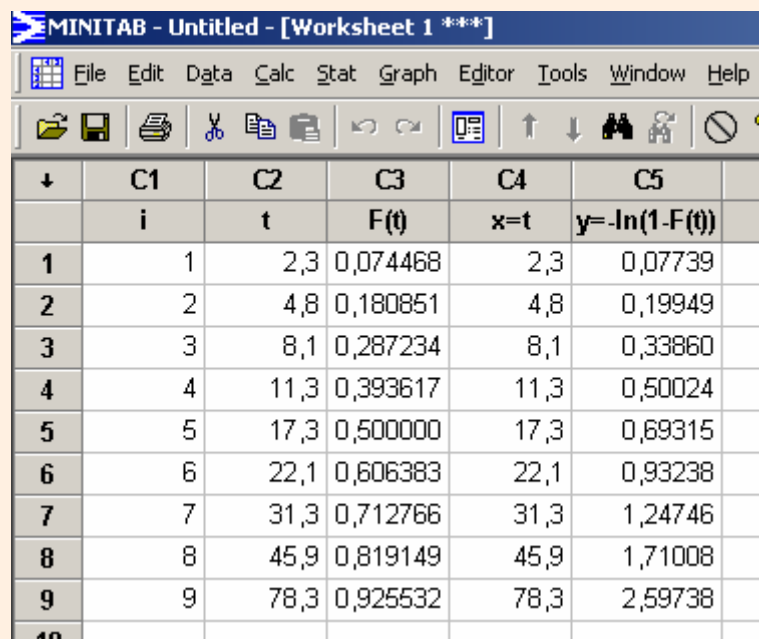
Éste es el llamado **método del rango mediano de Bernard** para el cálculo de  $\hat{F}(t_i)$ .

### Ejemplo 3: Gráfico de probabilidad de una exponencial

Se han obtenido los siguientes tiempos hasta el fallo (en días) de un motor diesel: 31.3, 45.9, 78.3, 22.1, 2.3, 4.8, 8.1, 11.3, y 17.3. Se sabe que, cada vez que se ha estropeado el motor, éste ha sido reparado y devuelto a un estado de "como

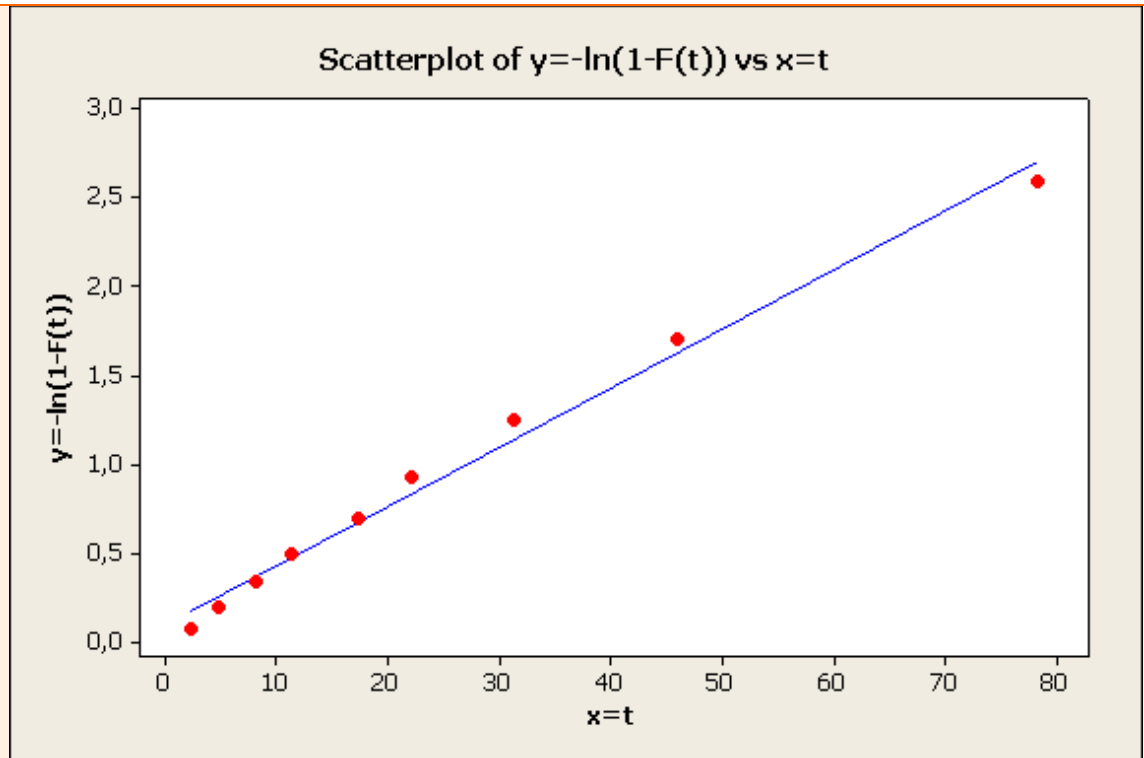
nuevo". Determina si tiene sentido suponer que los tiempos de fallo siguen una distribución aproximadamente exponencial.

