



Asignatura:

Ingeniería Industrial

Análisis no paramétrico de los datos

Índice de Contenidos

1	Introducción	2
2	Estimación de la fiabilidad en el caso de obs. completas.....	2
2.1	Estimación puntual	2
2.2	Estimación por intervalos	3
3	Estimación de la fiabilidad en el caso de obs. censuradas	5
3.1	Estimación puntual	5
3.2	Estimación por intervalos	8
4	Análisis no paramétrico con MINITAB.....	8

Análisis no paramétrico de los datos

1 Introducción

En ocasiones puede resultar ventajoso, o incluso necesario, comenzar el análisis de las observaciones con métodos analíticos y gráficos que no requieran de grandes supuestos previos sobre el modelo. Tales métodos no paramétricos permiten interpretar los datos obtenidos sin la distorsión que podría causar la elección de un modelo subyacente no demasiado acertado. En algunos casos, estos métodos no paramétricos serán suficientes para realizar el análisis de los datos. En otras ocasiones, sin embargo, supondrán un paso intermedio hacia un modelo más estructurado (paramétrico) que permita profundizar más en el análisis de las observaciones.

En este tema se presentaran algunas técnicas no paramétricas que permitirán estimar la función de distribución $F(t)$ (o, alternativamente, la función de supervivencia $R(t)=1-F(t)$) asociada a las observaciones. La técnica concreta a usar en cada caso dependerá del tipo de observaciones de que se disponga (completas o censuradas).

2 Estimación de la fiabilidad en el caso de obs. completas

2.1 Estimación puntual

El primer paso será introducir el estimador MV para la función de distribución (o, alternativamente, para la función de fiabilidad) de la variable que representa los tiempos de fallo del dispositivo:

Función de distribución empírica $\hat{F}(t)$ (obs. completas)

Sea T la variable aleatoria que representa el tiempo de fallo de un determinado dispositivo. Dada una muestra ordenada de observaciones completas, $t_1 < t_2 < \dots < t_n$, se define la **función de distribución empírica** de T , $\hat{F}(t)$, como:

$$\hat{F}(t) = \begin{cases} 0 & 0 < t < t_1 \\ \frac{i}{n} & t_i \leq t < t_{i+1} \quad i = 1, 2, \dots, n-1 \\ 1 & t_n \leq t < \infty \end{cases}$$

Observaciones: Función de distribución empírica $\hat{F}(t)$

- $\hat{F}(t)$ representa la fracción de dispositivos que han fallado antes del instante t
- Se puede probar que $\hat{F}(t)$ es el estimador MV de $F(t)$, la función de distribución real de T
- Obtenido $\hat{F}(t)$, resulta inmediato obtener el estimador MV para $R(t)$:
 $\hat{R}(t) = 1 - \hat{F}(t)$
- Es usual representar gráficamente $\hat{R}(t)$ mediante una función escalonada

Ejemplo 1: Estimación MV para $R(t)$ (observaciones completas)

Se inicia un test de vida sobre diez dispositivos idénticos, obteniéndose los siguientes tiempos de fallo: 89, 132, 202, 263, 321, 362, 421, 473, 575 y 663. Se desea obtener el estimador no paramétrico para $R(t)$ cuando $t = 350$ horas.

Usando la expresión de la función de distribución empírica, y puesto que en este caso $n = 10$ y hay sólo 5 observaciones cuyo tiempo de fallo sea inferior a 350 horas, se tiene que:

$$\hat{F}(350) = \frac{5}{10} = 0.5 \text{ y, por tanto, } \hat{R}(350) = 1 - \hat{F}(350) = 0.5$$

2.2 Estimación por intervalos

En cierto sentido, $F(t)$ se puede interpretar como la probabilidad de éxito, p , en una prueba de Bernoulli. Al analizar una muestra formada por n dispositivos idénticos, se puede considerar que la variable $r =$ "número de dispositivos que han fallado antes del instante t " se distribuye según una binomial de parámetros p y n , con lo que: $E[r] = n \cdot p = n \cdot F(t)$ y $Var[r] = n \cdot p \cdot (1 - p) = n \cdot F(t) \cdot (1 - F(t))$.

Conforme aumenta el tamaño muestral, n , la distribución binomial que sigue r tiende a comportarse como una distribución normal de media $\mu = n \cdot F(t)$ y varianza $\sigma^2 = n \cdot F(t) \cdot (1 - F(t))$ (puesto que $F(t)$ es desconocido, en la práctica usaremos $\mu \approx n \cdot \hat{F}(t)$ y $\sigma^2 \approx n \cdot \hat{F}(t) \cdot (1 - \hat{F}(t))$). Este hecho se puede utilizar para obtener un intervalo de confianza aproximado, a un nivel de confianza de $1 - \alpha$, para $p = F(t)$:

Intervalo de confianza para $F(t)$ (obs. completas)

Sea T la variable aleatoria que representa el tiempo de fallo de un determinado dispositivo. Dada una muestra ordenada de observaciones completas, $t_1 < t_2 < \dots < t_n$, un intervalo de confianza a nivel $1 - \alpha$ para $F(t)$ viene dado por:

$$\left(\hat{F}(t) - z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{F}(t) \cdot (1 - \hat{F}(t))}{n}}, \hat{F}(t) + z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{F}(t) \cdot (1 - \hat{F}(t))}{n}} \right)$$

donde z_p representa el percentil de orden p en una distribución normal tipificada.

