Diagnosis y Crítica del modelo
-Ajuste de distribuciones con Statgraphics-

Ficheros de datos: TiempoaccesoWeb.sf3 ; AlumnosIndustriales.sf3

1. Introducción

El objetivo de esta práctica es asignar un modelo de probabilidad a un conjunto de datos, de forma que el modelo elegido pueda interpretarse como la población de la que proceden esos datos. A esta búsqueda de un modelo de probabilidad a partir de una muestra de datos se le denomina **ajuste de una distribución**. Para que un modelo de probabilidad pueda considerarse que es un modelo razonable para explicar los datos, han de realizarse pruebas estadísticas. La realización de estas pruebas se denomina diagnosis o crítica del modelo. Por tanto, diremos que un modelo tendrá un buen **ajuste** a nuestros datos si supera con éxito la **diagnosis**.
La forma habitual para hacer ajuste de modelos es la siguiente. A partir del análisis de la muestra se comparará su distribución con la de algún modelo conocido (Normal, Poisson, Exponencial, etc) . Para evaluar si un modelo tiene un buen ajuste realizaremos el test de la Chi cuadrado.

Se utilizarán dos ficheros: TiempoaccesoWeb.sf3 y AlumnosIndustriales.sf3. Empezaremos analizando la variable Ordenador_Uni del fichero TiempoaccesoWeb.sf3. Esta variable tiene 55 medidas del tiempo, en segundos, que se tarda en acceder a la página Web de la Universidad desde un ordenador de su biblioteca. Veremos, como a partir de esta muestra, podemos encontrar un modelo de probabilidad que se ajuste a esos datos y que sirva como modelo poblacional del tiempo que tardamos cada vez que abrimos la página Web de la Universidad con ordenadores de su biblioteca. En segundo lugar analizaremos la variable Tiempo del fichero AlumnosIndustriales.sf3. Esta variable contiene el tiempo que tardan unos estudiantes en llegar a la Universidad.

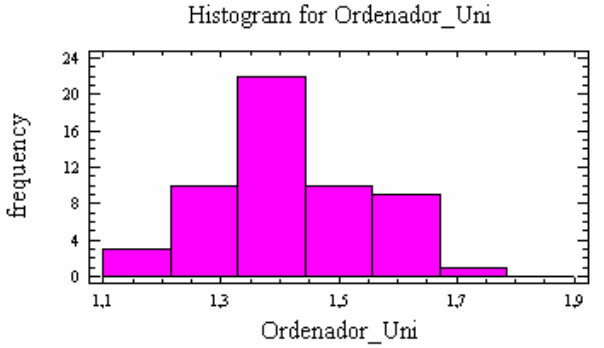
2. Ajuste del modelo. Variable Ordenador_Uni

2.1 Análisis descriptivo de los datos

Lo primero que haremos, será un estudio descriptivo de los datos (Medidas características, histograma). Así podemos hacernos una idea de la distribución de los datos.

Nos vamos a DESCRIBE/NUMERIC DATA/ONE VARIABLE ANALYSIS. Hacemos click en Summary Statistics y Frecuency Histogram. En Summary Statistics seleccionamos las medidas características más habituales (en Pane Options –botón derecho del ratón-)

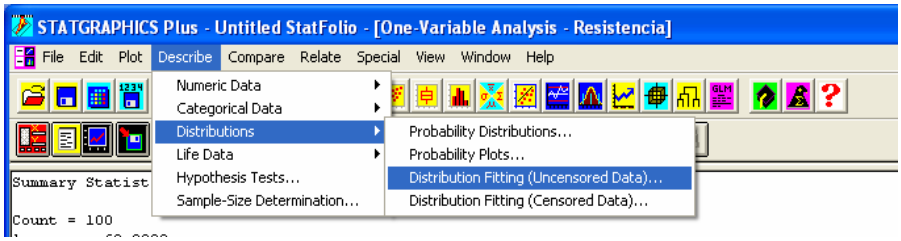
```
Summary Statistics for Ordenador_Uni

Count = 55
Average = 1,42482
Median = 1,416
Variance = 0,0156342
Standard deviation = 0,125037
Minimum = 1,158
Maximum = 1,678
Range = 0,52
Skewness = 0,0853185
Kurtosis = -0,292129
```