Lo primero es ordenar, de menor a mayor, los tiempos hasta el fallo. A continuación se tendrá que crear una tabla con todos los valores necesarios para poder representar el gráfico de probabilidad. Dado que en este caso no usaremos una plantilla especial para representar los puntos  $(t_i, \hat{F}(t_i))$ , será necesario calcular sus valores transformados  $(g_1(t_i), g_2(\hat{F}(t_i)))$ :



	C1	C2	C3	C4	C5
	i	t	F(t)	x=t	y=-ln(1-F(t))
1	1	2,3	0,074468	2,3	0,07739
2	2	4,8	0,180851	4,8	0,19949
3	3	8,1	0,287234	8,1	0,33860
4	4	11,3	0,393617	11,3	0,50024
5	5	17,3	0,500000	17,3	0,69315
6	6	22,1	0,606383	22,1	0,93238
7	7	31,3	0,712766	31,3	1,24746
8	8	45,9	0,819149	45,9	1,71008
9	9	78,3	0,925532	78,3	2,59738
10					

Finalmente, bastará con representar la nube de puntos  $(g_1(t_i), g_2(\hat{F}(t_i)))$ , i.e., la nube de puntos  $(x_i, y_i)$ :



En la imagen anterior se observa la nube de puntos y su recta de regresión asociada. Se observa que, en efecto, la nube de puntos sigue un patrón bastante lineal, por lo cual es coherente pensar que las observaciones puedan seguir una distribución exponencial.

#### Ejemplo 4: Gráfico de probabilidad con censura

Se considera ahora el caso de una compañía que fabrica cubiertas para motores, cubiertas que pueden estropearse rápidamente si se ven sometidas a temperaturas elevadas. Supóngase que se lleva a cabo un experimento, de duración determinada, consistente en:

- a) someter 50 cubiertas a 80°C de temperatura y registrar, en meses, sus respectivos tiempos de fallo (variable T80)
- b) someter 40 cubiertas a 100°C de temperatura y registrar, en meses, sus respectivos tiempos de fallo (variable T100)

Algunas de las cubiertas que se empezaron a estudiar, o bien fallaron debido a causas distintas a la temperatura, o bien no continuaron en el estudio por motivos

diversos y, por tanto, se desconoce el instante en que fallaron (observaciones censuradas a derecha). Por ello, para cada cubierta se han considerado también dos variables, C80 y C100, que especifican si los tiempos obtenidos pertenecen a observaciones completas (1) o bien a observaciones censuradas (0).

La tabla siguiente muestra los datos registrados en el experimento:

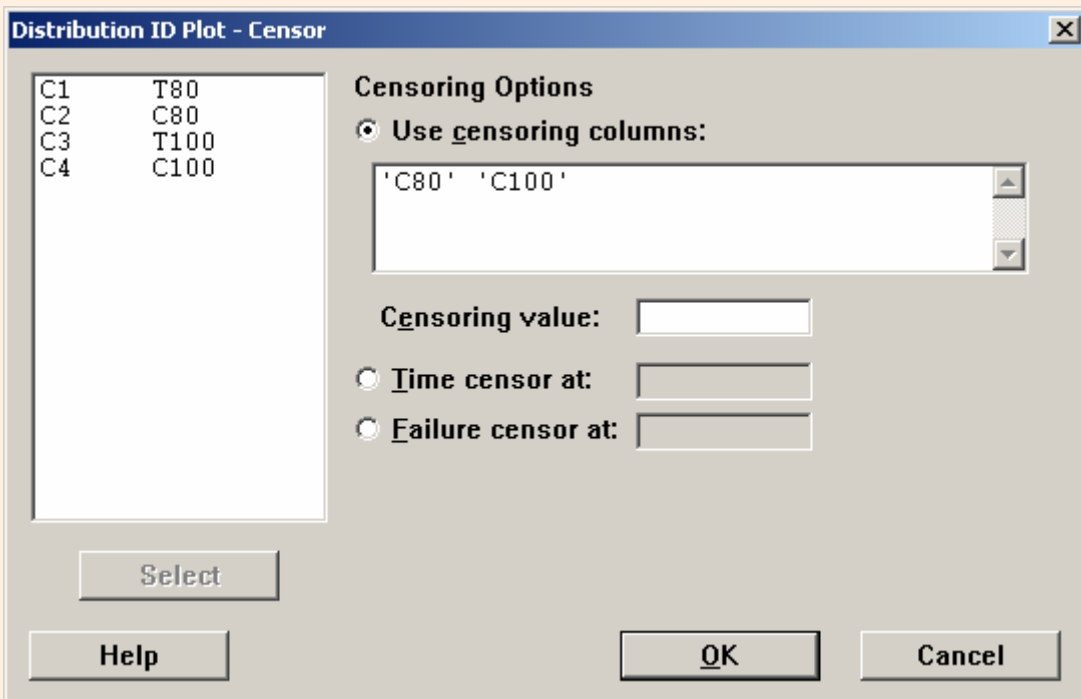
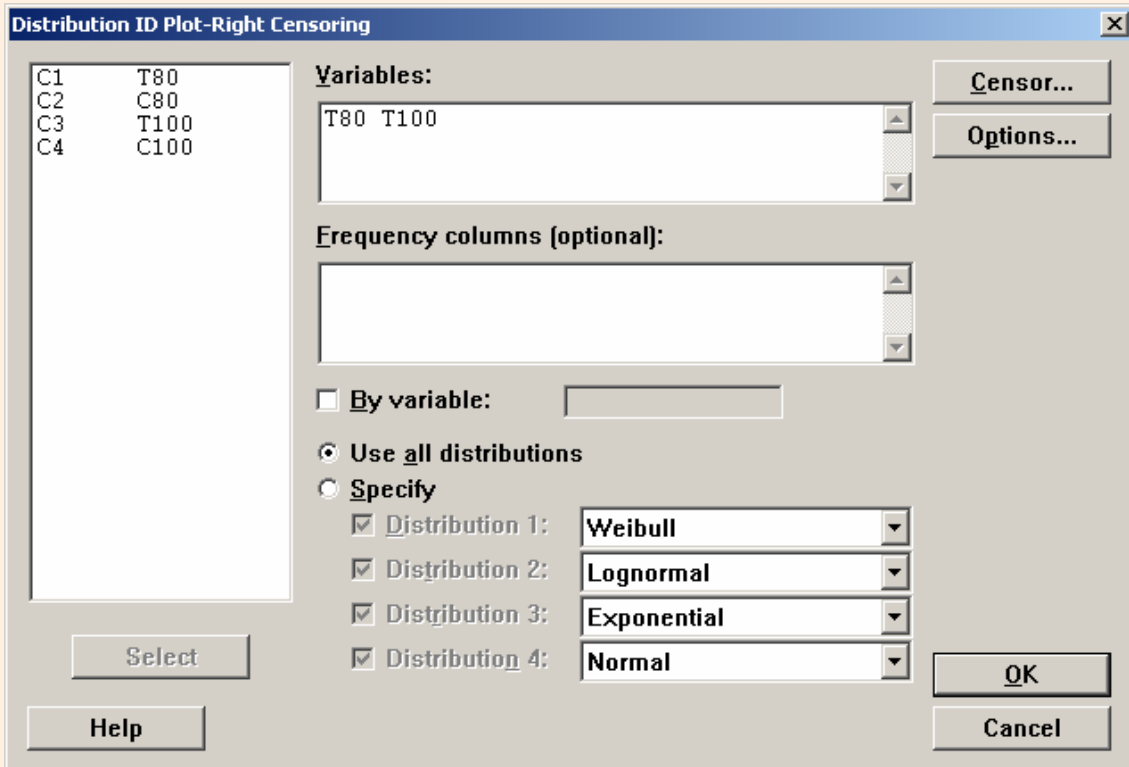
Obs	T80	C80	Obs	T80	C80		Obs	T100	C100	Obs	T100	C100
1	50	1	26	51	1		1	101	0	21	27	1
2	60	1	27	27	1		2	11	1	22	62	0
3	53	1	28	52	1		3	48	1	23	54	1
4	40	1	29	48	1		4	32	1	24	84	0
5	51	1	30	79	0		5	36	1	25	45	1
6	99	0	31	48	1		6	22	1	26	10	1
7	35	1	32	67	1		7	72	1	27	97	0
8	55	1	33	66	1		8	69	1	28	6	1
9	74	1	34	27	1		9	35	1	29	37	1
10	101	0	35	59	1		10	29	1	30	38	1
11	56	1	36	48	1		11	18	1	31	40	1
12	45	1	37	77	0		12	38	1	32	30	1
13	61	1	38	58	1		13	39	1	33	64	0
14	92	0	39	51	1		14	68	1	34	46	1
15	73	0	40	97	0		15	36	1	35	46	1
16	51	1	41	34	1		16	18	1	36	24	1
17	49	1	42	79	0		17	25	1	37	76	1
18	24	1	43	91	0		18	14	1	38	18	1
19	37	1	44	41	1		19	77	0	39	16	1
20	31	1	45	64	1		20	47	1	40	45	1
21	67	1	46	81	0							
22	62	1	47	105	0							
23	100	0	48	84	0							
24	58	1	49	54	1							
25	46	1	50	23	1							

En esta ocasión, vamos a utilizar MINITAB para que nos genere varios gráficos de probabilidad a partir de los cuales intentar identificar una distribución que, al menos en apariencia, se ajuste bien a los datos:

Una vez introducidos los datos en MINITAB, seleccionamos la opción **Stat > Reliability/Survival > Distribution Analysis (Right Censoring) > Distribution ID Plot... :**

En las pantallas siguientes, se eligen tanto las variables que contienen los tiempos de fallo como las columnas en las que se indica si ha habido o no censura (cada