Ejemplo 2: Estimación por intervalos para $R(t)$ (observaciones completas)

Usando las diez observaciones del ejemplo 1, se desea hallar un intervalo de confianza, a un nivel del 95%, para $R(t)$ cuando $t = 350$ horas.

Como ya vimos en el ejemplo 1, $\hat{F}(350) = 0.5$. Por tanto, un intervalo de confianza aproximado para $F(350)$, a un nivel de confianza del 95%, vendrá dado por:

$$\left(0.5 - 1.96 \cdot \sqrt{\frac{0.5 \cdot (1-0.5)}{10}}, 0.5 + 1.96 \cdot \sqrt{\frac{0.5 \cdot (1-0.5)}{10}} \right) \text{ i.e.: } (0.1900, 0.6099)$$

Así que un intervalo aproximado para $R(350)$, a un nivel de confianza del 95%, vendrá dado por:

$$(0.3901, 0.8100)$$

3 Estimación de la fiabilidad en el caso de obs. censuradas

Los métodos de estimación puntual y por intervalos considerados en el apartado anterior para el caso de observaciones completas no son, en general, aplicables para el caso de datos con censura. Para este tipo de muestras, resulta necesario recurrir a otros métodos más sofisticados como, p.e., el llamado estimador de Kaplan-Meier o estimador producto-límite, que resulta ser el estimador MV de la función de supervivencia $R(t)$.

3.1 Estimación puntual

Estimador Kaplan-Meier o producto-límite para $R(t)$ (obs. con censura)

Sea T la variable aleatoria que representa el tiempo de fallo de un determinado

dispositivo. Se considera una muestra formada por n dispositivos, de los cuales sólo se observan k tiempos de fallo distintos: $t_1 < t_2 < \dots < t_k$. Sea n_i el número de dispositivos bajo observación justo antes del instante t_i y d_i el número de dispositivos que fallan justo en el instante t_i . Se define el **estimador de Kaplan-Meier** o **estimador producto-límite** para la función de supervivencia de T , $R(t)$ como:

$$\hat{R}(t) = \prod_{j:t_j \leq t} \left(1 - \frac{d_j}{n_j} \right)$$

Observaciones: estimador Kaplan-Meier

- $\hat{R}(t)$ representa la fracción de dispositivos que han sobrevivido al instante t
- Se puede probar que $\hat{R}(t)$ es el estimador MV de $R(t)$
- Obtenido $\hat{R}(t)$, resulta inmediato obtener el estimador MV para $F(t)$:
 $\hat{F}(t) = 1 - \hat{R}(t)$
- Es usual representar gráficamente $\hat{R}(t)$ mediante una función escalonada

Ejemplo 3: Estimación MV para $R(t)$ (censura múltiple)

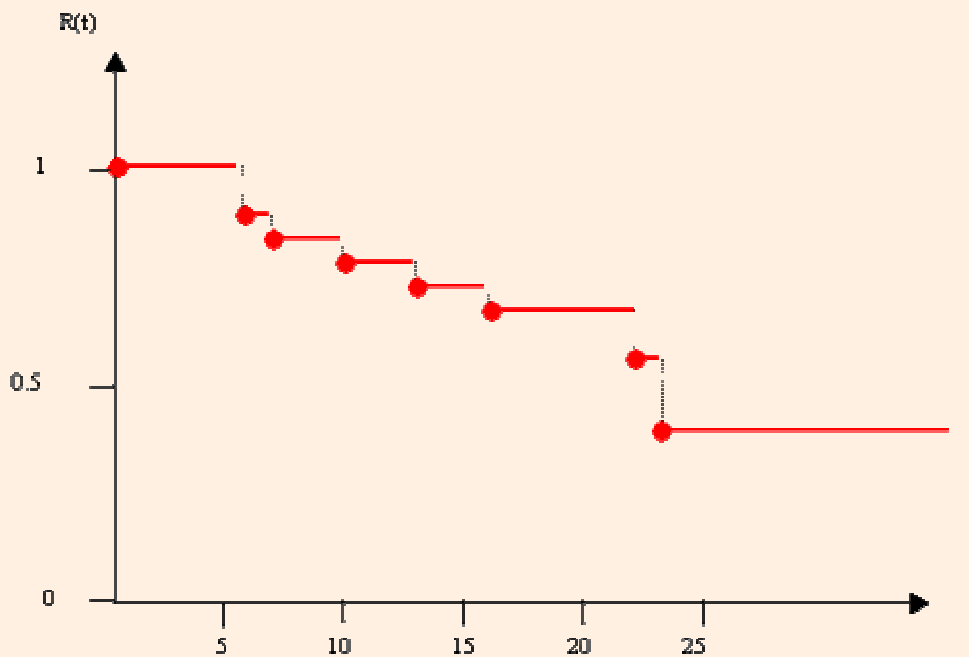
Se inicia un test de vida sobre veintiún dispositivos idénticos, obteniéndose los siguientes tiempos de fallo y tiempos de censura (estos últimos, marcados con un asterisco): 6, 6, 6, 6*, 7, 9*, 10, 10*, 11*, 13, 16, 17*, 19*, 20*, 22, 23, 25*, 32*, 32*, 34*, 35*. Se desea obtener el estimador Kaplan-Meier para $R(t)$.