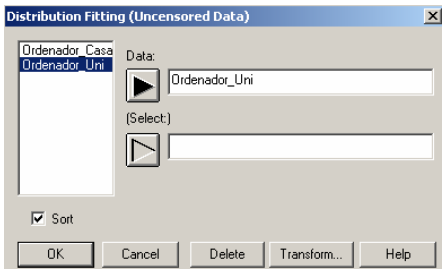
Vemos que el histograma se parece a una Normal. Es unimodal y bastante simétrico (Skewness=0.08) aunque menos apuntado que la normal (Kurtosis=-0.29). Esto nos conduce a pensar que una normal podría proporcionar un ajuste suficientemente bueno a estos datos y ser utilizada para explicar las distribuciones de tiempos de acceso.

2.2 Diagnosis del modelo elegido

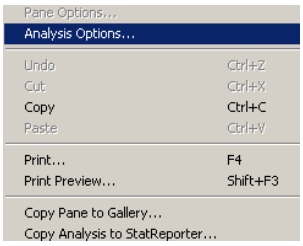
Para evaluar el ajuste de un modelo vamos a DESCRIBE/DISTRIBUTION FITTING/UNCESORED DATA



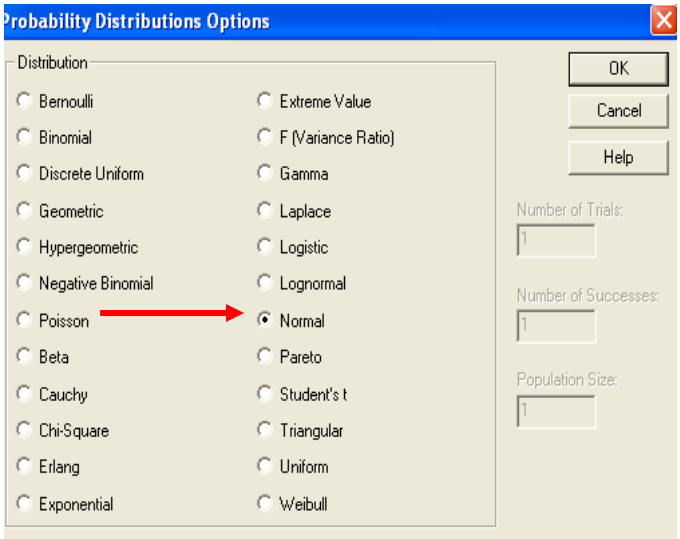
Se abre entonces la ventana para introducir la variable a la que queremos ajustar una distribución. Seleccionamos Ordenador_Uni.



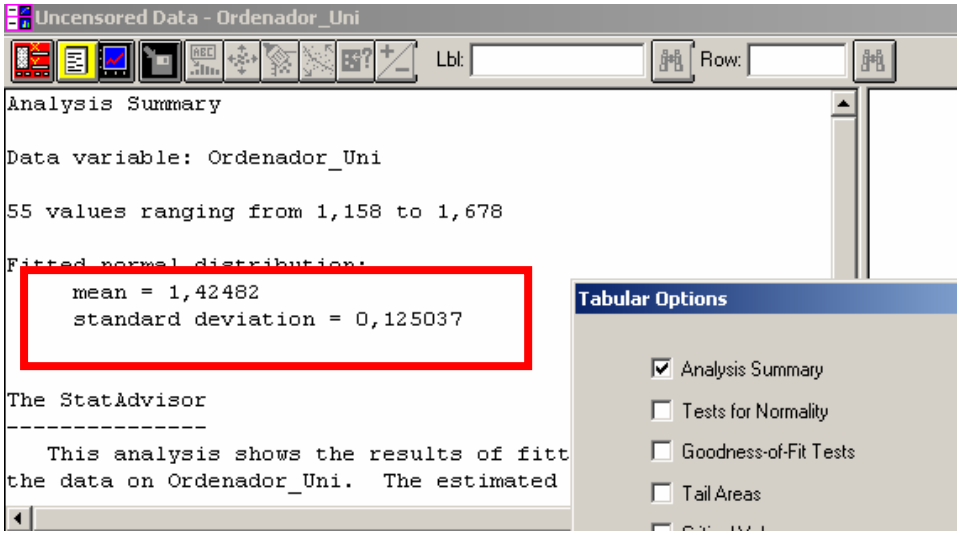
Seleccionamos ahora el modelo de distribución escogido. Para eso nos colocamos en cualquier ventana y pulsamos el botón derecho, y pulsamos Analysis Options.



