Análisis de datos de entrada

Paula Escudero

Bl 38-407

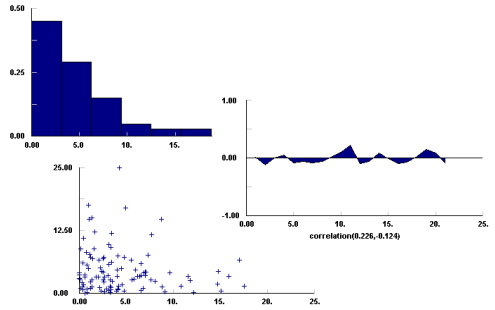
email: pescuder@eafit.edu.co

Escuela de Ciencias Aplicadas e Ingeniería

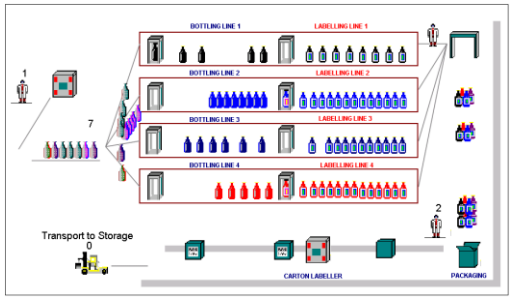
Análisis de Datos de Entrada

Sistema

Análisis Estadístico



Modelo de Simulación



- Tiempos de procesamiento
- Tiempos de transporte
- Tiempos de cambio de referencia
- Cantidades demandadas
- Tiempo entre fallas
- Tiempos de reparación

- Tiempos de abastecimiento
- Cantidades Defectuosas
- Tiempos de Atención
- Tiempo entre llegadas
- Tipo de Clientes
- .....

Tarea

1. Leer: Biller, B., & Nelson, B. L. (2002, December). Answers to the top ten input modeling questions. In *Proceedings of the Winter Simulation Conference* (Vol. 1, pp. 35-40). IEEE. https://www.informs-sim.org/wsc02papers/005.pdf
2. Leer: Capítulo 7 del libro: Robinson, Stewart. Simulation: the practice of model development and use. Palgrave Macmillan, 2014. **Hacer los ejercicios: E7.2, E7.3, E 7.4, E 7.6**

Análisis de los datos de entrada

1. Obtención de datos
2. Análisis de Independencia de los datos.
3. Evaluación de la Homogeneidad de los datos.
4. Ajuste de distribuciones teóricas:
    i. Identificar la distribución con los datos (Hipótesis):
        a. Histogramas
        b. Selección de la familia de Distribuciones posibles
        c. Gráficas de Probabilidad (Q-Q Plot)
    ii. Estimación de parámetros
    iii. Pruebas de Bondad de Ajuste
5. Selección de distribuciones en ausencia de datos históricos

Obtención de datos

- Categorías de disponibilidad de datos
  - Categoría A: Disponibles
    - Asegurarse de que los datos sean precisos y en el formato correcto para el modelo de simulación
  - Categoría B: No disponibles pero se pueden obtener
    - Se necesita gente o sistemas electrónicos para monitorear la operación
    - Posiblemente entrevistas con expertos
    - Asegurarse de que los datos sean precisos y en el formato correcto para el modelo de simulación
  - Categoría C: No disponibles, No obtenibles
    - El sistema del mundo real no existe
    - Tiempo limitado para recoger datos de eventos que ocurren esporádicamente
    - Demanda?

Datos existentes

ANÁLISIS DE LA INDEPENDENCIA DE LOS DATOS

- Ejemplos de datos correlacionados:
  - Tiempos de ejecución de operaciones manuales (Aprendizaje o Fatiga).
  - Tiempos de corte con herramientas que se desgastan.
  - Tiempos de espera en un servidor.
  - Demanda (Efecto Bullwhip, ciclo de vida del producto)
  - Ocurrencia de productos defectuosos en un proceso.

  Muchas de las pruebas estadísticas utilizadas asumen independencia de los datos (Chi cuadrado, estimación de máxima verosimilitud, etc).