En este caso tenemos $k = 7$ tiempos de fallo distintos y 12 observaciones censuradas. Construimos la siguiente tabla para calcular el estimador K-P:

i	$[t_i, t_{i+1})$	n_i	d_i	$1 - \frac{d_i}{n_i}$	$\hat{R}(t)$
-----	------------------	-------	-------	-----------------------	--------------

0	[0, 6)	21	--	--	1
1	[6, 7)	21	3	0.8571	0.8571
2	[7, 10)	17	1	0.9412	0.8067
3	[10, 13)	15	1	0.9333	0.7529
4	[13, 16)	12	1	0.9167	0.6902
5	[16, 22)	11	1	0.9091	0.6275
6	[22, 23)	7	1	0.8571	0.5378
7	[23, inf.)	6	1	0.8333	0.4482

La gráfica siguiente muestra cómo sería la representación gráfica de $\hat{R}(t)$:



Conviene notar la forma escalonada de la función $\hat{R}(t)$ (la función da un salto cada vez que se llega a un nuevo tiempo de fallo, permaneciendo constante entre un tiempo de fallo y el instante previo al siguiente).

Ejemplo 4: Estimación MV para $F(t)$ (censura múltiple)

Se inicia un test de vida sobre dieciséis dispositivos idénticos, obteniéndose los siguientes tiempos de fallo y tiempos de censura (marcados con un asterisco): 31.7, 39.2, 57.2, 65.0*, 65.8, 70, 75*, 75.2*, 87.5*, 88.3*, 94.2*, 101.7*, 105.8, 109.2*, 110, 130*. Se desea obtener el estimador Kaplan-Meier para $F(t)$.

En este caso tenemos $k = 7$ tiempos de fallo distintos y 9 observaciones censuradas. Construimos la siguiente tabla para calcular el estimador K-P:

i	$[t_i, t_{i+1})$	n_i	d_i	$1 - \frac{d_i}{n_i}$	$\hat{R}(t)$	$\hat{F}(t)$
0	[0, 31.7)	16	--	--	1	0
1	[31.7, 39.2)	16	1	0.9375	0.9375	0.0625
2	[39.2, 57.2)	15	1	0.9333	0.8750	0.1250
3	[57.2, 65.8)	14	1	0.9286	0.8125	0.1875
4	[65.8, 70)	12	1	0.9167	0.7448	0.2552
5	[70, 105.8)	11	1	0.9091	0.6771	0.3229
6	[105.8, 110)	4	1	0.7500	0.5078	0.4922
7	[110, inf.)	2	1	0.5000	0.2539	0.7461

3.2 Estimación por intervalos

Usando un razonamiento análogo al realizado para el caso de observaciones completas, y utilizando un resultado conocido como Fórmula de Greenwood (que proporciona el estimador MV para la varianza de $\hat{R}(t)$), es posible obtener la expresión de un intervalo de confianza aproximado para $\hat{R}(t)$ (o, alternativamente, $\hat{F}(t)$). Los detalles técnicos de la construcción de dicho intervalo quedan, sin embargo, fuera de los objetivos del presente tema, por lo que en la práctica utilizaremos software estadístico para obtenerlo.

4 Análisis no paramétrico con MINITAB

Para realizar un análisis no paramétrico de los datos es posible utilizar las opciones que ofrece MINITAB a tal efecto (**Reliability/Survival > Distribution Analysis > Nonparametric Distribution Analysis...**). Estas opciones incluyen varios estimadores puntuales y por intervalos, entre los que se incluye el de Kaplan-Meier.

Al realizar un análisis no paramétrico de los datos, se le deben indicar al programa los **inputs** siguientes:

- Columna que contiene las observaciones
- Tipo de censura: a derecha o arbitraria (de cualquier tipo)
- Método estadístico (Kaplan-Meier o Actuarial) que se desea emplear para realizar las estimaciones de la función de fiabilidad
- Nivel de confianza para la estimación, por intervalos, de la función de fiabilidad
- Otros *inputs* opcionales (gráficos de supervivencia, etc.)

