

---

# Diagnos y Crítica del modelo

## -Ajuste de distribuciones con Statgraphics Centurion-

---

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 **diagnos** o **crítica del modelo** (del modelo de distribución que ajustamos con los datos). Por tanto, diremos que un modelo tendrá un buen **ajuste** a nuestros datos si supera con éxito la **diagnos**.

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.

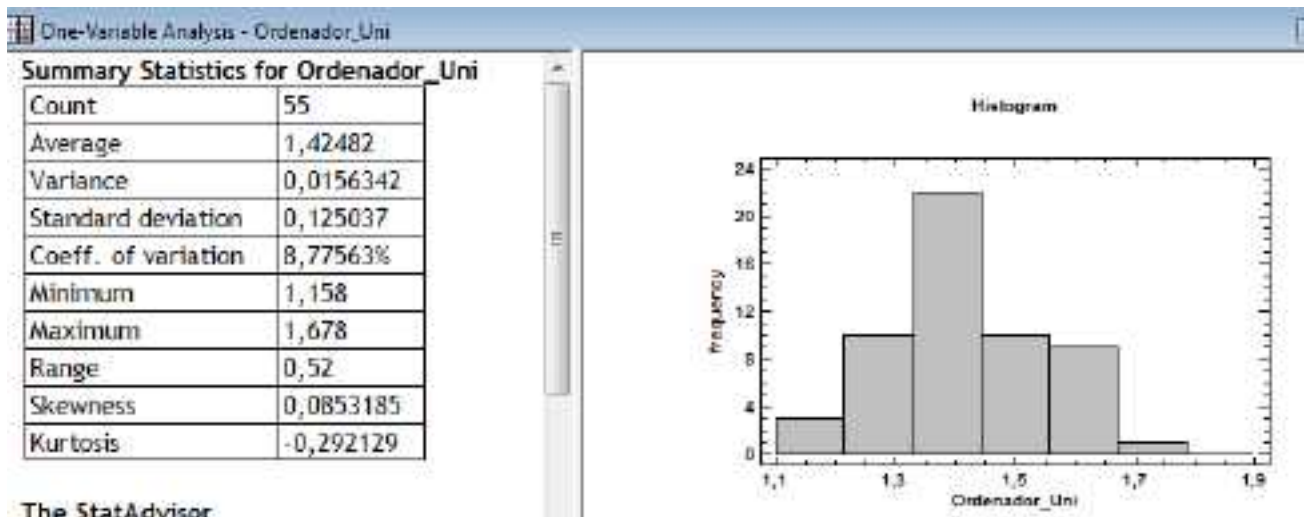
Se utilizarán dos ficheros: [TiempoaccesoWeb.sf3](#) y [AlumnosIndustriales.sf3](#). Empezaremos analizando la variable Ordenador\_Uni del fichero [TiempoaccesoWeb.sf3](#). Esta variable tiene 55 medidas del tiempo, en segundos, que se tarda en acceder a la página Web de la Universidad desde un ordenador 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. De los resultados que nos ofrece el Statgraphics, seleccionamos Summary Statistics y Frequency Histogram. En Summary Statistics seleccionamos las medidas características más habituales (en Pane Options -botón derecho del ratón-)



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 -en una normal es 0-). Esto nos conduce a plantearnos si una normal podría proporcionar un ajuste suficientemente bueno a estos datos y ser utilizada para modelizar las distribución del tiempo de acceso.

