

# **Material para clase de análisis de los datos de entrada DES**

**Clase Modelación y Simulación V  
Ingeniería Matemática**

**Profesora: Paula Escudero**

**Medellín  
Septiembre 6 de 2019**

## Tabla de contenido

1	ANÁLISIS DE LOS DATOS DE ENTRADA.....	1
1.1	Análisis exploratorio de los datos usando Excel .....	1
1.1.1	Tablas dinámicas para analizar comportamientos.....	1
1.1.2	TAREA: Hacer un análisis exploratorio de los datos.....	2
1.2	Análisis estadístico de los datos usando RStudio.....	2
1.2.1	Realizar en RStudio.....	2
1.2.2	TAREA: Análisis estadístico de los datos .....	3
1.3	Pruebas de independencia y de bondad de ajuste .....	2
1.3.1	Estadística descriptiva .....	3
1.3.2	Análisis de independencia.....	3
1.3.3	Histograma de frecuencias.....	3
1.3.4	Ajuste de distribuciones .....	4
1.3.5	Pruebas de bondad de ajuste.....	7
1.3.6	Análisis y conclusiones de las pruebas de independencia y bondad de ajuste ....	8
1.3.7	TAREA: pruebas de independencia y bondad de ajuste.....	8
2	Tarea.....	10

# 1 ANÁLISIS DE LOS DATOS DE ENTRADA

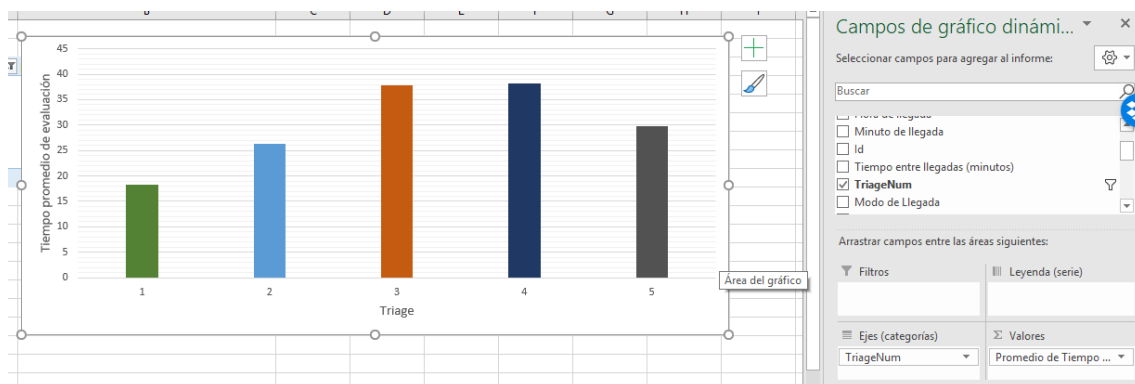
Es importante realizar un buen análisis de los datos del problema para poder determinar como modelar las entradas del modelo. Inicialmente se puede comenzar con un análisis exploratorio de los datos para entender el comportamiento general de éstos y luego realizar un análisis estadístico para probar las hipótesis planteadas sobre los datos.

Para esta clase se usarán los datos del archivo “DatosHospitalDescriptiva.xlsx” o “DatosHospitalDescriptivaText.txt” que se encuentran en formato texto. Se propone usar tablas dinámicas de Excel para gráficamente el comportamiento de las diferentes variables y plantear hipótesis sobre éstos y luego usar Rstudio para probar las hipótesis planteadas.

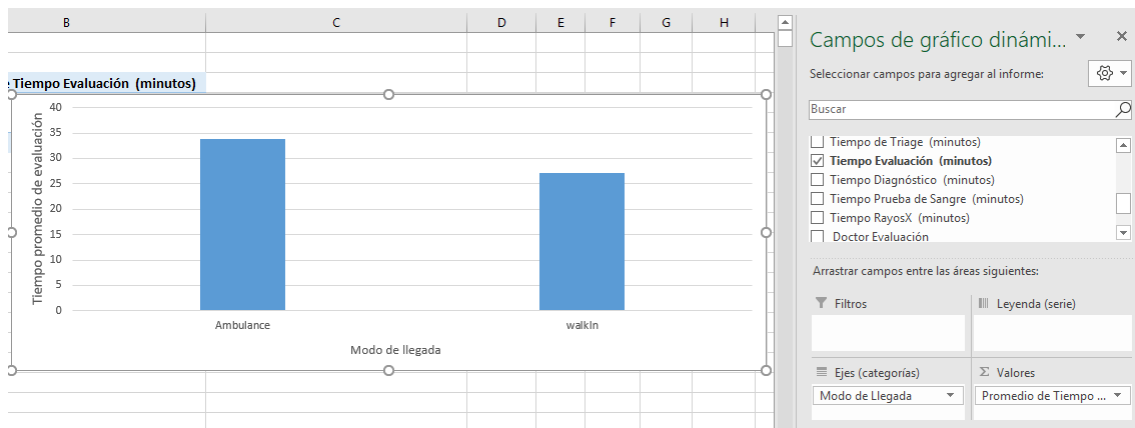
## 1.1 Análisis exploratorio de los datos usando Excel