Aparece entonces la ventana para seleccionar el modelo de probabilidad. Seleccionamos la Normal (es la que aparece seleccionada por defecto)



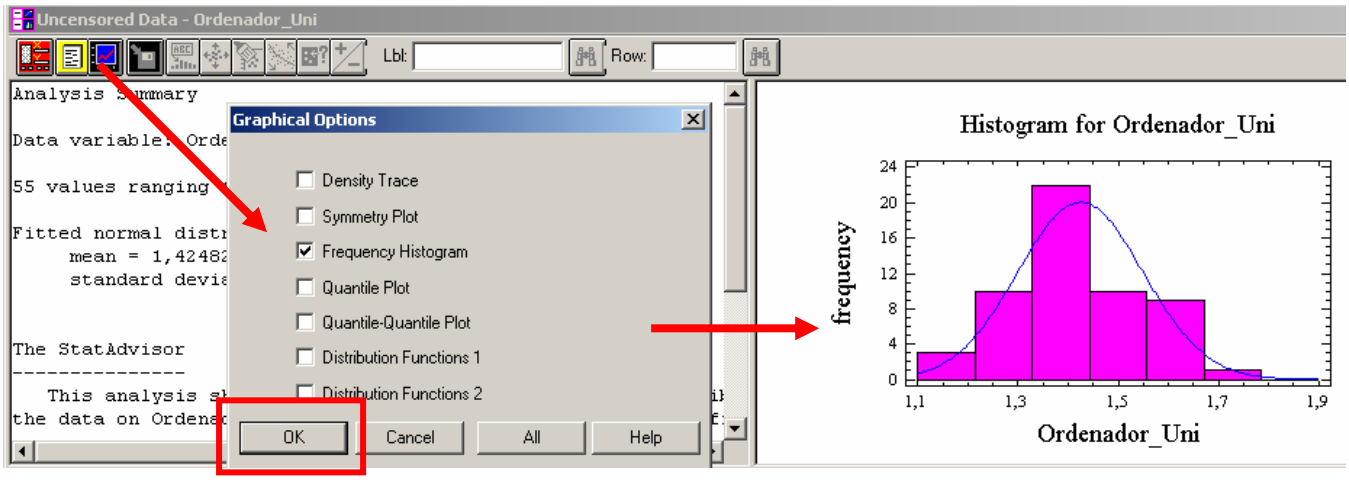
La estimación de los parámetros del modelo la enconrtamos en las Opciones Tabulares , Analysis Summary.



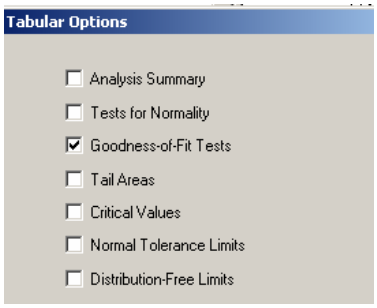
Los parámetros de la normal estimados son entonces $\hat{\mu} = 1.42; \hat{\sigma}^2 = 0.125^2$ que corresponden con los obtenidos anteriormente al describir las variables. El modelo estimado es por tanto

$$X \sim N(1.42, 0.125^2)$$

Este ejercicio de estimación no nos informa de si la normal es o no un modelo apropiado. Para hacer la diagnosis del modelo seleccionamos primeramente Frequency Histogram entre las Opciones Gráficas. El resultado obtenido es



Este gráfico nos presenta nuestro histograma junto con la función de densidad del modelo teórico. Cuánto más se aproxime la curva a nuestros datos, mejor será el ajuste. Esta figura es muy útil pues nos permite visualizar el ajuste. Finalmente hacemos el Test de bondad de ajuste de la Chi-cuadrado. Vamos a Tabular Options y seleccionamos GOODNESS-OF-FIT-TEST (Test de bondad del ajuste)



El resultado es el siguiente:

Lbl: Row:

Goodness-of-Fit Tests for Ordenador_Uni

Chi-Square Test

Lower Limit	Upper Limit	Observed Frequency	Expected Frequency	Chi-Square
at or below	1,26458	6	5,50	0,05
1,26458	1,31958	6	5,50	0,05
1,31958	1,35925	3	5,50	1,14
1,35925	1,39314	7	5,50	0,41
1,39314	1,42482	7	5,50	0,41
1,42482	1,4565	8	5,50	1,14
1,4565	1,49039	4	5,50	0,41
1,49039	1,53005	2	5,50	2,23
1,53005	1,58506	6	5,50	0,05
above	1,58506	6	5,50	0,05