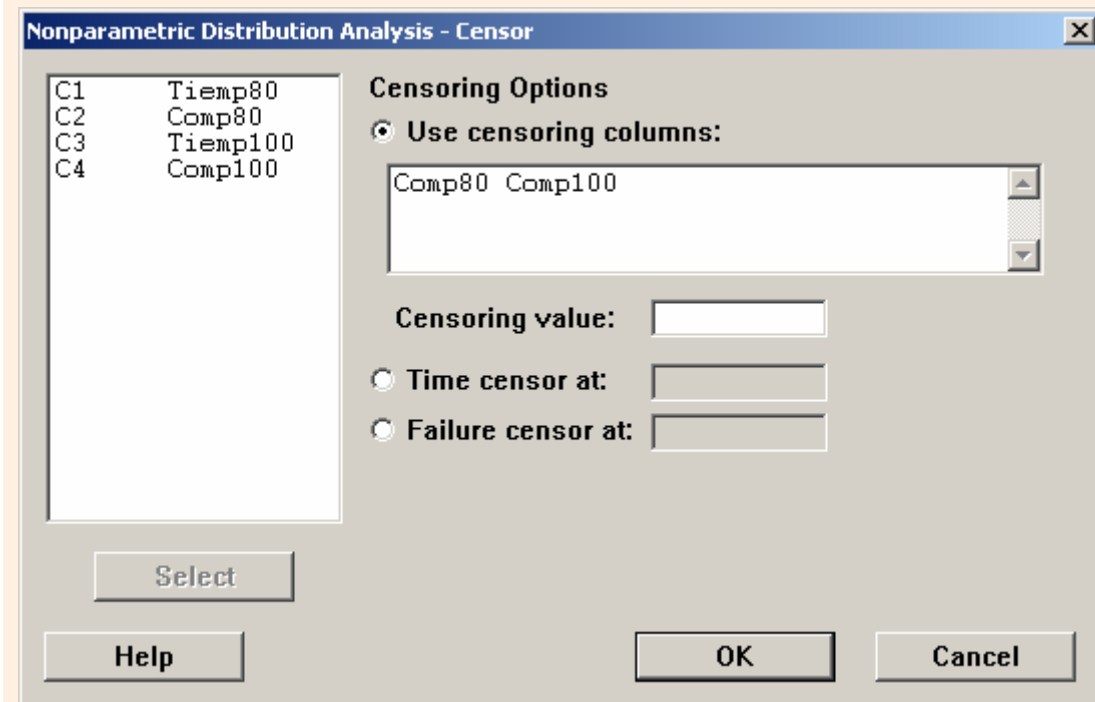
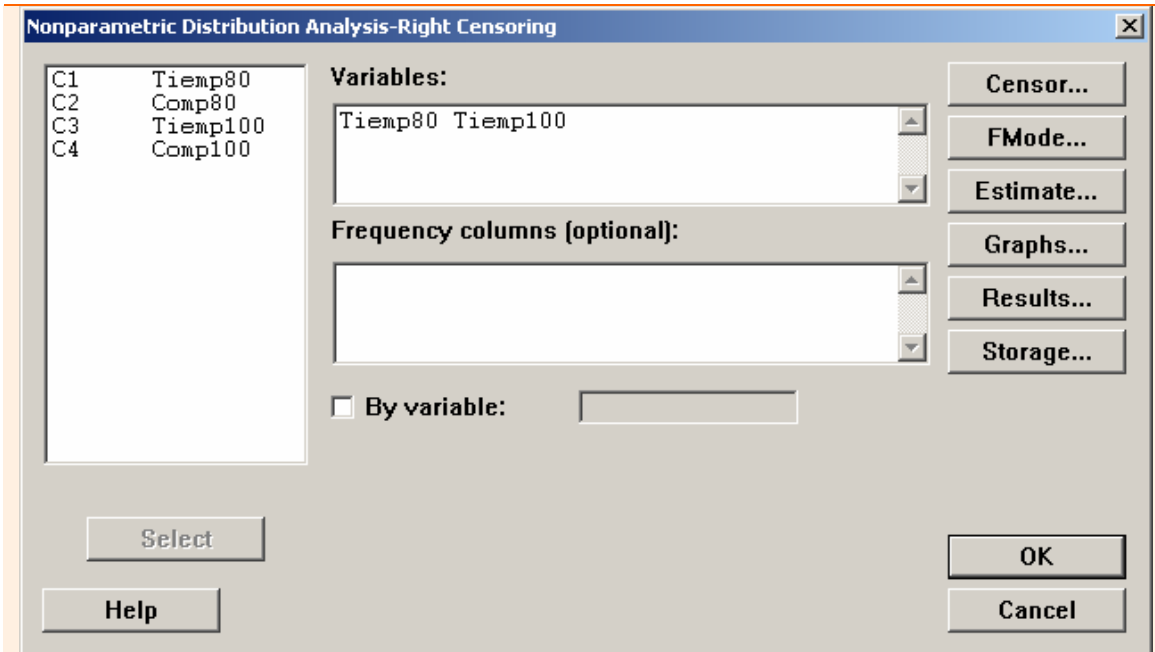
Por su parte, el programa ofrece los **outputs** siguientes:

- Información sobre las observaciones (número de observaciones completas y censuradas y tipo de censura)
- Estimaciones, puntuales y por intervalos para la función de fiabilidad
- Opcionalmente, los gráficos solicitados