### 1.1.1 Tablas dinámicas para analizar comportamientos

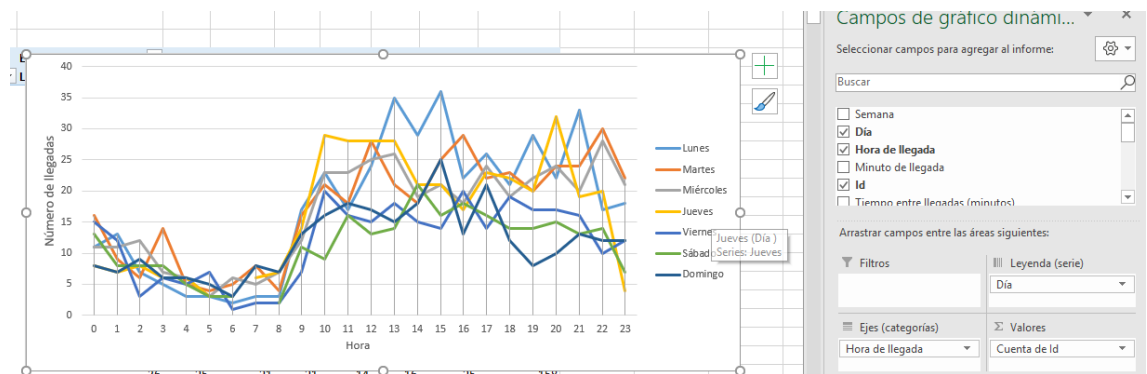
1. Por ejemplo, podría querer mirar si el tiempo promedio de evaluación es similar para cada valor del triage.



- ¿Qué conclusiones puede sacar de esta gráfica?
2. ¿Será que el tiempo de evaluación parece ser distinto para cada modo de llegada?



### 3. ¿Será que las llegadas variarían con el día y la hora del día?



- ¿Se podría hacer una sola distribución del tiempo entre llegadas para todo el día? ¿Qué conclusiones podría usted sacar de estos datos? ¿cómo modelaría el tiempo entre llegadas? (Por ejemplo, ¿cree usted que podría usar siempre una distribución exponencial durante el día y cambiar el parámetro  $\lambda$  cada hora?)

#### 1.1.2 TAREA: Hacer un análisis exploratorio de los datos

- Mediante un análisis exploratorio de los datos determine si podría existir una relación entre el tiempo de registro y alguna otra variable como ***Hora del día, Día, Triage, Modo de llegada (ambulancia o caminando).***
- Realizar el mismo ejercicio para el tiempo del triage, evaluación y diagnóstico.
- Tome nota de las conclusiones obtenidas para realizar un análisis estadístico que compruebe sus hipótesis.

#### 1.2 Análisis estadístico de los datos usando RStudio

##### 1.2.1 Realizar en RStudio

- Importe los datos de estadística descriptiva

```
> dataHosp <- read.delim("~/DatosHospitalDescriptivaText.txt")
> view(dataHosp)
```

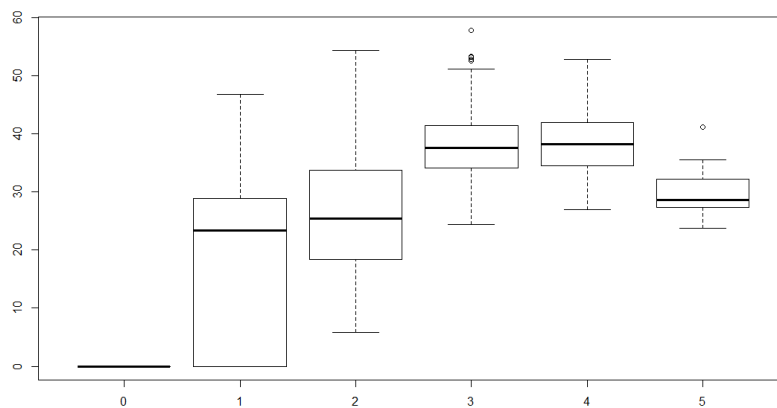
2.

Extraer variables de interés: por ejemplo si usted quiere probar la homogeneidad del tiempo de evaluación con relación al triage, entonces:

```
> teval = dataHosp$Tiempo.Evaluación...minutos.
> triage = dataHosp$Triagenum
```

3. Puede realizar un diagrama de cajas para mirar la relación:

```
> boxplot(teval ~ triage, data=dataHosp)
```



- ¿Qué conclusiones saca de este análisis preliminar?
- ¿Será que todos los datos provienen de la misma población?
- ¿Será que la distribución del tiempo de evaluación es la misma para cada tipo de triage?
- ¿qué características tendría cada distribución (simétrica, sesgada, etc)?
- ¿se pueden unir todos los datos para cada triage para hacer una sola distribución?
- ¿Cuáles valores del triage parecen tener la misma distribución?