Chi-Square = 5,90881 with 7 d.f. P-Value = 0,550436

Estimated Kolmogorov statistic DPLUS = 0,0880407
Estimated Kolmogorov statistic DMINUS = 0,0872779
Estimated overall statistic DN = 0,0880407
Approximate P-Value = 0,787465

El resultado del Test de la Chi-cuadrado se resume en los tres valores siguientes:

- **Chi-square** = **5.9088**, que representa el valor del estadístico calculado en el test

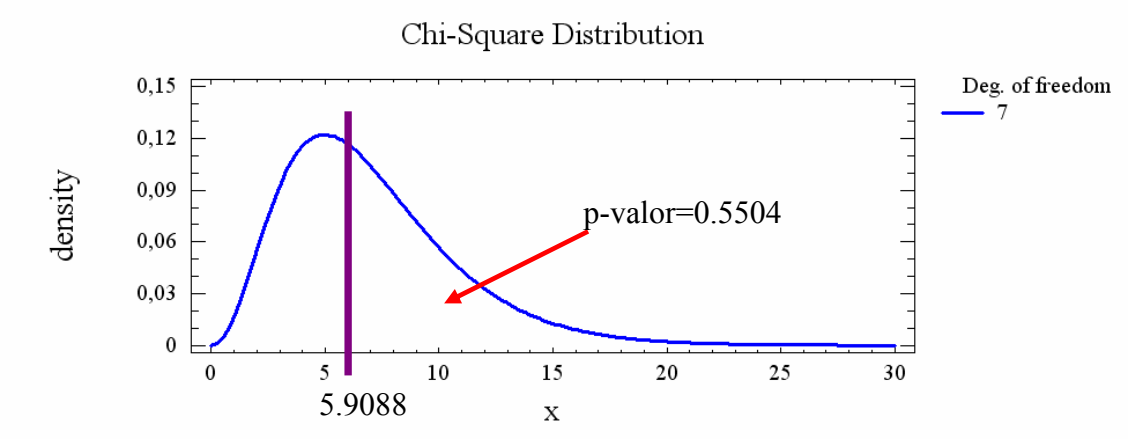
$$\chi^2 = \sum \frac{(\text{Frec.Obs.} - \text{Frec.Esp.})^2}{\text{Frec.Esp.}}$$

  Este estadístico resume la discrepancia entre el histograma y la curva de la normal. Cuanto mayor sea este valor, peor es el ajuste de nuestros datos al modelo elegido.

- **d.f** (degrees of freedom)= 7, que son grados de libertad de la distribución Chi-Cuadrado que se usa de referencia para valorar el ajuste de la distribución. Los grados de libertad se calculan como df= k- υ-1, donde:
  k= número de intervalos, en este caso 10
  υ= número de parámetros del modelo escogido, en este caso 2 (media y varianza)

- **p_value** =0.5504 (p valor). Probabilidad que queda a la derecha el valor del estadístico calculado en la distribución de referencia. En nuestro caso, es el área que queda a la derecha del valor 5.9088 en la distribución χ^2_{k-v-1}.

La teoría estadística nos dice que cuanto peor es el ajuste del modelo elegido, el estadístico χ^2 dará un valor mayor, y que la referencia para evaluar cómo de grande es ese estadístico en cada caso es la distribución χ^2_{k-v-1}. Una forma sencilla de valorar la bondad del ajuste es calcular el área que queda a la derecha del valor del estadístico χ^2 en la distribución χ^2_{k-v-1}. Ese área es precisamente el p-valor. La figura siguiente ilustra este resultado para nuestro caso.



Si el p-valor es inferior a 0.05 se considera que el estadístico está ya en zonas de muy poca probabilidad, y por tanto concluimos que el ajuste no es satisfactorio. Por el contrario, si el p-valor es mayor de 0.05 consideramos que el ajuste es suficientemente bueno, y que el modelo elegido puede usarse como modelo para la población. En nuestro caso, el p-valor es 0.55 por lo que concluimos que la normal es un modelo razonable para explicar la distribución de nuestros datos.