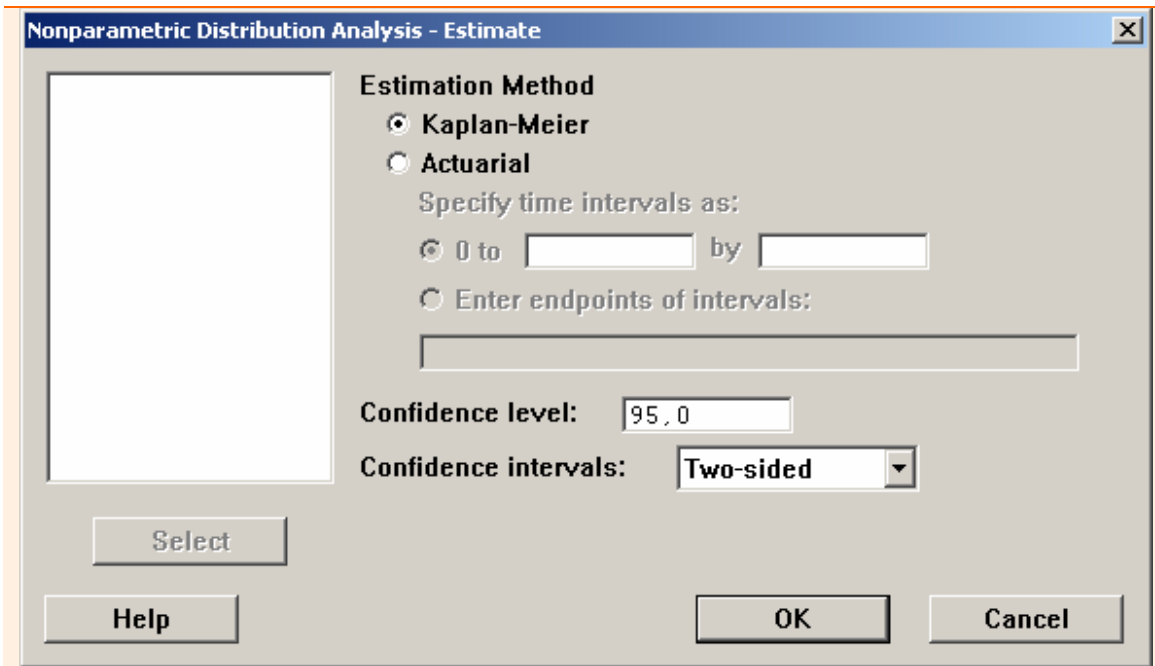
Ejemplo 5: Análisis no paramétrico con MINITAB (obs. con censura a derecha)

Continuando con el ejemplo de la compañía que fabrica cubiertas para motores, introducido en el capítulo 3 y analizado con técnicas paramétricas en el 4, vamos a realizar ahora el análisis no paramétrico de los datos con ayuda de MINITAB.

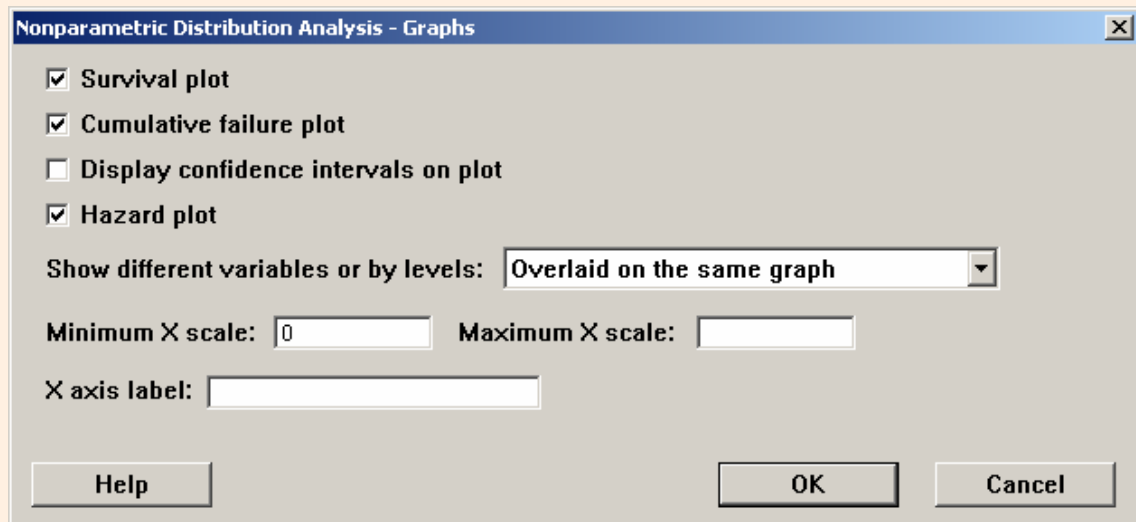
En primer lugar, puesto que se trata de observaciones censuradas a derecha, elegimos la opción **Reliability/Survival > Distribution Análisis (Right Censoring) > Nonparametric Distribution Analysis...** y especificamos las columnas que contienen los datos (tiempos de fallo observados para ambos grupos y calificadores de censura respectivos):



A continuación, especificamos el método de estimación que deseamos utilizar (Kaplan-Meier en este caso) y el nivel de confianza para la estimación por intervalos (usaremos un nivel del 95%):