4. Realizar la prueba de Kruskal Wallis para concluir si los datos del tiempo de evaluación son homogéneos o dependen del triage.

```
> kruskal.test(teval ~ triage, data=dataHosp)

kruskal-wallis rank sum test

data:  teval by triage
kruskal-wallis chi-squared = 786.37, df = 5, p-value < 2.2e-16
.
```

5. Concluya sobre los resultados del test.

## 1.2.2 TAREA: Análisis estadístico de los datos

1. Con base en el análisis exploratorio de los datos, hacer un análisis estadístico para comprobar si existe una relación entre el tiempo de registro y alguna otra variable como:
  - a. Hora del día

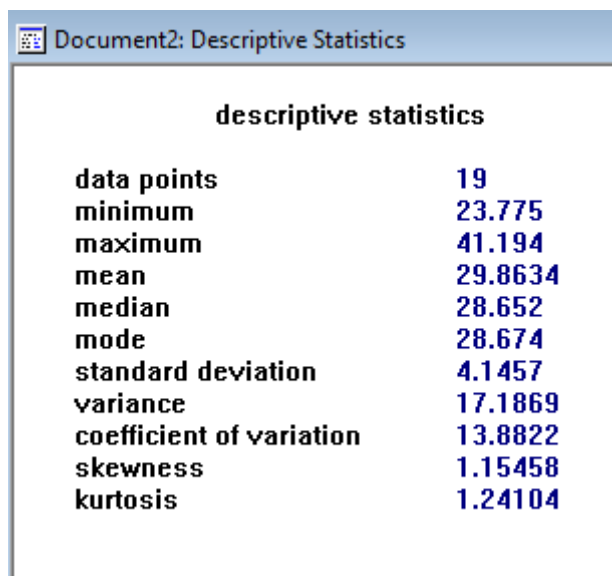
- b. Día
  - c. Triage
  - d. Modo de Llegada (ambulancia o caminando)
2. Realice el mismo ejercicio para el tiempo del triage, evaluación y diagnóstico.
3. Determine cuales variables va a analizar para realizar las pruebas de bondad de ajuste correspondientes. Por ejemplo, si usted concluyó que el tiempo de evaluación es diferente para cada tipo de triage usted debe seleccionar una muestra de la base de datos para hacer el análisis.

### 1.3 Pruebas de independencia y de bondad de ajuste

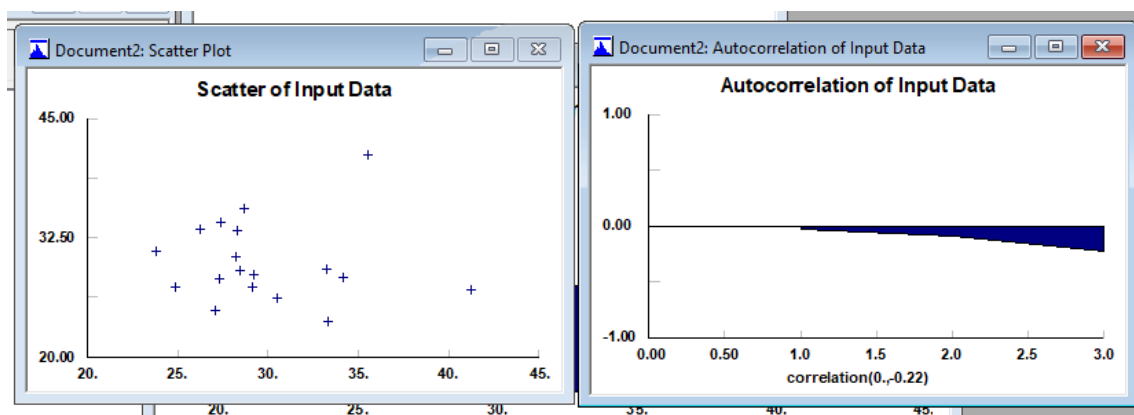
Por ejemplo, si usted determinó que el tiempo de evaluación es diferente para cada triage, pero no depende ni del día ni de la hora, entonces usted puede obtener una muestra para cada tipo de triage. Por ejemplo, la distribución para el triage =5 pude ser obtenida filtrando el tiempo de evaluación con triage = 5