## 2.2 Diagnósis del modelo elegido

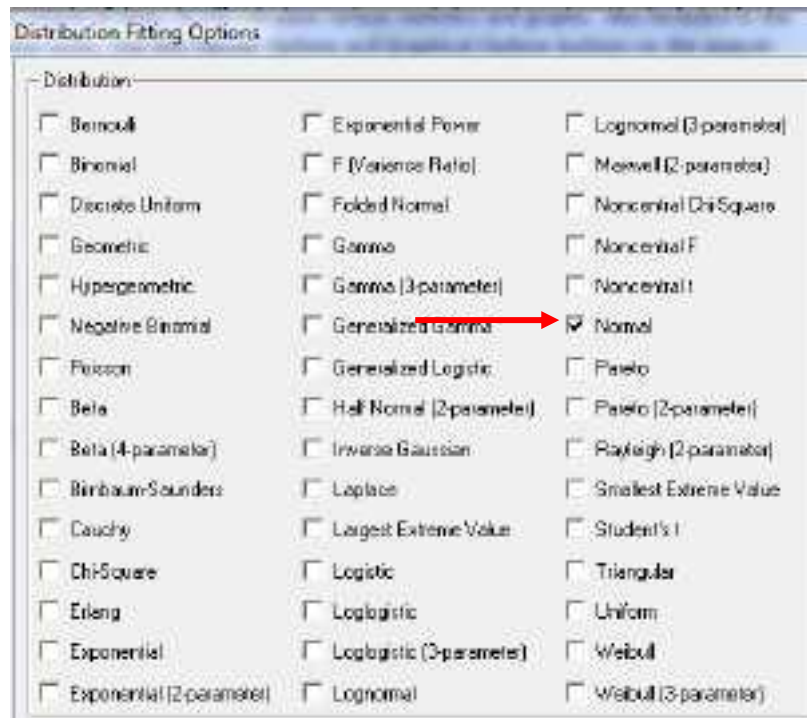
Para evaluar el ajuste de un modelo vamos a DESCRIBE/DISTRIBUTION FITTING/UNCESORED DATA




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

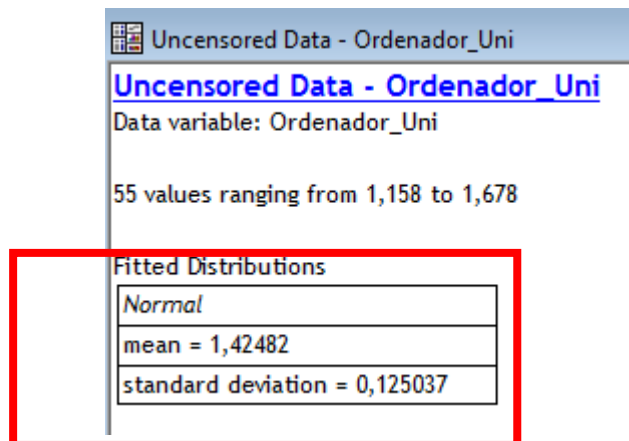


Aparece entonces la ventana para seleccionar el modelo de probabilidad. Seleccionamos la Normal (es la que aparece seleccionada por defecto)



Esta ventana con los modelos está también accesible pulsando el botón derecho del ratón y seleccionando Analysis Options.

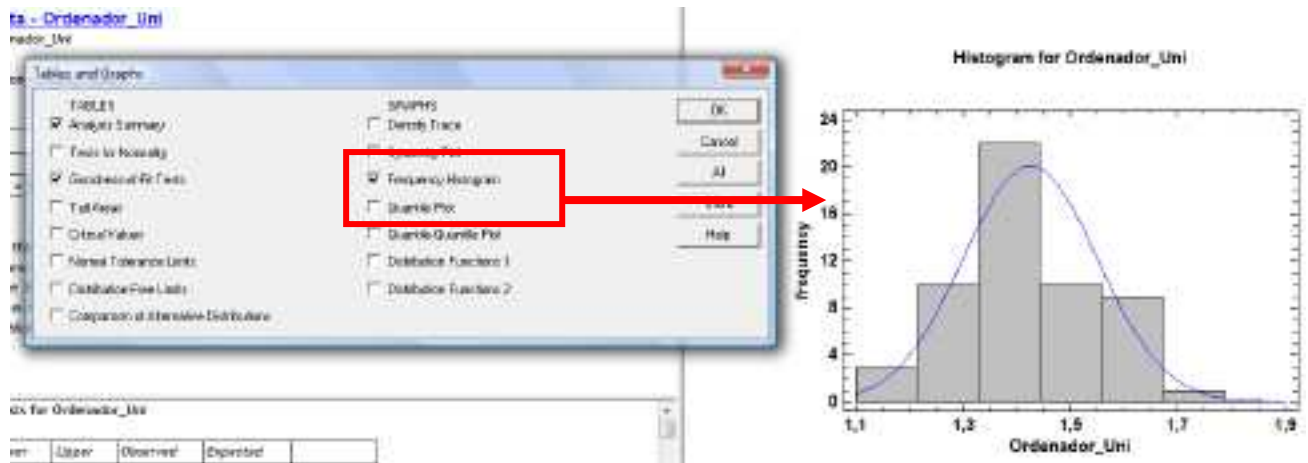
La estimación de los parámetros del modelo la encontramos en Tables and Graphs , Analysis Summary.