El Statgraphics realiza otros tests de bondad de ajuste. Los resultados de todos ellos pueden interpretarse también a través de sus p-valores de la misma forma que con el tests de la chi-cuadrado. Por ejemplo, puede observar que en el test de Kolmogorov-Smirnov el p-valor es 0.7674, también mayor de 0.05.

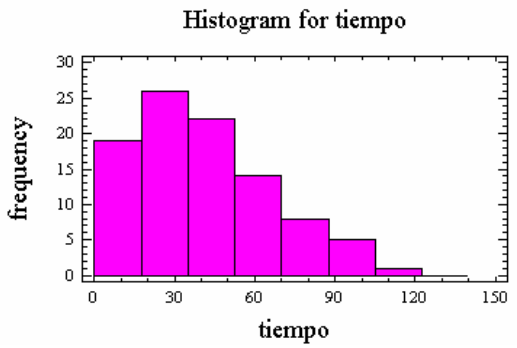
3. Ajuste de un modelo para la variable ¨Tiempo¨

Vamos a repetir el estudio anterior, con la variable Tiempo del fichero AlumnosIndustriales.sf3. Esta variable es el tiempo que tardan unos estudiantes en llegar a la Universidad. El tamaño de la muestra es 95.

3.1 Análisis descriptivo de los datos

Después de cargar el fichero AlumnosIndustriales.sf3 en Statgraphics procedemos a hacer un resumen estadístico de nuestra variable. La descripción estadística de la variable se realiza como antes en Describve/Numeric Data/One Variable Analysis. El resultado se muestra en la siguiente figura. En la construcción del histograma se ha puesto que el límite inferior sea 0, ya que se trata de valores de tiempo que son no negativos.

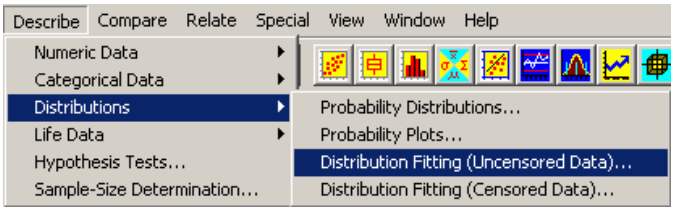
```
Summary Statistics for tiempo

Count = 95
Average = 41,4211
Median = 40,0
Variance = 611,97
Standard deviation = 24,738
Minimum = 1,0
Maximum = 120,0
Range = 119,0
Skewness = 0,651076
Stnd. skewness = 2,5907
Kurtosis = 0,0915265
Stnd. kurtosis = 0,182097
```

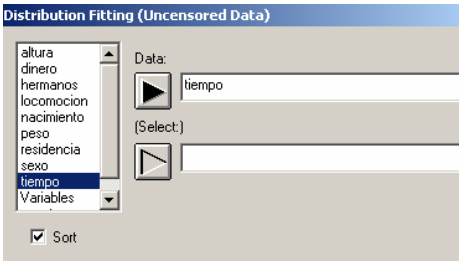


Los datos son unimodales, con asimetría positiva. La zona de la moda tiene un apuntamiento en forma de campana. Tenemos dos opciones bara asignar un modelo de probabilidad a esta variable. En primer lugar probaremos ajustar un modelo con asimetría positiva como la distribución Weibull, o una distribución lognormal. En segundo lugar, intentaremos ajustar una normal a una transfromación de los datos que corrijan su asimetría. Por ejemplo a la raíz cuadrada (ajustar una normal al logaritmo de una variable es lo mismo que ajustar una lognormal a la variable original).

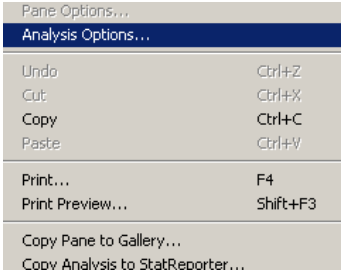
3.2 Ajuste de una Weibull

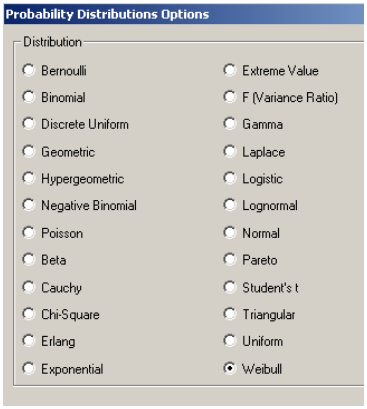
Como en el ejemplo anterior, vamos a Describe/Distributions/Distribution Fitting (Uncensored data), y alli seleccionamos la variable Tiempo.