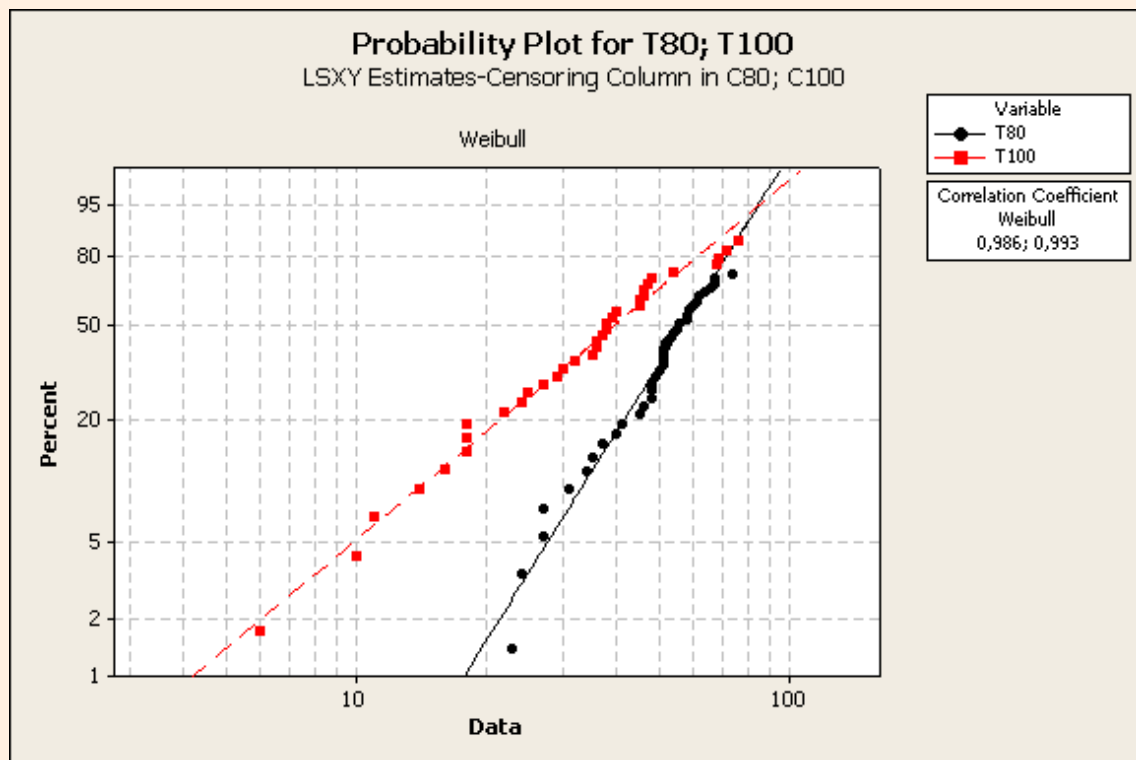
una de éstas se asociará a una variable según el orden de introducción de las mismas):

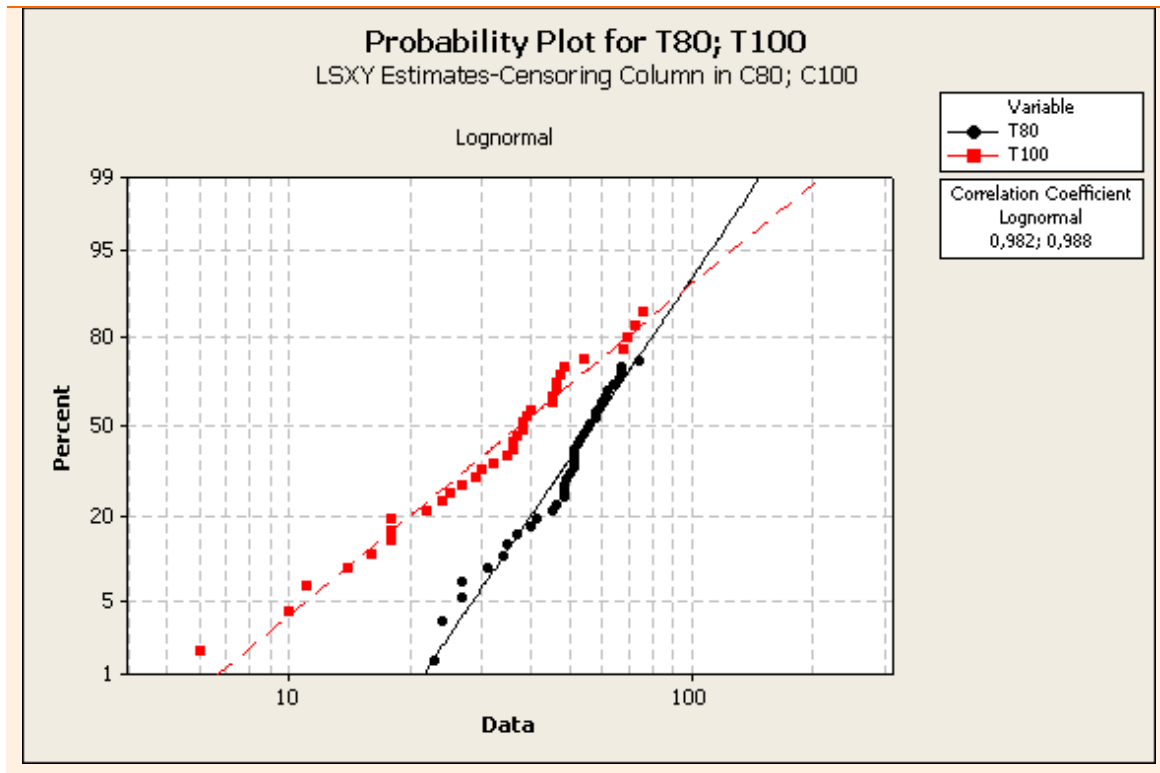


Observar que una alternativa al uso de columnas indicadoras de censura sería fijar

bien el tiempo que ha de transcurrir o bien el número de observaciones que han de fallar como indicador de censura (censura por tiempo o por fallos).