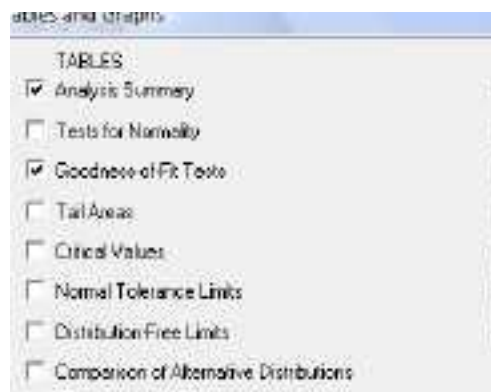
Los parámetros de la normal estimados son entonces  $\hat{\mu} = 1.42$ ;  $\hat{\sigma}^2 = 0.125^2$  que corresponden con los obtenidos anteriormente al describir las variables. El modelo estimado es por tanto

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

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 de Tables and Graphs. El resultado obtenido es



Este gráfico nos presenta nuestro histograma junto con la función de densidad del modelo teórico con los parámetros estimados con los datos. Cuanto más se aproxime la curva a nuestros datos, mejor será el ajuste. Esta figura es muy útil pues nos permite visualizar el ajuste, y entender mejor el resultado del test chi-cuadrado. Finalmente hacemos el Test de bondad de ajuste de la Chi-cuadrado. Vamos a Tables and Graphs y seleccionamos GOODNESS-OF-FIT-TEST (Test de bondad de ajuste)



El resultado que proporciona el Statgraphics en este punto es el siguiente:

#### Chi-Square Test

	<i>Lower</i>	<i>Upper</i>	<i>Observed</i>	<i>Expected</i>	
	<i>Limit</i>	<i>Limit</i>	<i>Frequency</i>	<i>Frequency</i>	<i>Chi-Square</i>
at or below		5,70019	0	2,89	2,89
	5,70019	5,77242	7	2,89	5,82
	5,77242	5,82132	3	2,89	0,00
	5,82132	5,86032	4	2,89	0,42
	5,86032	5,89389	1	2,89	1,24
	5,89389	5,92417	2	2,89	0,28
	5,92417	5,95234	1	2,89	1,24
	5,95234	5,97922	4	2,89	0,42
	5,97922	6,00538	1	2,89	1,24
	6,00538	6,03131	8	2,89	9,00
	6,03131	6,05747	3	2,89	0,00
	6,05747	6,08435	2	2,89	0,28
	6,08435	6,11252	2	2,89	0,28
	6,11252	6,1428	4	2,89	0,42
	6,1428	6,17638	3	2,89	0,00
	6,17638	6,21537	2	2,89	0,28
	6,21537	6,26427	3	2,89	0,00
	6,26427	6,3365	0	2,89	2,89
above	6,3365		5	2,89	1,53

Chi-Square = 28,2545 with 16 d.f. P-Value = 0,0294753

#### Kolmogorov-Smirnov Test

	<i>Normal</i>
DPLUS	0,0662769
DMINUS	0,0696088
DN	0,0696088
P-Value	0,952603

#### Anderson-Darling A^2

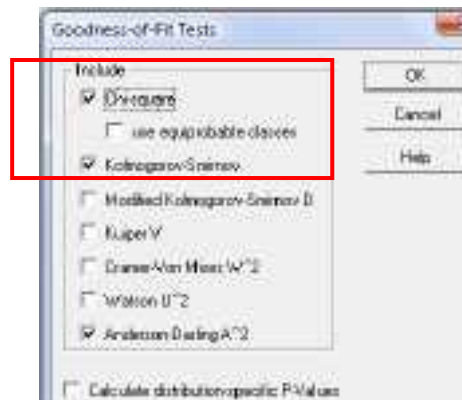
	<i>Normal</i>
A^2	0,499884
Modified Form	0,499884
P-Value	>=0.10

Lo que se muestra es el resultado de varios tests de bondad de ajuste. **Hay que tener cuidado con esta versión del Statgraphics por la forma en que realiza el test de la Chi-cuadrado, pues la opción que ofrece por defecto no es muy conveniente.** El test de la chi-cuadrado mide la discrepancia entre el histograma y el modelo que se propone. El resultado del test depende de cómo se construya este histograma. El Statgraphics Centurion ofrece dos opciones para la realización de este histograma:

Opción 1: Realizar un histograma con un número de clases especificado por un criterio automático, con clases de diferente amplitud de forma que la frecuencia esperada según el modelo que se evalúa sea la misma (clases equiprobables). Esta es la opción que el Statgraphics Centurion nos muestra por defecto.