	A	B	C	D	E	F	G	H	I	J	K	L	M
	Semana	Día	Hora de Llegada	Minuto de Llegada	Id	Tiempo entre Llegadas	TriageNum	Modo de Llegada	Tiempo en el sistema (minutos)	Tiempo de Registro (minutos)	Tiempo de Triage (minutos)	Tiempo Evaluación (minutos)	Tiempo Diagnóstico (minutos)
46	0	Lunes	13	20.81				Ambulance	120.855	0.000	0.000	28.397	4.955
168	0	Martes	8	31.78				Ambulance	76.896	0.000	0.000	29.093	10.253
329	0	Miércoles	11	1.112				Ambulance	88.282	0.000	0.000	27.367	29.118
385	0	Miércoles	17	35.50				Ambulance	156.951	0.000	0.000	34.141	10.319
429	0	Miércoles	23	21.45				Ambulance	51.927	0.000	0.000	28.321	2.971
130	0	Miércoles	23	35.92				Ambulance	94.251	0.000	0.000	33.219	16.067
523	0	Jueves	19	49.23				Ambulance	206.274	0.000	0.000	29.174	107.147
546	0	Jueves	23	45.36				Ambulance	102.635	0.000	0.000	28.652	36.508
560	0	Viernes	1	50.89				Ambulance	150.363	0.000	0.000	35.491	37.410
935	0	Domingo	18	2.147				Ambulance	114.665	0.000	0.000	41.194	12.567
228	1	Martes	20	58.49				Ambulance	101.304	0.000	0.000	27.082	21.153
284	1	Miércoles	10	11.19				Ambulance	137.37	0.000	0.000	24.848	36.934
321	1	Miércoles	16	48.06				Ambulance	108.014	0.000	0.000	27.325	12.421
490	1	Jueves	20	23.37				Ambulance	135.511	0.000	0.000	28.174	42.260
614	1	Viernes	17	41.59				Ambulance	164.704	0.000	0.000	30.479	25.337
759	1	Sábado	20	23.58				Ambulance	109.59	0.000	0.000	26.223	16.842