Si los puntos representados en el gráfico están suficientemente próximos a la recta, podremos considerar a la distribución teórica asociada como una buena candidata al ajuste. Por lo que se observa en los gráficos siguientes, la distribución que mejor se ajusta a los datos parece ser la log-normal (resulta conveniente prestar atención especial a los valores de los extremos):





### 3 Descripción gráfica de los datos

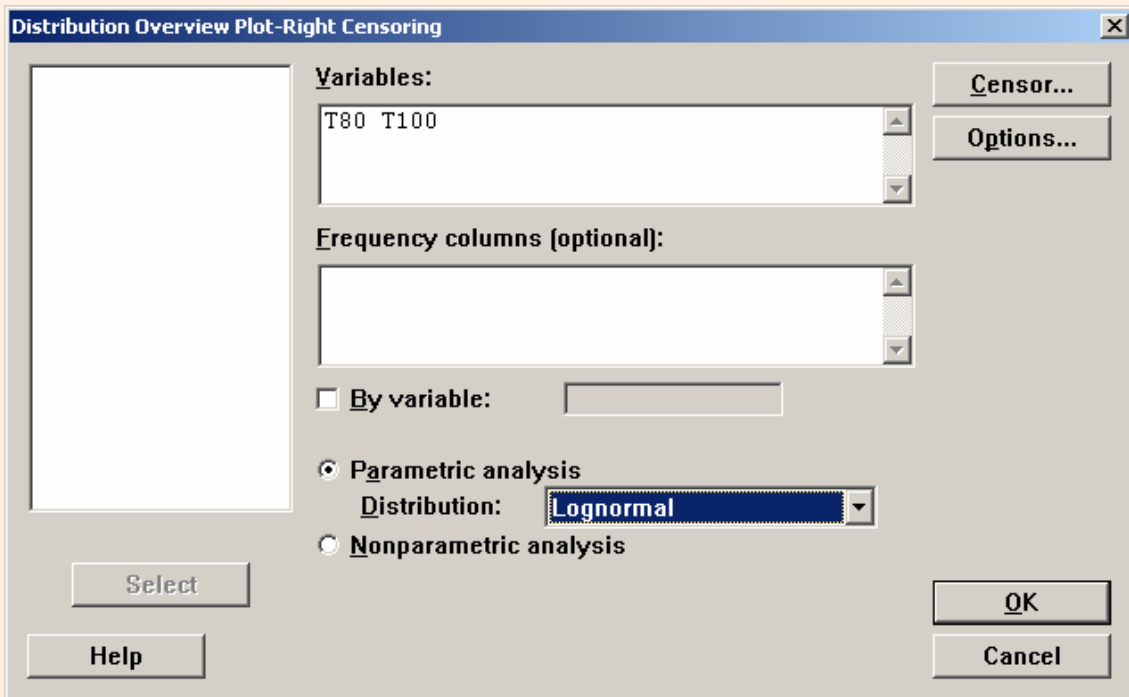
Tras haber usado los gráficos de probabilidad para tratar de encontrar alguna distribución teórica que se ajuste bien a los tiempos de fallo observados, será conveniente realizar una descripción gráfica de las observaciones. En caso de haber hallado alguna distribución teórica que, al menos en apariencia, se ajuste bien a las observaciones, se optará por realizar una descripción paramétrica de las mismas. Si, por el contrario, las observaciones no parecen ajustarse a ninguna distribución teórica, se optará por recurrir a una descripción no paramétrica.