Opción 2: Realizar el mismo histograma (o similar) que se dibuja en la parte derecha de los resultados, basado en un número de clases especificado por el usuario y con clases de igual amplitud (esta opción se selecciona en Pane Options).

En teoría la Opción 1 muestra, en general, resultados más precisos que la Opción 2. Sin embargo, para que esta opción muestre resultados válidos el número de clases debe proporcionar una frecuencia esperada superior a 5, y la frecuencia observada debe ser también superior a 5. Como puede verse en las figuras anteriores, el Statgraphics Centurion parece que no respeta estas recomendaciones, debido a que tiende a seleccionar un número demasiado elevado de clases. Por tanto, es mejor utilizar la opción 2. Para ello pulsamos el botón derecho del ratón, y en Pane Options eliminamos la opción de 'use equiprobable cases':



y el nuevo resultado es:

#### Goodness-of-Fit Tests for Ordenador\_Casa

##### Chi-Square Test

	<i>Lower</i>	<i>Upper</i>	<i>Observed</i>	<i>Expected</i>	
	<i>Limit</i>	<i>Limit</i>	<i>Frequency</i>	<i>Frequency</i>	<i>Chi-Square</i>
at or below		5,74286	5	4,42	0,08
	5,74286	5,88571	10	9,32	0,05
	5,88571	6,02857	16	14,91	0,08
	6,02857	6,17143	14	14,38	0,01
	6,17143	6,31429	5	8,36	1,35
above	6,31429		5	3,63	0,52

Chi-Square = 2,08464 with 3 d.f. P-Value = 0,555023

##### Kolmogorov-Smirnov Test

	<i>Normal</i>
DPLUS	0,0662769
DMINUS	0,0696088
DN	0,0696088
P-Value	0,952603

##### Anderson-Darling A^2

	<i>Normal</i>
A^2	0,499884
Modified Form	0,499884
P-Value	>=0.10

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

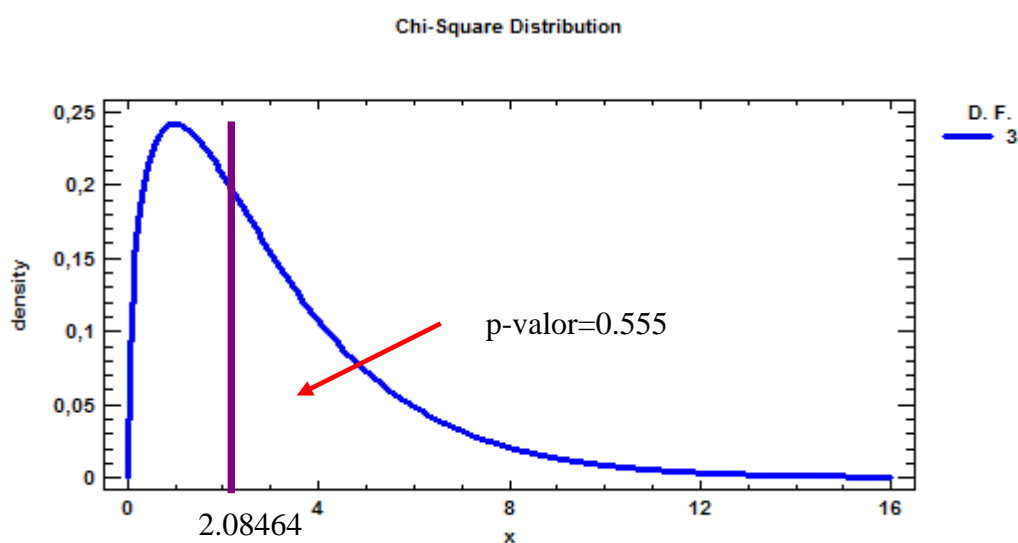
- **Chi-square = 2.08464**, que representa el valor del estadístico calculado en el test

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

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)= 3**, 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  $df = k - v - 1$ , donde:  
k= número de intervalos, en este caso 6. El número de clases es, en principio, el mismo que en el histograma que aparece en la parte derecha de la pantalla de resultados. No obstante, si hubiese clases con menos de 5 observaciones puede decidir agruparlas, por lo que el número de clases puede ser inferior al del histograma de la derecha  
v= número de parámetros del modelo escogido, en este caso 2 (media y varianza)
- **p\_value =0.555** (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  $\chi^2_{k-v-1}$ .