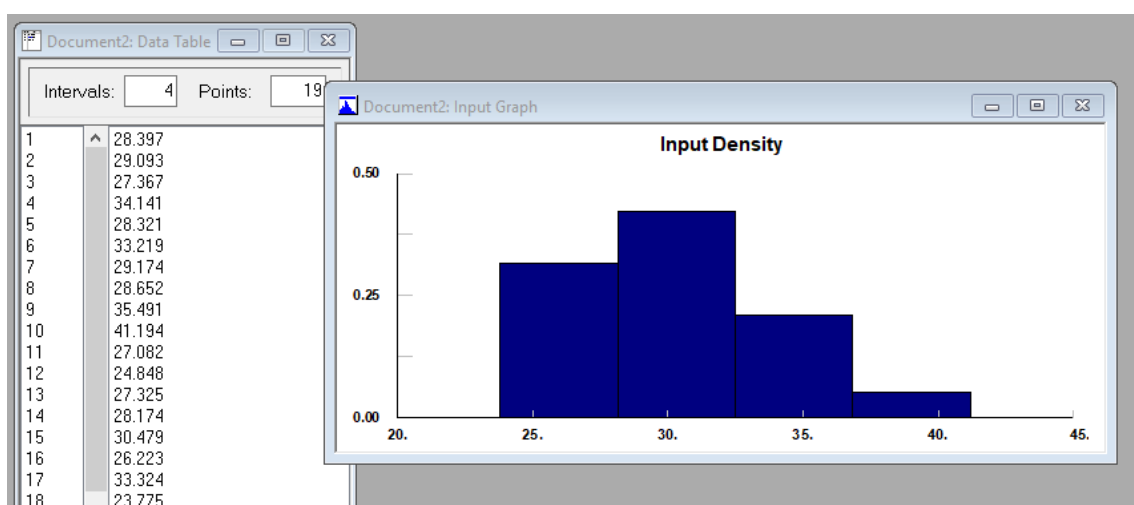
### 1.3.1 Estadística descriptiva



### 1.3.2 Análisis de independencia

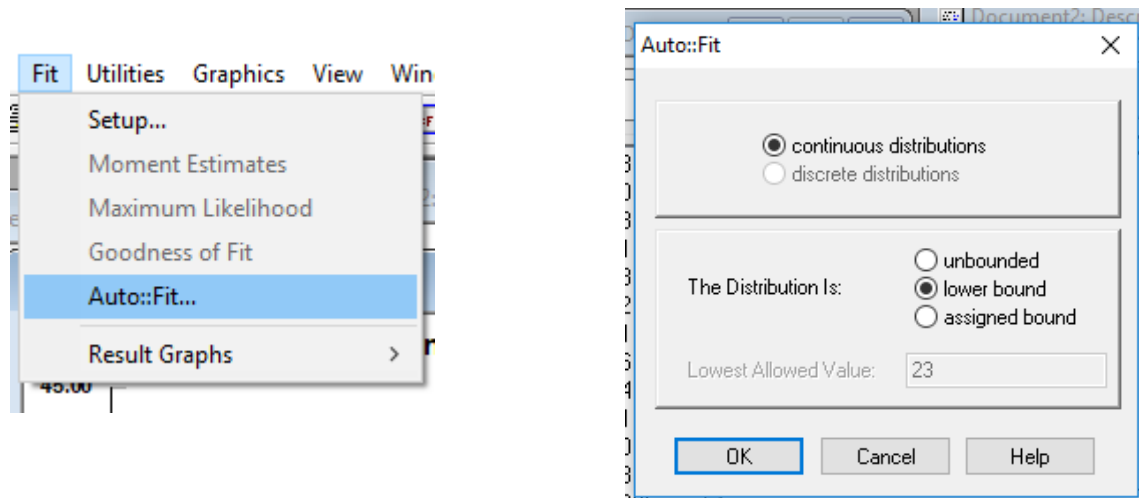


### 1.3.3 Histograma de frecuencias



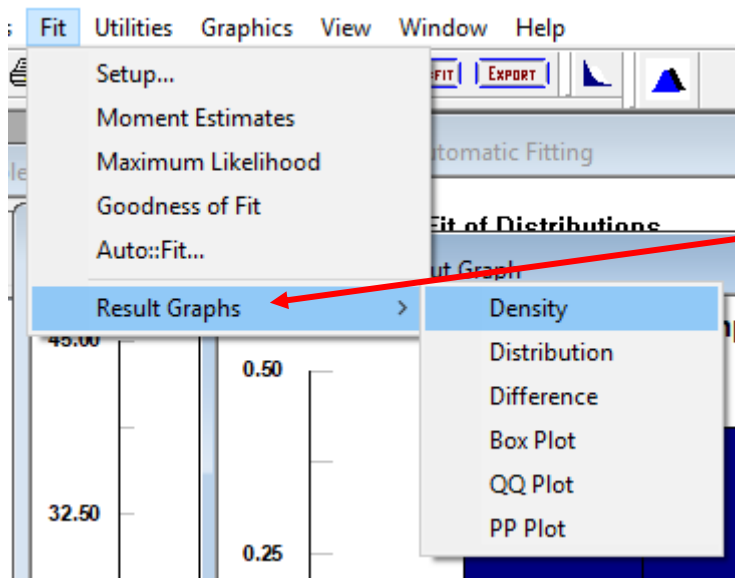
### 1.3.4 Ajuste de distribuciones

Realice el ajuste de distribuciones y las gráficas correspondientes.