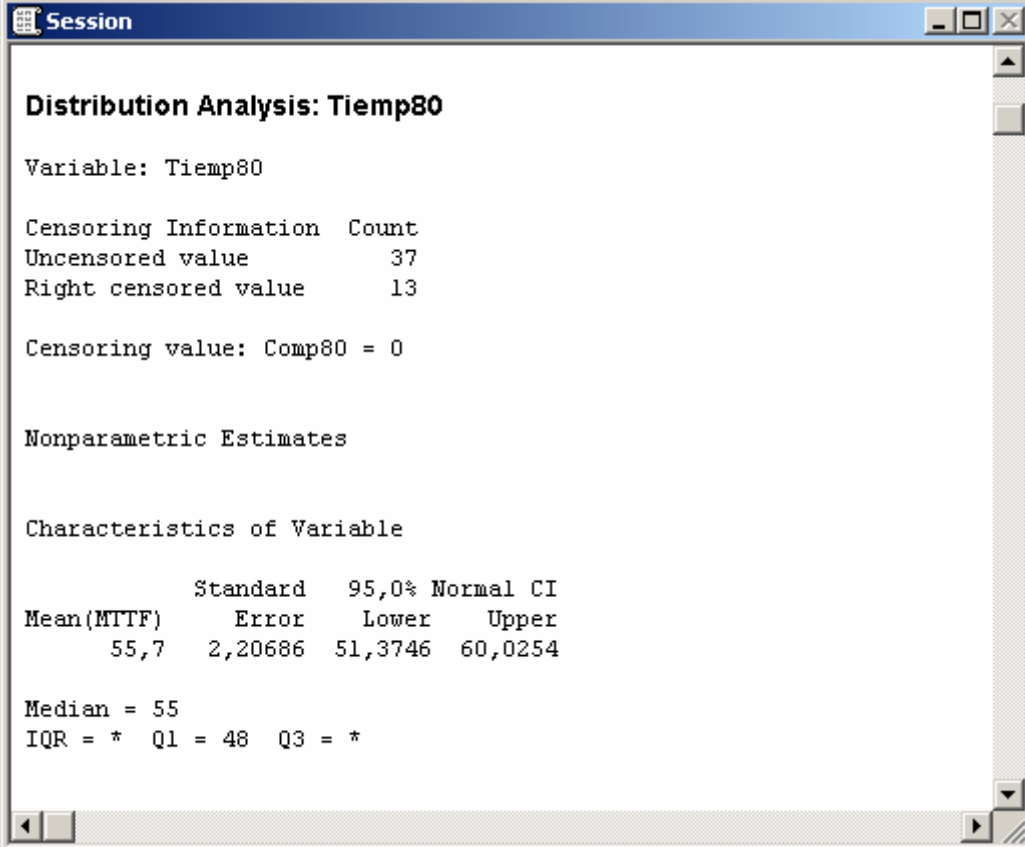


Finalmente, optaremos también por solicitar tres gráficos, el de la función de supervivencia, el de la función de distribución (que será el inverso del anterior) y el de la función tasa de fallo:



El programa ofrece un *output* para cada una de las variables consideradas (tiempo de fallo de los cubiertas a 80°C y tiempo de fallo de las cubiertas a 100°C). A continuación, analizaremos la información que se nos ofrece sobre las cubiertas a 80°C (la información que se ofrece sobre las cubiertas a 100°C se puede analizar de forma análoga):

En primer lugar, el programa proporciona el número de observaciones completas y censuradas, el valor que indica la existencia de censura (ésta viene indicada por un valor 0 en la columna Comp80), y algunos estadísticos descriptivos de la variable (tiempo medio hasta el fallo, mediana, etc.):



```
Session

Distribution Analysis: Tiemp80

Variable: Tiemp80

Censoring Information  Count
Uncensored value      37
Right censored value  13

Censoring value: Comp80 = 0

Nonparametric Estimates

Characteristics of Variable

                Standard  95,0% Normal CI
Mean(MTTF)      Error    Lower  Upper
                55,7    2,20686  51,3746  60,0254

Median = 55
IQR = *  Q1 = 48  Q3 = *
```

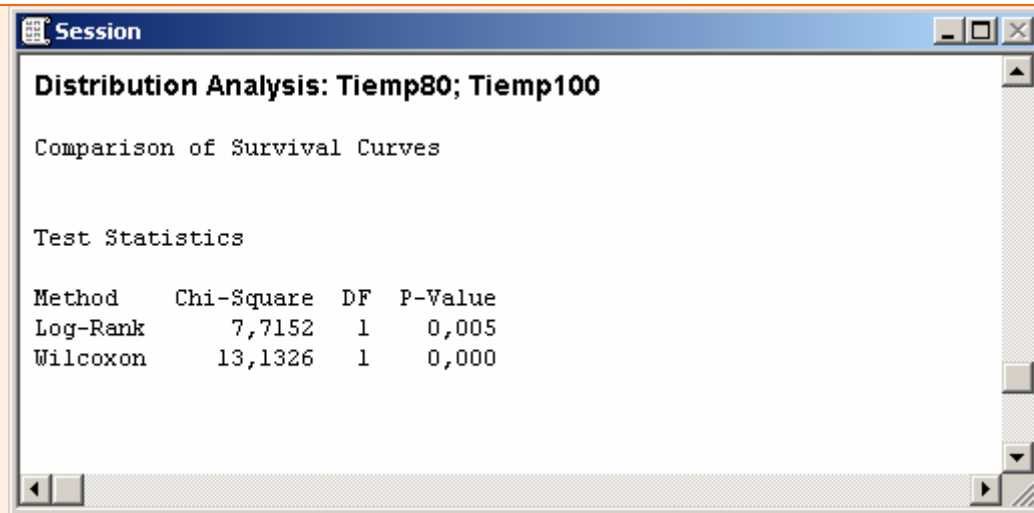
De los resultados se concluye que la mediana de la variable “tiempo de fallo” es, para una temperatura de 80° C, de 55 meses (i.e., el 50% de los dispositivos habrán fallado tras 55 meses de funcionamiento a 80°C). Mirando el *output* respectivo para la variable Tiemp100, se observa que, a 100°C, dicho tiempo es de 38 meses, lo que permite afirmar que el incremento de temperatura conlleva una disminución del tiempo mediano hasta el fallo de, aproximadamente, 17 meses.