La teoría estadística nos dice que cuanto peor es el ajuste del modelo elegido, el estadístico  $\chi^2$  dará un valor mayor, y que la referencia para evaluar cómo de grande es ese estadístico en cada caso es la distribución  $\chi^2_{k-v-1}$ . Una forma sencilla de valorar la bondad del ajuste es calcular el área que queda a la derecha del valor del estadístico  $\chi^2$  en la distribución  $\chi^2_{k-v-1}$ . 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 0.555 por lo que concluimos que la normal es un modelo muy razonable para explicar la distribución de nuestros datos.

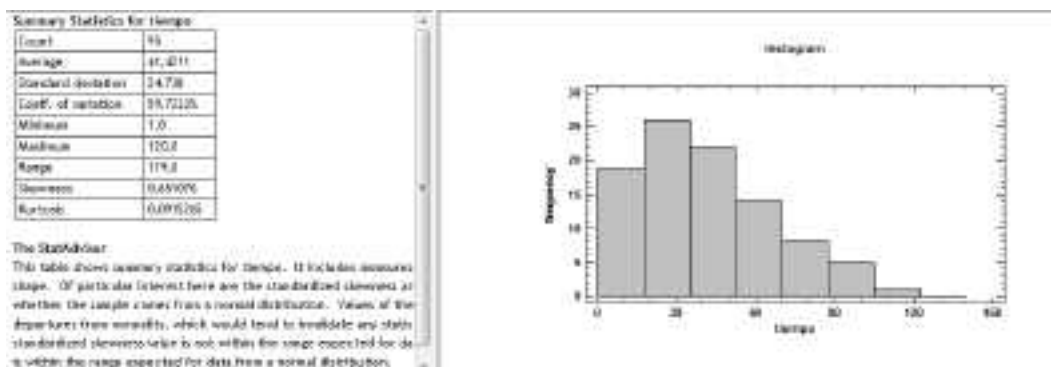
El Statgraphics realiza otros tests de bondad de ajuste. Los resultados de todos ellos pueden interpretarse también a través de sus p-valores de la misma forma que con el tests de la chi-cuadrado. Por ejemplo, puede observar que en el test de Kolmogorov-Smirnov el p-valor es también mayor de 0.05.

### 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 Describe/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.



Los datos son unimodales, con asimetría positiva. La zona de la moda tiene un apuntamiento en forma de campana. Tenemos dos opciones para 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 transformación 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

La distribución llamada Weibull se emplea mucho en la práctica para modelizar duraciones. Podría entonces ser útil para modelizar la variable 'Tiempo'. Como en el ejemplo anterior, vamos a Describe/Distributions/Distribution Fitting (Uncensored data), y allí seleccionamos la variable Tiempo.



**Distribution Fitting Options**

**Distribution**

<input type="checkbox"/> Bernoulli	<input type="checkbox"/> Exponential Power	<input type="checkbox"/> Lognormal (3-parameter)
<input type="checkbox"/> Binomial	<input type="checkbox"/> F (Variance Ratio)	<input type="checkbox"/> Maxwell (2-parameter)
<input type="checkbox"/> Discrete Uniform	<input type="checkbox"/> Folded Normal	<input type="checkbox"/> Noncentral Chi-Square
<input type="checkbox"/> Geometric	<input type="checkbox"/> Gamma	<input type="checkbox"/> Noncentral F
<input type="checkbox"/> Hypergeometric	<input type="checkbox"/> Gamma (3-parameter)	<input type="checkbox"/> Noncentral t
<input type="checkbox"/> Negative Binomial	<input type="checkbox"/> Generalized Gamma	<input type="checkbox"/> Normal
<input type="checkbox"/> Poisson	<input type="checkbox"/> Generalized Logistic	<input type="checkbox"/> Pareto
<input type="checkbox"/> Beta	<input type="checkbox"/> Half Normal (2-parameter)	<input type="checkbox"/> Pareto (2-parameter)
<input type="checkbox"/> Beta (4-parameter)	<input type="checkbox"/> Inverse Gaussian	<input type="checkbox"/> Rayleigh (2-parameter)
<input type="checkbox"/> Birnbaum-Saunders	<input type="checkbox"/> Laplace	<input type="checkbox"/> Smallest Extreme Value
<input type="checkbox"/> Cauchy	<input type="checkbox"/> Largest Extreme Value	<input type="checkbox"/> Student's t
<input type="checkbox"/> Chi-Square	<input type="checkbox"/> Logistic	<input type="checkbox"/> Triangular
<input type="checkbox"/> Erlang	<input type="checkbox"/> Loglogistic	<input type="checkbox"/> Uniform
<input type="checkbox"/> Exponential	<input type="checkbox"/> Loglogistic (3-parameter)	<input checked="" type="checkbox"/> Weibull
<input type="checkbox"/> Exponential (2-parameter)	<input type="checkbox"/> Lognormal	<input type="checkbox"/> Weibull (3-parameter)