En Analysis Options (botón derecho del ratón) accedemos a la ventana para elegir la distribución. Elegimos Weibull.





El Statgraphics nos proporciona entonces las estimaciones de los parámetros de esta distribución

```
Analysis Summary

Data variable: tiempo

95 values ranging from 1,0 to 120,0

Fitted Weibull distribution:
     shape = 1,70898
     scale = 46,3503
```

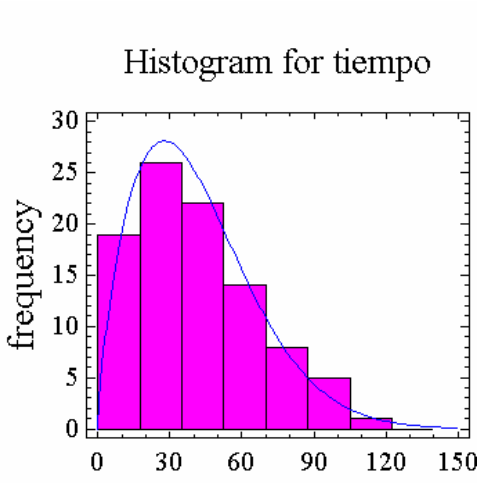
En Tabular Options seleccionamos Goodness-of-Fit Test, y en Graphical Options seleccionamos Frequency Histogram. Obtenemos el siguiente resultado (de nuevo poniendo 0 como origen del histograma)

Goodness-of-Fit Tests for tiempo

Chi-Square Test

	Lower Limit	Upper Limit	Observed Frequency	Expected Frequency
at or below		11,1058	9	7,92
	11,1058	17,1213	10	7,92
	17,1213	22,3583	8	7,92
	22,3583	27,3305	3	7,92
	27,3305	32,2843	10	7,92
	32,2843	37,4035	5	7,92
	37,4035	42,88	6	7,92
	42,88	48,9725	11	7,92
	48,9725	56,1123	7	7,92
	56,1123	65,2015	11	7,92
	65,2015	78,9524	8	7,92
above	78,9524		7	7,92

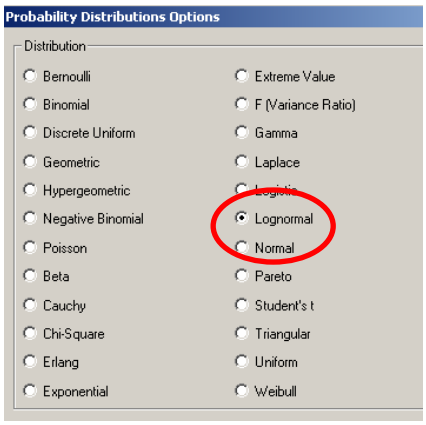
Chi-Square = 8,45263 with 9 d.f. P-Value = 0,489252



Tanto el histograma con la Weibull superpuesta como el p-valor del test de la Chi-cuadrado nos muestran que el ajuste es bastante satisfactorio. Por tanto podemos utilizar la distribución Weibull para modelizar los tiempos de llegada a la universidad.

3.2 Ajuste de una Lognormal

Pulsando el botón derecho del ratón, seleccionamos Analysis Options y elegimos ahora la distribución Lognormal



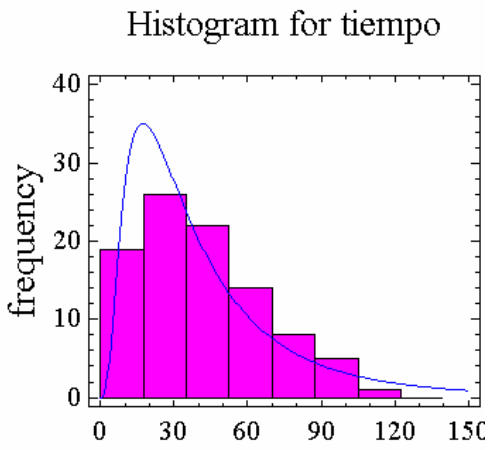
obteniéndose los siguientes resultados.