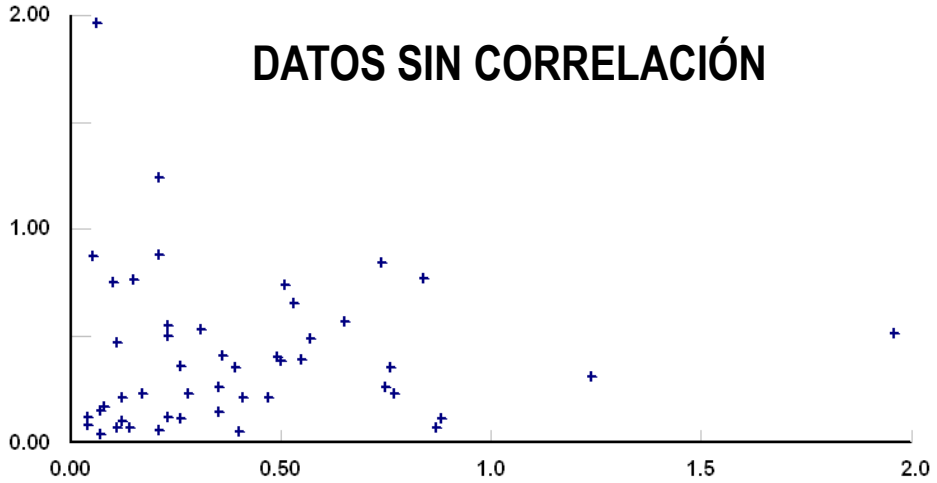
  Si no se cumple se debe modelar la dependencia de los datos.
- Técnicas para evaluar la independencia de los datos:
  - Diagramas de Dispersión (x_i,x_{i+1}) para x_1, $x_{2,}$...,x_{n-1}.
  - Diagrama de Autocorrelación.

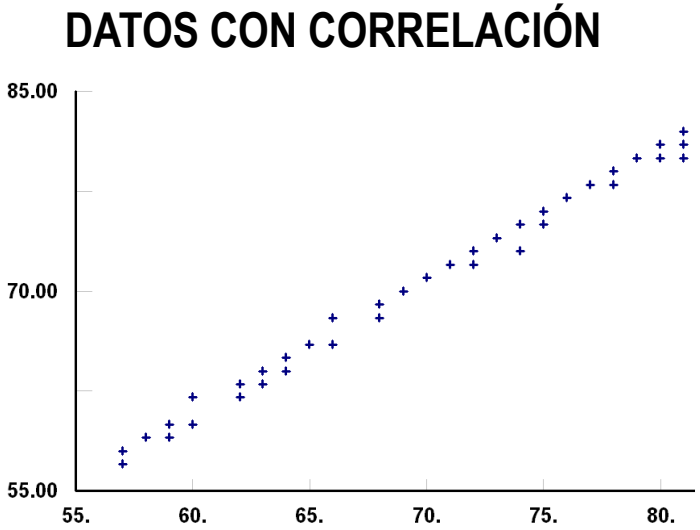
Herramientas

1. StatFit
   - Ver manual 1 en el link
   - No hay licencias para estudiantes
   - Manual 2 (en Eafit Interactiva)
2. Rstudio
3. Excel tablas dinámicas

Diagrama de Dispersión

- Si los datos son independientes se espera que estén dispersos aleatoriamente en el primer cuadrante.





- Si están positivamente o negativamente correlacionados se verán como una línea con pendiente positiva o negativa según el caso.

- En Stat:Fit® se usa **Statistics- Independence-Scatter Plot**

Diagrama de Autocorrelación (1/2)

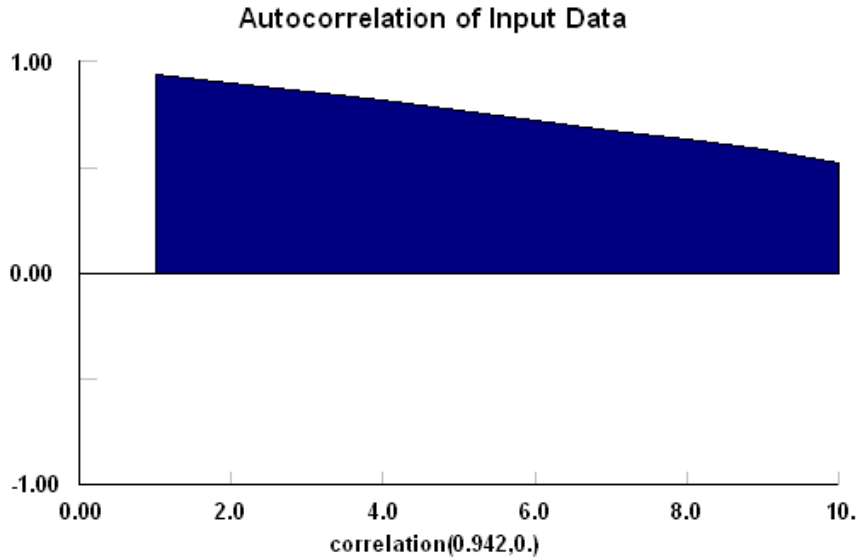
- El diagrama de dispersión solo evalúa la correlación entre datos consecutivos, pero la correlación puede ser cada *j* datos. Se usa la correlación muestral ρ, (rho) para evaluar la independencia de los datos.

$$\rho = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(X_{i+j} - \overline{X}\right)}{\sigma^2\left(n - j\right)} \qquad -1 \leq \rho \leq 1$$