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 Tables and Graphs seleccionamos Goodness-of-Fit Test, y en Pane Options seleccionamos que no se usen clases equiprobables (como tenemos 95 datos, podemos usar 8-9 clases)

Goodness-of-Fit Tests for Tiempo

Chi-Square Test

Lower Limit	Upper Limit	Observed Frequency	Expected Frequency	Chi-Square
at or below	18,75	19	18,32	0,03
	18,75	26	26,42	0,40
	37,5	24	23,14	0,09
	56,25	19	13,86	1,71
	75,0	5	6,36	0,29
above	81,75	2	2,39	0,57

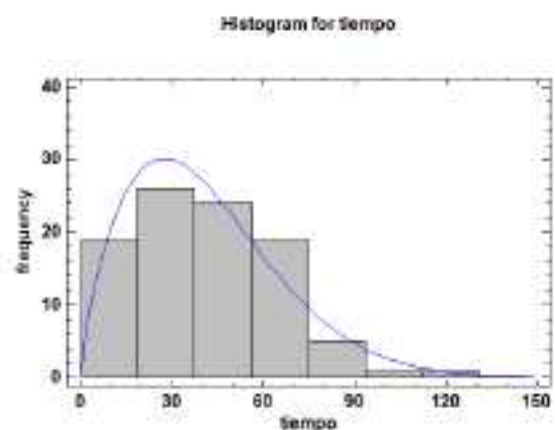
Chi-Square = 3,20981 with 4 d.f. P-Value = 0,50965

Kolmogorov-Smirnov Test

	Weibull
DP1US	0,0015627
DMPUS	0,0044992
DN	0,0066992
P-Value	0,401257

Anderson-Darling A\*2

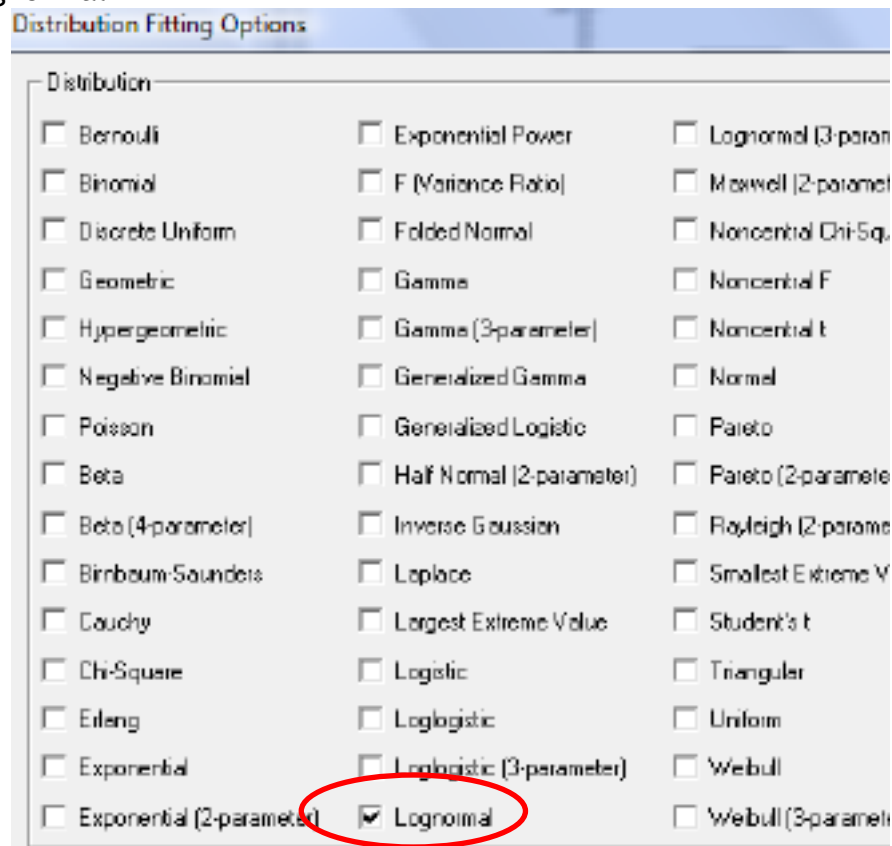
	Weibull
A*2	0,581257
Modified Form	0,581257
P-Value	>0,18