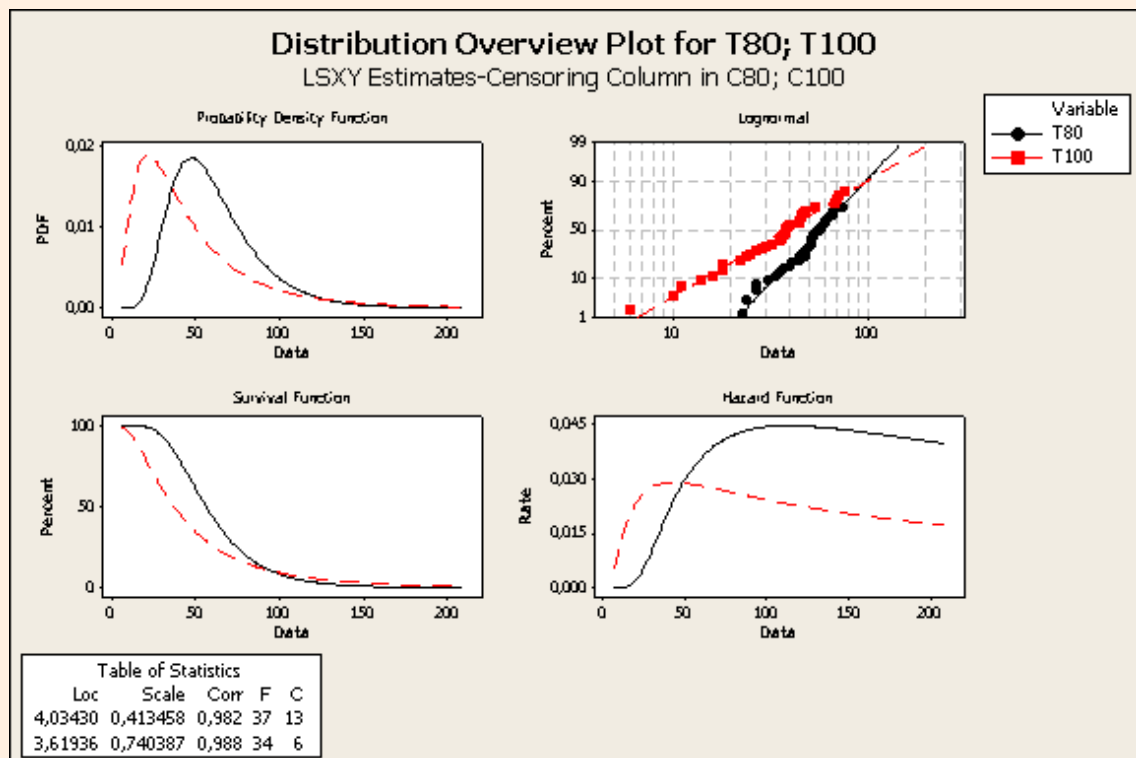
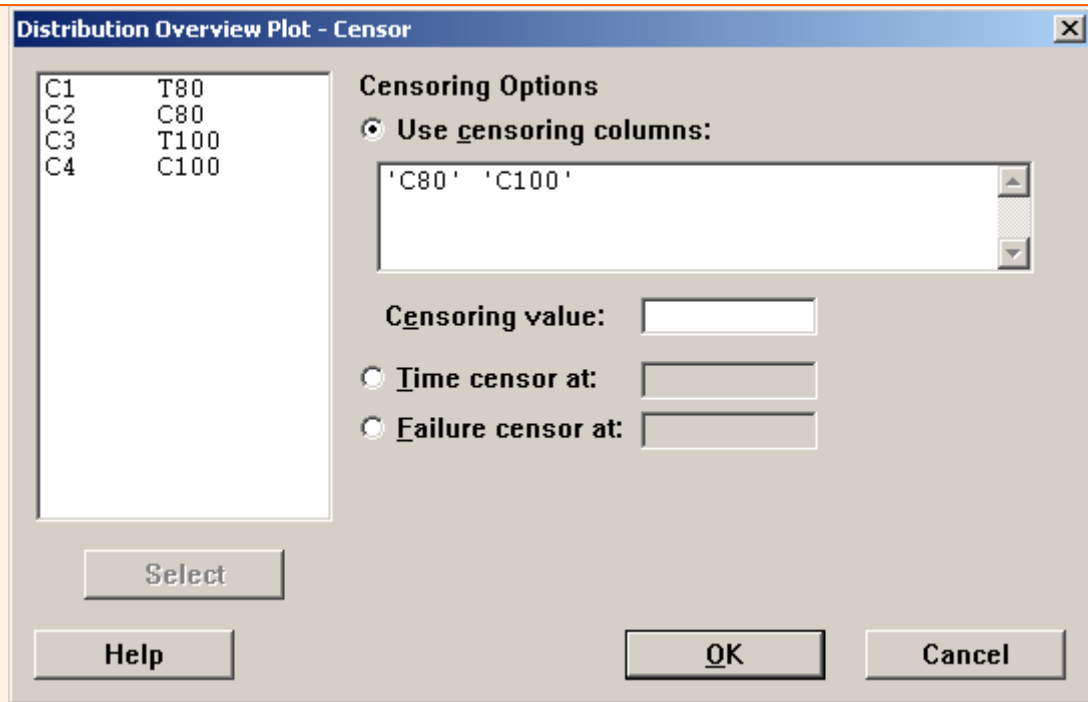
Sea cual sea la aproximación escogida (paramétrica o no paramétrica), resultará necesario aplicar sofisticadas técnicas estadísticas que serán introducidas con detalle en capítulos posteriores. Así, por ejemplo, si se opta por realizar un análisis paramétrico de los datos, será imprescindible estimar los parámetros concretos que definen la distribución teórica seleccionada para el ajuste; por su parte, en el caso de optar por un análisis no paramétrico, será necesario estimar la función de supervivencia mediante el método de Kaplan-Meier. En este punto, sin embargo, podríamos aprovechar las capacidades del *software* estadístico para realizar,

avanzándonos a la posterior explicación teórica de las técnicas que éste utiliza, una primera descripción gráfica de las observaciones.

### Ejemplo 5: Descripción gráfica paramétrica

Siguiendo con el ejemplo 4, el de las cubiertas para motores, usaremos la distribución log-normal (que, según se vio, parecía ajustarse bien a las observaciones) para describir gráficamente los datos con ayuda de MINITAB. Para ello, usaremos la opción **Stat > Reliability/Survival > Distribution Analysis (Right Censoring) > Distribution Overview Plot...** :





Las cuatro gráficas anteriores describen la distribución de los tiempos de fallo de las cubiertas para los dos niveles diferentes de temperatura considerados (80 y 100 grados). A partir de las mismas, es posible determinar, p.e., cuánto más probable resulta el que las cubiertas fallen si se encuentran sometidas a una temperatura de