Goodness-of-Fit Tests for tiempo

Chi-Square Test

Lower Limit	Upper Limit	Observed Frequency	Expected Frequency
at or below	10,9447	9	7,92
10,9447	15,2151	10	7,92
15,2151	19,1922	0	7,92
19,1922	23,2835	8	7,92
23,2835	27,7263	3	7,92
27,7263	32,7597	10	7,92
32,7597	38,7067	5	7,92
38,7067	46,0925	17	7,92
46,0925	55,9181	7	7,92
55,9181	70,5347	12	7,92
70,5347	98,0565	12	7,92
above	98,0565	2	7,92

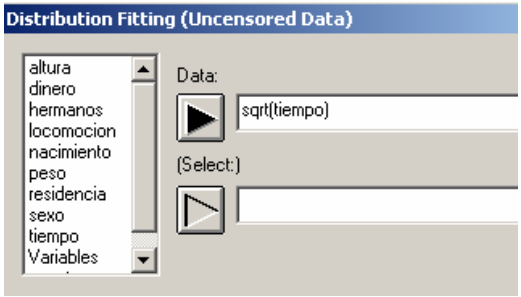
Chi-Square = 32,4528 with 9 d.f. P-Value = 0,00016623



Ahora el ajuste no es bueno. El p-valor del test de la chi-cuadrado es ya muy bajo. El histograma nos muestra el motivo, y es que la lognormal es más apuntada que nuestros datos. La lognormal no es un modelo adecuado para esta variable.

3.3 Ajuste de una Normal a una transformación

La variable tiempo es asimétrica positiva, sin embargo su raíz cuadrada es ya bastante simétrica. Si ajustamos una Normal a la raíz cuadrada obtenemos los siguientes resultados

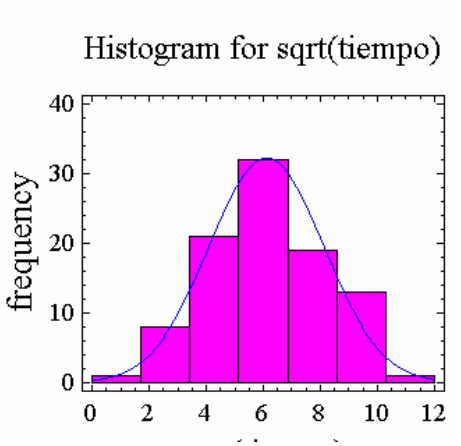


Goodness-of-Fit Tests for sqrt(tiempo)

Chi-Square Test

Lower Limit	Upper Limit	Observed Frequency	Expected Frequency	Chi-Square
at or below	3,3348	9	7,92	0,15
3,3348	4,1708	10	7,92	0,55
4,1708	4,76008	8	7,92	0,00
4,76008	5,25045	3	7,92	3,05
5,25045	5,69362	10	7,92	0,55
5,69362	6,11693	5	7,92	1,07
6,11693	6,54025	6	7,92	0,46
6,54025	6,98341	11	7,92	1,20
6,98341	7,47378	7	7,92	0,11
7,47378	8,06307	11	7,92	1,20
8,06307	8,89906	8	7,92	0,00
above 8,89906		7	7,92	0,11

Chi-Square = 8,45304 with 9 d.f. P-Value = 0,489212



que presenta un ajuste casi tan bueno como el de la Weibull. En Tabular Options seleccionamos Analysis Sumary para ver los parámetros estimados para este modelo.

Analysis Summary

Data variable: sqrt(tiempo)

95 values ranging from 1,0 to 10,9545

Fitted normal distribution:
mean = 6,11693
standard deviation = 2,01167

4. Ejemplo de aplicación del modelo ajustado

El disponer de un modelo que sea adecuado para representar a la población de la que hemos obtenido los datos observados es muy útil. Permite, entre otras cosas, calcular probabilidades de sucesos de forma más precisa que utilizando la frecuencia de aparición de dicho suceso en la muestra observada.

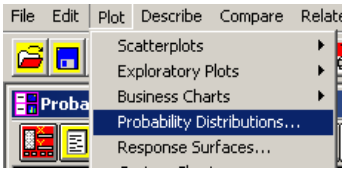
En esta sección vamos a calcular la probabilidad de que un alumno viva a más de una hora de la Universidad. Lo podemos hacer tanto con la distribución Weibull como con la Normal aplicada a la raíz cuadrada de la variable. Ambos modelos no darán los mismos resultados, pero esperaremos que no sean muy diferentes.