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

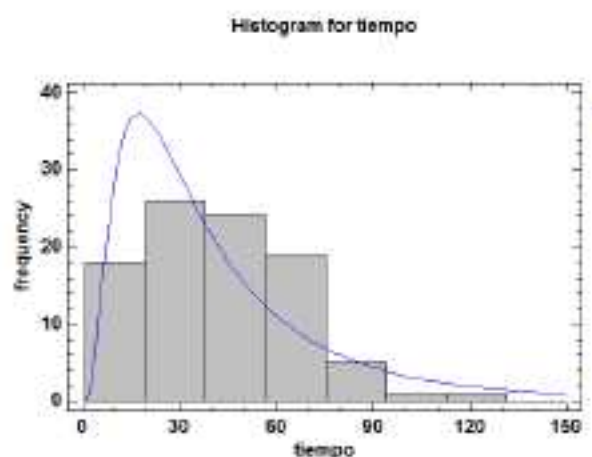
	Lower Limit	Upper Limit	Observed Frequency	Expected Frequency	Chi-Square
at or below		19.625	19	14.61	1.18
	19.625	38.25	38	18.32	0.60
	38.25	56.875	24	17.01	3.86
	56.875	75.5	19	9.21	18.33
	75.5	94.125	5	5.15	0.01
	94.125	112.75	1	1.09	1.37
	112.75	131.375	1	1.04	1.37
above	131.375		0	2.61	2.61

Chi-Square = 20.6174 with 5 d.f. P-Value = 0.000341

Kolmogorov-Smirnov Test

	Lognormal
DPLUS	0.080267
DMINUS	0.140001
DN	0.140001
D-Value	0.040001

Anderson-Darling A-S



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 (como son 95 obs. realizamos 10 clases)



Goodness-of-Fit Tests for sqrt(tiempo)

Chi-Square Test

	Lower Limit	Upper Limit	Observed Frequency	Expected Frequency	Chi-Square
at or below		2.4	4	3.00	0.38
	2.4	3.6	8	6.95	0.58
	3.6	4.8	18	14.34	0.94
	4.8	6.0	30	23.95	0.40
	6.0	7.2	22	21.66	0.01
	7.2	8.4	14	15.86	0.32
	8.4	9.6	12	8.22	1.74
above	9.6		2	3.16	0.57

Chi-Square = 5.18125 with 5 d.f. P-Value = 0.40000

Kolmogorov-Smirnov Test

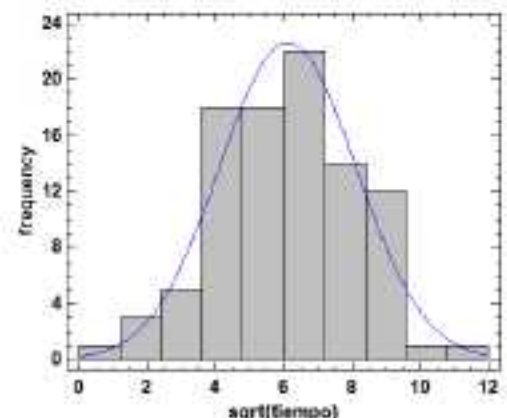
	Normal
DF(1.5)	0.03774268
Dev(4.8)	0.0281511
DN	0.0381911
P-value	0.557685

Anderson-Darling A\*2

	Normal
A*2	0.514372
Modified P-value	0.514372
P-value	0.16



Histogram for sqrt(tiempo)



que presenta un ajuste casi tan bueno como el de la Weibull. En Tables and Graphs seleccionamos Analysis Summary para ver los parámetros estimados para este modelo.

#### Uncensored Data - sqrt(tiempo)

Data variable: sqrt(tiempo)

95 values ranging from 1,0 to 10,9545

#### Fitted Distributions

Normal
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 tarde más de una hora en llegar a 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 esperamos que no sean muy diferentes.