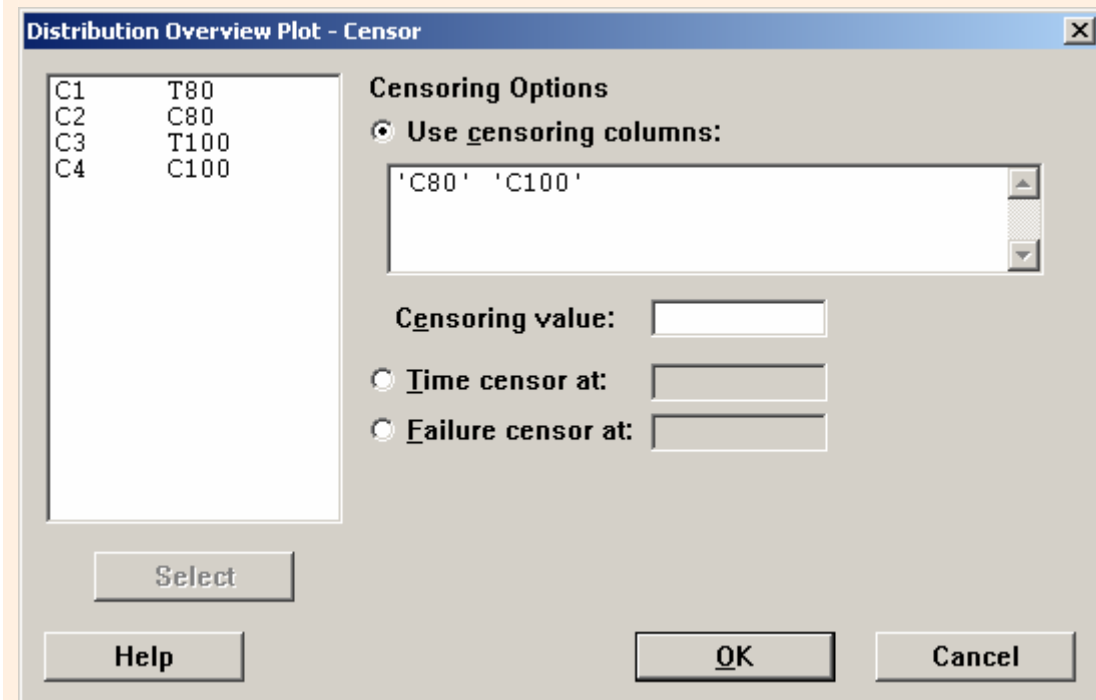
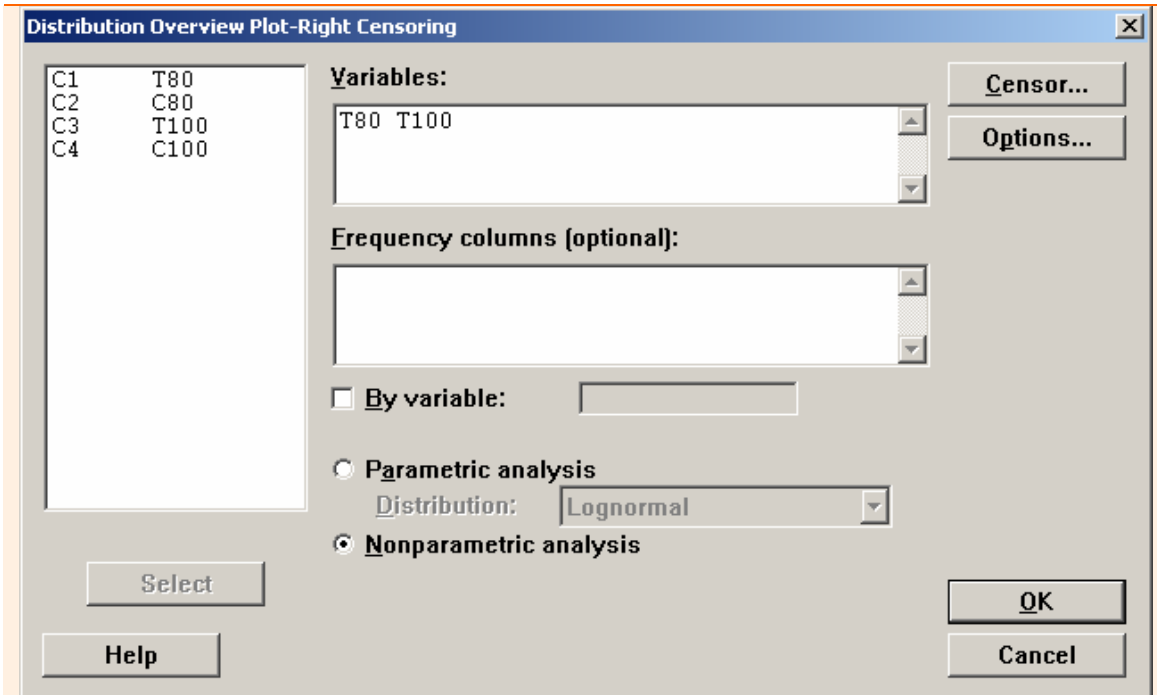
100° C que si lo están a una de 80° C.