En la siguiente parte del *output*, el programa ofrece la tabla de estimadores Kaplan-Meier, la cual incluye tanto estimaciones puntuales como por intervalo para la función de fiabilidad en distintos intervalos temporales. En la tabla se detallan, para cada intervalo temporal, el número de dispositivos en observación y el número de

dispositivos que han fallado. De esta tabla se deduce, por ejemplo, que a 80° C un 90% de las cubiertas seguirán funcionando correctamente tras 31 meses (mirando la tabla correspondiente para el caso en que la temperatura es de 100°C, se observa que dicho porcentaje de cubiertas sólo sobrevivirían unos 14 meses).

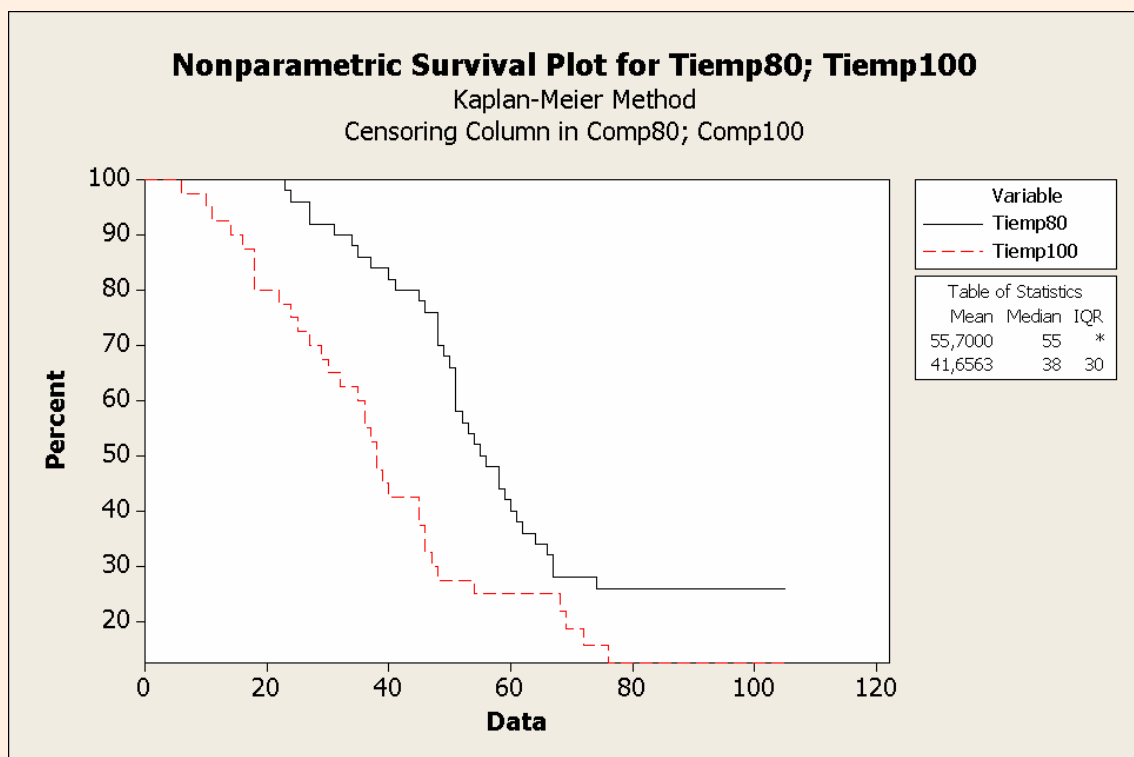
Time	Number at Risk	Number Failed	Survival Probability	Standard Error	95,0% Normal CI Lower	95,0% Normal CI Upper
23	50	1	0,980000	0,0197990	0,941195	1,000000
24	49	1	0,960000	0,0277128	0,905684	1,000000
27	48	2	0,920000	0,0383667	0,844803	0,99520
31	46	1	0,900000	0,0424264	0,816846	0,98315
34	45	1	0,880000	0,0459565	0,789927	0,97007
35	44	1	0,860000	0,0490714	0,763822	0,95618
37	43	1	0,840000	0,0518459	0,738384	0,94162
40	42	1	0,820000	0,0543323	0,713511	0,92649
41	41	1	0,800000	0,0565685	0,689128	0,91087
45	40	1	0,780000	0,0585833	0,665179	0,89482
46	39	1	0,760000	0,0603987	0,641621	0,87838
48	38	3	0,700000	0,0648074	0,572980	0,82702
49	35	1	0,680000	0,0659697	0,550702	0,80930
50	34	1	0,660000	0,0669925	0,528697	0,79130
51	33	4	0,580000	0,0697997	0,443195	0,71680
52	29	1	0,560000	0,0701997	0,422411	0,69759
53	28	1	0,540000	0,0704840	0,401854	0,67815
54	27	1	0,520000	0,0706541	0,381521	0,65848
55	26	1	0,500000	0,0707107	0,361410	0,63859
56	25	1	0,480000	0,0706541	0,341521	0,61848
58	24	2	0,440000	0,0701997	0,302411	0,57759
59	22	1	0,420000	0,0697997	0,283195	0,55680
60	21	1	0,400000	0,0692820	0,264210	0,53579
61	20	1	0,380000	0,0686440	0,245460	0,51454
62	19	1	0,360000	0,0678823	0,226953	0,49305
64	18	1	0,340000	0,0669925	0,208697	0,47130
66	17	1	0,320000	0,0659697	0,190702	0,44930
67	16	2	0,280000	0,0634980	0,155546	0,40445
74	13	1	0,258462	0,0621592	0,136632	0,38029