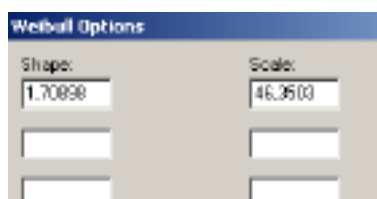
### 4.1 Cálculo con la Weibull

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 Tables and Graphs seleccionamos la Función de distribución. 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

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

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

Variable      Upper Tail Area (>)
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%.

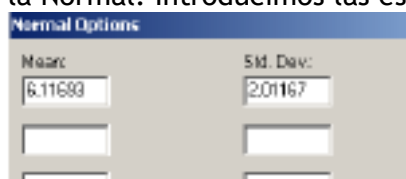
## 4.2 Cálculo con la Normal y la variable transformada

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

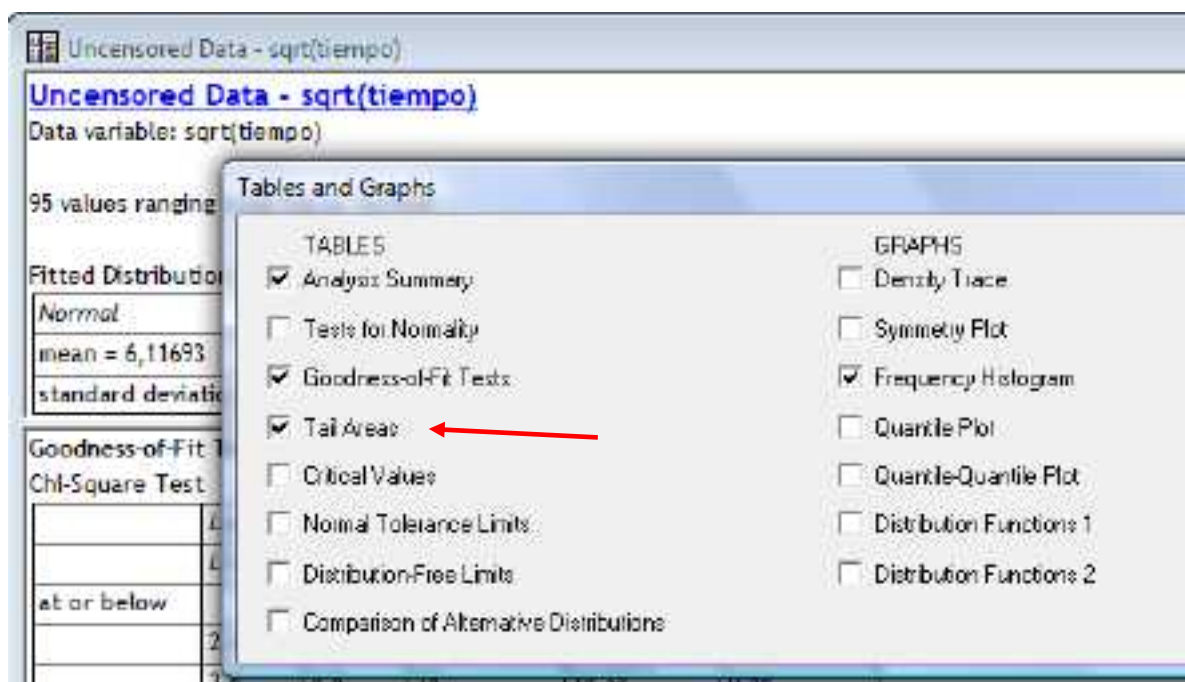
Variable      Lower Tail Area (<)
              Dist. 1      Dist. 2
7.746         0.720976

Variable      Probability Density
              Dist. 1      Dist. 2
7.746         0.142873

Variable      Upper Tail Area (>)
              Dist. 1      Dist. 2
7.746         0.279024
```

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.

Estas probabilidades también pueden calcularse dentro de las opciones del ajuste de distribuciones, en la opción 'Tail areas'



Se obtiene entonces la siguiente ventana de resultados

**Tail Areas for sqrt(tiempo)**  
Normal distribution

X	Lower Tail Area (<)	Upper Tail Area (>)
4,89355	0,271545	0,728455
5,50524	0,380535	0,619465
6,11693	0,499997	0,500003
6,72862	0,619465	0,380535
7,746	0,790976	0,209024

The StatAdvisor  
This pane calculates tail areas for the fitted normal distribution. It

**Tail Areas Options**

Critical Values:

4,89355  
5,50524  
6,11693  
6,72862  
7,746

OK  
Cancel  
Help

donde el valor de X puede cambiarse en Pane Options.