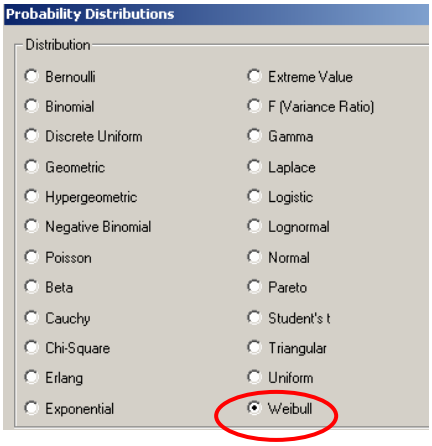
4.1 Cálculo con la Weibull

Como hemos visto anteriormente, la Weibull que hemos ajustado a los datos ha producido las siguientes estimaciones de los parámetros:

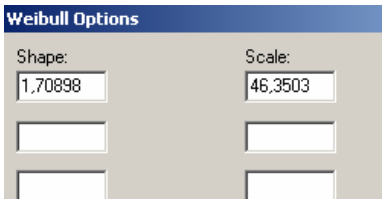
$$\text{shape} = \hat{\beta} = 1.70898$$
$$\text{scale} = \hat{\alpha} = 46.3503$$

Vamos entonces a calcular la probabilidad deseada para esa distribución (ver guión sobre modelos de distribución). En Statgraphics no vamos a Plot/Probability Distributions y allí seleccionamos la Weibull.

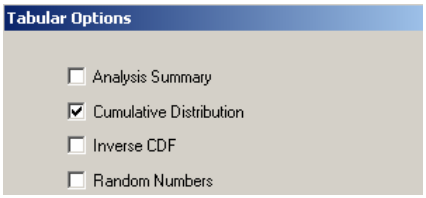




Una vez seleccionada la Weibull introducimos los parámetros que hemos estimado pulsando el botón derecho del ratón y Analysis Options



En Tabular Options seleccionamos la Función de distribución (Cumulative Distribution)



Ahora seleccionamos Pane Options (botón derecho del ratón) y allí ponemos los 60 minutos, que es el suceso en el que estamos interesados. El resultado es el siguiente

```
Cumulative Distribution
-----------------------

Distribution: Weibull

              Lower Tail Area (<)
Variable      Dist. 1       Dist. 2
60            0,788693

              Probability Density
Variable      Dist. 1       Dist. 2
60            0,00935567

              Upper Tail Area (>)
Variable      Dist. 1       Dist. 2
60            0,211307
```

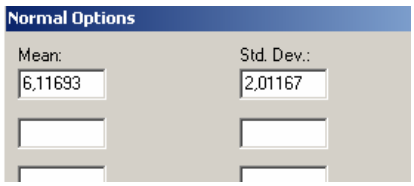
Por tanto podemos concluir que la probabilidad de que un alumno viva a más de una hora de la universidad es del 21,1%.

4.2 Cálculo con la Normal y la variable transformada

Como vimos antes, la raíz cuadrada del tiempo se ajusta muy bien a la Normal. Para calcular probabilidades debemos hacerlas sobre la variable transformada. Por tanto la probabilidad de tardar más de 60 minutos será equivalente a tardar más de $\sqrt{60} = 7.746$ en unidades transformadas. Vimos más arriba que la distribución normal ajustada a los datos tiene los siguientes parámetros estimados

$$\hat{\mu} = 6.11693$$
$$\hat{\sigma} = 2.01167$$

Calculamos entonces la probabilidad deseada para esa distribución. Vamos a Plot/Probability Distributions y allí seleccionamos la Normal. Introducimos las estimaciones de los parámetros.



y ahora calculamos la probabilidad deseada P(X>7.746), obteniéndose el siguiente resultado

```
Cumulative Distribution
-----------------------

Distribution: Normal

                Lower Tail Area (<)
Variable        Dist. 1         Dist. 2
7,746           0,790976

                Probability Density
Variable        Dist. 1         Dist. 2
7,746           0,142873

                Upper Tail Area (>)
Variable        Dist. 1         Dist. 2
7,746           0,209024
```

Por tanto, con este otro modelo, la probabilidad de que acudan alumnos que vivan a más de una hora de distancia es de 20,9% y que, como era de esperar, es casi lo mismo que con el otro modelo.