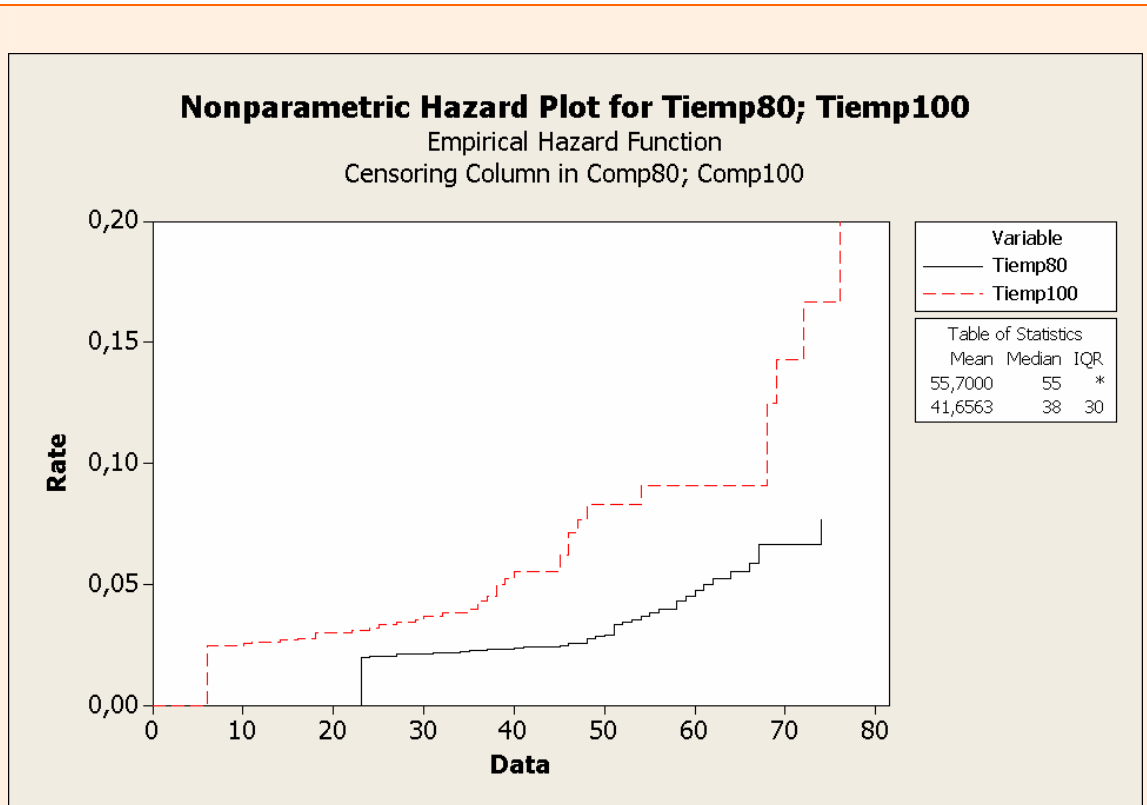
Finalmente, MINITAB realiza dos contrastes de hipótesis (Log-Rank y Wilcoxon) para contrastar la hipótesis nula de que ambas muestras (Tiemp80 y Tiemp100) son iguales:



En el *output* anterior se obtiene un p-valor significativo, tanto para el test Log-Rank como para el test de Wilcoxon (considerando $\alpha = 0,05$), por lo que se confirma la existencia de diferencias significativas entre los tiempos de fallo a 80°C y a 100°C.

Finalmente, obtenemos los gráficos solicitados, que permiten comparar visualmente el comportamiento de los dos grupos considerados (cubiertas a 80°C y cubiertas a 100°C):





En todos los gráficos se puede observar claramente como el factor temperatura afecta significativamente al tiempo de vida de las cubiertas (por ejemplo, mirando el gráfico de supervivencia se observa que la función de supervivencia desciende más rápidamente en el caso de las cubiertas a 100°C que en el caso de las cubiertas a 80°C).