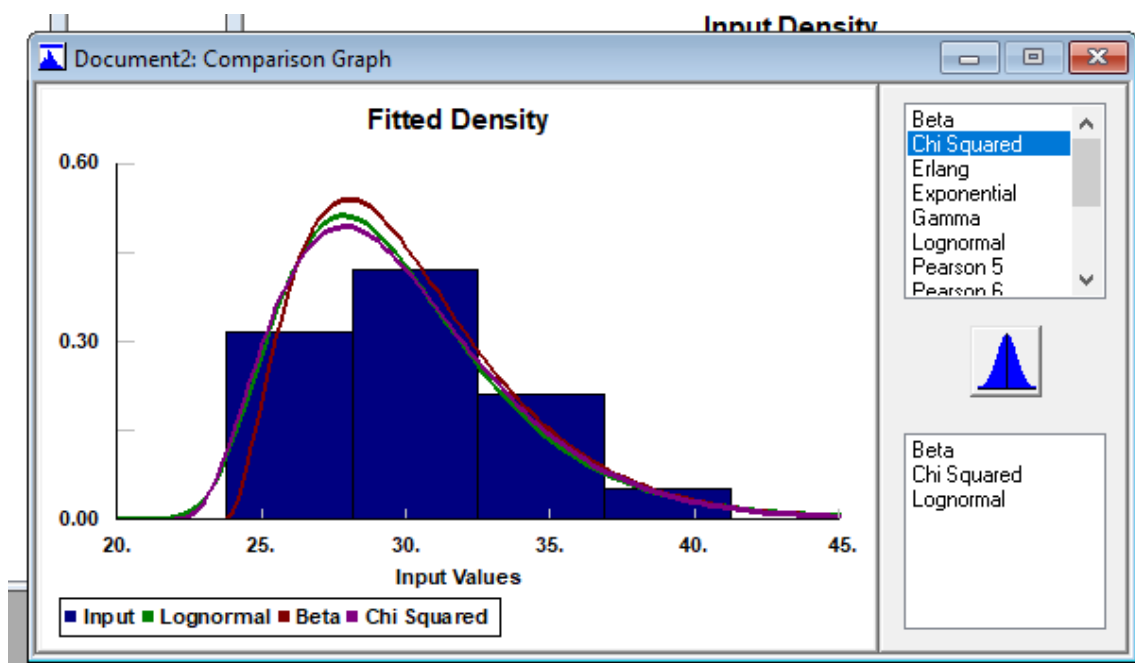


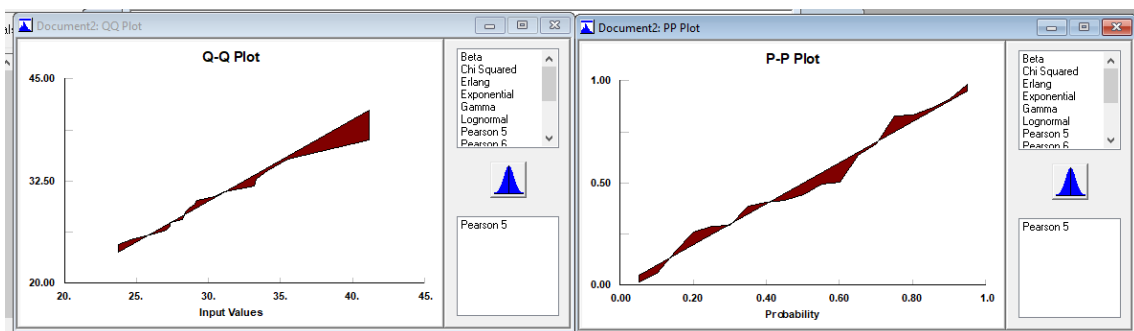
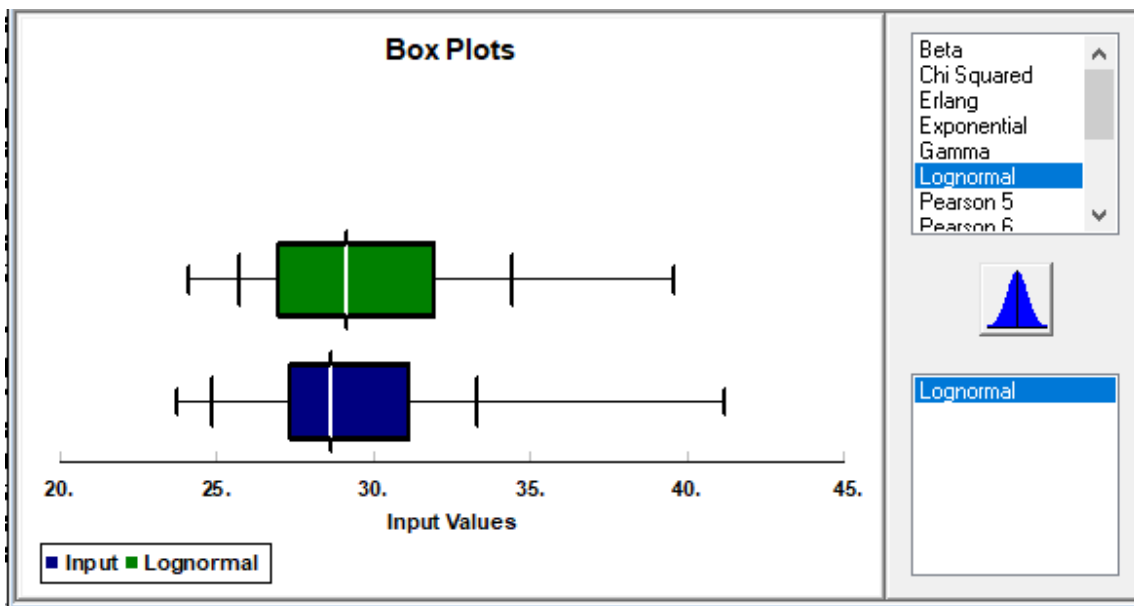
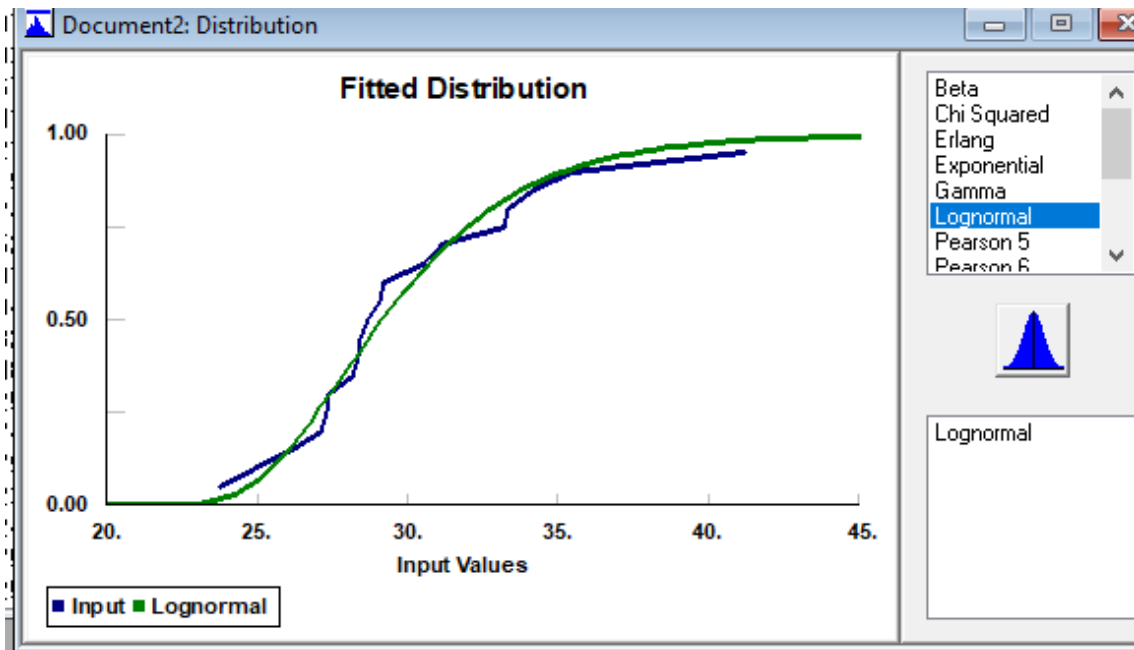
Auto::Fit of Distributions		
distribution	rank	acceptance
Pearson 5(16.7, 12.4, 151)	100	do not reject
<u>Lognormal(19.5, 2.26, 0.379)</u>	<u>98.8</u>	<u>do not reject</u>
Gamma(22.3, 3.55, 2.14)	96.6	do not reject
Chi Squared(22., 7.83)	94.6	do not reject
Weibull(23.3, 1.66, 7.3)	92.3	do not reject
Erlang(22.3, 4., 1.9)	89.4	do not reject
Pearson 6(23.8, 20.7, 3.7, 12.9)	82.6	do not reject
Beta(23.8, 592, 2.86, 250)	66.	do not reject
Rayleigh(22.7, 5.81)	60.4	do not reject
Triangular(22.6, 42.5, 27.3)	14.1	do not reject
Exponential(23.8, 6.09)	1.59	do not reject
Power Function(23.7, 46.6, 0.602)	0.809	do not reject
Uniform(23.8, 41.2)	3.85e-002	reject