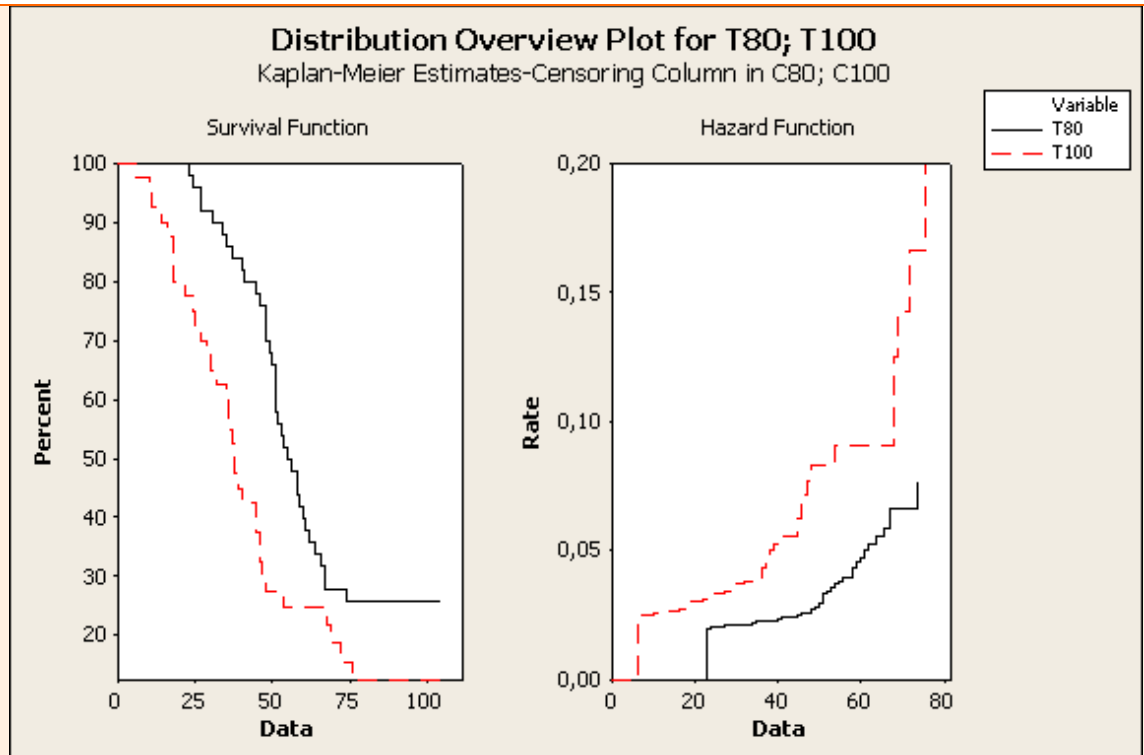
Así, p.e., se observa (a partir del gráfico de supervivencia) que, tras 50 meses, sólo sobrevivirán (aproximadamente) un 30% de las cubiertas sometidas a 100° C, mientras que el porcentaje de supervivientes a los 50 meses sube hasta (aproximadamente) un 65% para las cubiertas sometidas a 80° C.

Obsérvese, además, que el *output* de MINITAB ya proporciona las estimaciones para los parámetros de localización y escala que definen exactamente la distribución log-normal que sigue cada conjunto de tiempos de fallo (los asociados a temperaturas de 80°C y los asociados a temperaturas de 100°C). Se explicará en capítulos posteriores cómo se obtienen dichas estimaciones.

### **Ejemplo 6: Descripción gráfica no paramétrica**

Siguiendo con el ejemplo 4, supondremos ahora que no hemos sido capaces de encontrar ninguna distribución teórica que se ajuste bien a las observaciones y, por tanto, usaremos una aproximación no paramétrica para describir gráficamente los datos con ayuda de MINITAB. Para ello, usaremos nuevamente la opción **Stat > Reliability/Survival > Distribution Analysis (Right Censoring) > Distribution Overview Plot...** :





A partir de las funciones de supervivencia se aprecia que hay una sustancial diferencia entre los tiempos de fallo de las cubiertas a 80° C y los de las cubiertas a 100° C: claramente, a una temperatura de 80° C la mayor parte de las cubiertas perdura durante más tiempo que a 100° C.

Por su parte, la gráfica de las tasas de riesgo muestra dos funciones crecientes, siendo la de mayor pendiente la asociada a las cubiertas que soportan más temperatura.

Nuevamente, se aprecia cómo transcurridos 50 meses, solo sobrevivirán aprox. un 30% de las cubiertas sometidas a 100° C, mientras que este porcentaje llega al 65-70% para cubiertas a 80° C. Notar, además, que aproximadamente un 50% de las cubiertas a 100° C habrán fallado entre los 35 y 40 meses. Por otro lado, en el caso de las cubiertas a 80° C, un 50% de las mismas sobrevivirá hasta los 55-60 meses.

Observar que, al realizar una aproximación no paramétrica a los datos, MINITAB hace uso del estimador Kaplan-Meier, el cual se explicará detalladamente en capítulos posteriores.

Al finalizar este capítulo conviene recordar que los gráficos de probabilidades constituyen un método visual para encontrar (o descartar) distribuciones teóricas **candidatas** a proporcionar buenos ajustes a las observaciones. Antes de poder dar por realmente bueno un ajuste de las observaciones mediante una distribución teórica, será imprescindible aplicar métodos formales (como los contrastes de hipótesis sobre la bondad del ajuste) que corroboren la percepción visual.