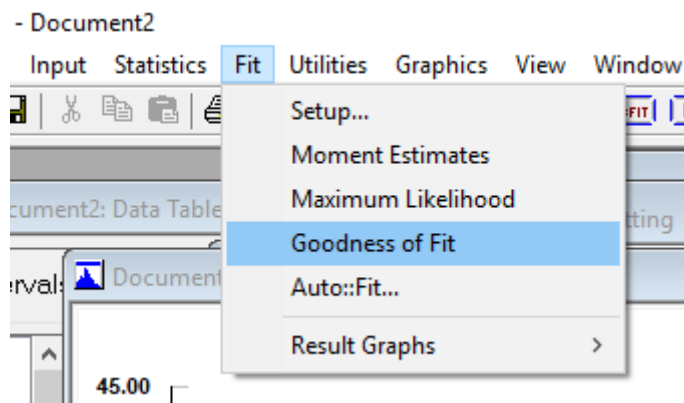


Realice las gráficas pertinentes. Por ejemplo, densidad, distribución, Box Plot, QQ plot, PP Plot y analícelas.





### 1.3.5 Pruebas de bondad de ajuste



#### goodness of fit

data points	19
estimates	maximum likelihood estimates
accuracy of fit	3.e-004
level of significance	5.e-002

#### summary

distribution	Kolmogorov Smirnov	Anderson Darling
Beta	0.169	0.271
Chi Squared	0.134	0.257
Erlang	0.139	0.278
Exponential	0.261	1.86
Gamma	0.131	0.258
Lognormal	0.129	0.237
Pearson 5	0.127	0.232
Pearson 6	0.15	0.24
Power Function	0.277	2.12
Rayleigh	0.17	0.356
Triangular	0.22	0.853
Uniform	0.322	3.79
Weibull	0.132	0.306

<b>Lognormal</b>		
minimum	=	19.547
mu	=	2.2619
sigma	=	0.379231
<b>Kolmogorov-Smirnov</b>		
data points		19
ks stat		0.129
alpha		5.e-002
ks stat(19,5.e-002)		0.301
p-value		0.872
result		DO NOT REJECT
<b>Anderson-Darling</b>		
data points		19
ad stat		0.237
alpha		5.e-002
ad stat(5.e-002)		2.49
p-value		0.977
result		DO NOT REJECT

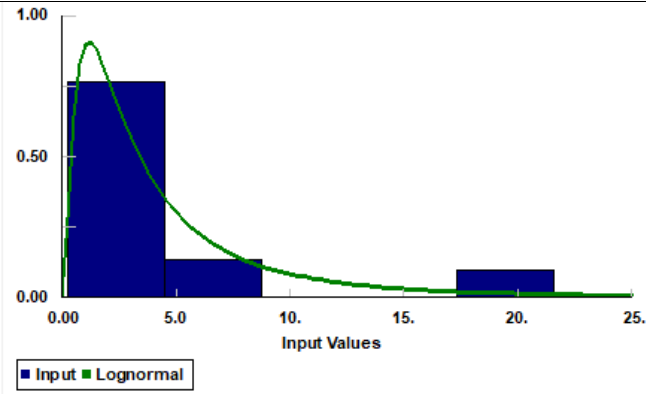
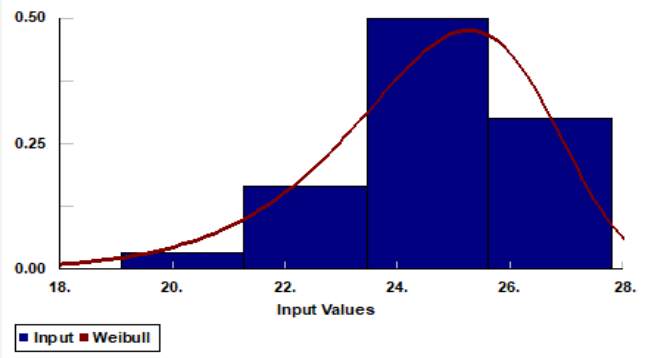
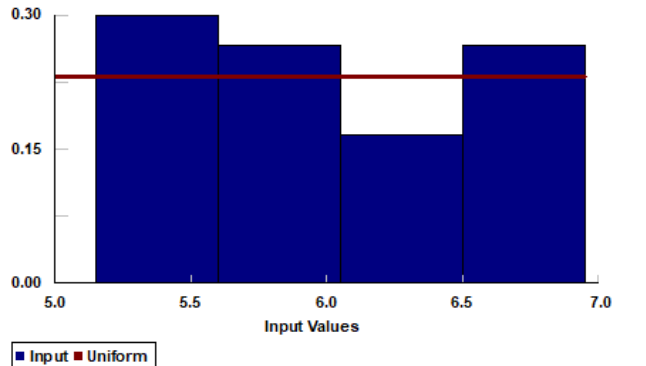
### 1.3.6 Análisis y conclusiones de las pruebas de independencia y bondad de ajuste

Concluya sobre los análisis realizados.

### 1.3.7 TAREA: pruebas de independencia y bondad de ajuste

Realice las pruebas de independencia y bondad de ajuste para al menos dos distribuciones obtenidas de los datos anteriormente analizados.

A continuación, se propone una plantilla para resumir los resultados de las pruebas de bondad de ajuste para las variables de interés. (las variables que se muestran en la tabla son ejemplos que no representan los datos de este ejercicio)

Nombre de la variable	Tamaño de la muestra	Distribución de probabilidad que mejor se ajusta	Estadístico y Valor P	Parámetros	Estadísticas descriptivas	Histograma
Tiempo entre llegadas a la caja registradora	30 datos	Lognormal	Kolmogorov-Smirnov VP 0.428 Anderson-Darling VP 0.650	Mínimo: 0 Mu: 1.12523 Sigma: 0.992513	<b>descriptive statistics</b>  data points 30 minimum 0.231335 maximum 21.5401 mean 4.92759 median 3.26137 mode 3.38759 standard deviation 5.50369 variance 30.2906 coefficient of variation 111.691 skewness 2.19883 kurtosis 3.65696	 <p>Input Lognormal</p>
Tiempo de registro en caja	30	Weibull	Kolmogorov-Smirnov VP 0.935 Anderson-Darling VP 0.977	Mínimo: - 1.0254 Alpha: 15.711 Beta: 26.4061	<b>descriptive statistics</b>  data points 30 minimum 19.0916 maximum 27.7873 mean 24.5164 median 24.4746 mode 23.9733 standard deviation 1.97851 variance 3.9145 coefficient of variation 8.07016 skewness -0.619114 kurtosis 0.373556	 <p>Input Weibull</p>
Tiempo de entrega del pedido	30	Uniforme	Kolmogorov-Smirnov VP 0.988 Anderson-Darling VP 0.965	Mínimo: 5 Máximo: 6.9436	<b>descriptive statistics</b>  data points 30 minimum 5.14767 maximum 6.94936 mean 5.97518 median 5.95019 mode 5.99084 standard deviation 0.580945 variance 0.337498 coefficient of variation 9.72264 skewness 0.179577 kurtosis -1.30002	 <p>Input Uniform</p>

## 2 Tarea

Considere los datos que se encuentran en el archivo “DatosOperadoresPunto3.xlsx”. Suponga que esos son muestras de los datos reales de los tiempos de atención de los operadores 1 y 2.

- Verifique la independencia de los datos. Para eso use diagramas de dispersión y autocorrelación e interprételos (Use StatFit for Simul8).
- Evalúe la homogeneidad de los datos usando la prueba de Kruskal-Wallis.
- Grafique el histograma de frecuencias para cada una de las variables.
- Haga una prueba de bondad de ajuste usando el estadístico de Kolmogorov-Smirnov (K-S) para cada una de las variables. (Use StatFit for Simul8). Escriba los resultados de la prueba para la distribución seleccionada y los parámetros de esta distribución.

Variable	Distribución	Valor P del K-S Test	Parámetros de la distribución
Tiempo Atención Operador 1			
Tiempo Atención Operador 2			