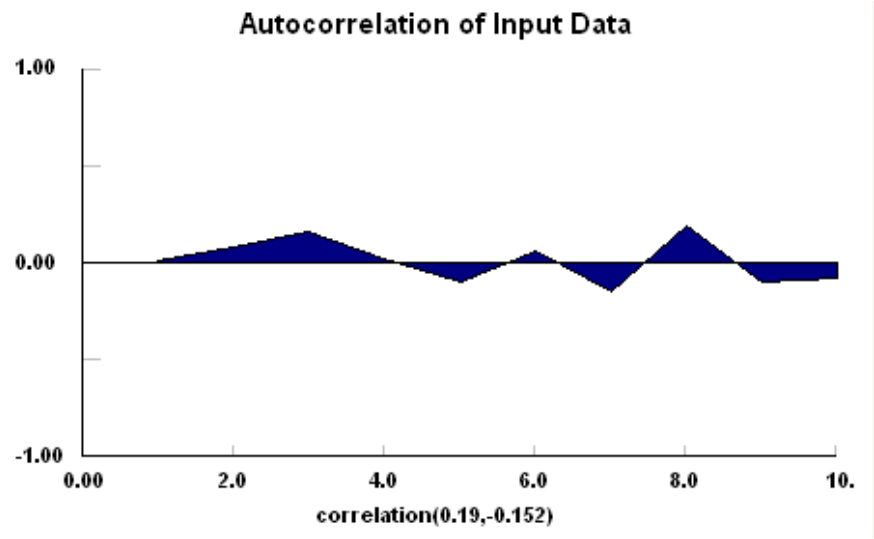
- Se calcula solo hasta $^1/_5$ de los datos ya que cuando se tienen menos pares los cálculos no son tan confiables.
- Valores de ρ cercanos a los extremos indican que los datos están correlacionados.

Diagrama de Autocorrelación (2/2)

DATOS CON CORRELACIÓN



DATOS SIN CORRELACIÓN

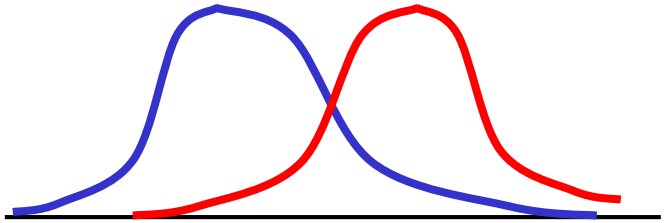


- En Stat:Fit® se usa **Statistics- Independence- Autocorrelation**

EVALUACIÓN DE LA HOMOGENEIDAD DE LOS DATOS

- Algunas veces se recogen diferentes conjuntos de datos de un mismo fenómeno aleatorio, y es necesario preguntarse si es posible unirlos para modelar un solo fenómeno o no. Es decir,
  - Ejemplos de situaciones en las que puede haber datos no homogéneos:
    - Tiempo entre llegadas de los clientes a diferentes horas del día.
    - Tiempos de servicio de diferentes servidores.
    - Tiempos de operación entre diferentes procesos, operarios, turnos, etc.
    - Porcentaje de productos defectuosos resultantes de procesos similares.
    - Tiempos de entrega de diferentes proveedores
    - ...

Muestras No Homogéneas



- Técnicas para evaluar la homogeneidad de los datos:
  - Pruebas no paramétricas (Kruskal-Wallis o Wilcoxon), ó paramétricas (Anova)
  - Informalmente (Ajuste de distribuciones y comparación)

Prueba de Kruskal-Wallis (1/4)

Ho: Las muestras provienen de la misma población (i.e.,son homogéneas)

H_1: Al menos una de las muestras presenta observaciones mayores a las demás

k: Número de muestras

n_i: Número de observaciones en la i-ésima muestra

X_{ij}: j-ésima observación en la i-ésima muestra

n: Número total de observaciones

Hay mas de 5 datos por muestra, las muestras son aleatorias e independientes.

Prueba de Kruskal-Wallis (2/4)

Para construir el estadístico Kruskal – Wallis se organizan los n datos de forma ascendente, y se le asigna al menor un rango igual a 1, al segundo menor un rango de 2, y así sucesivamente se asigna un rango a cada una de las n observaciones. En caso de haber observaciones iguales, a estas observaciones se les asigna el promedio de sus rangos normales:

Ejemplo:

Variable Y	Rango normal	Rango
212	10	10
199	5	5.5
202	7	7.5
194	3	3
188	2	2
187	1	1
199	6	5.5
196	4	4
206	9	9
202	8	7.5

En Rstudio:
Kruskal.test(var1 ~ var2, data = Data)

Prueba de Kruskal-Wallis (3/4)

$R(x_{ij})$: Rango Asignado a la observación x_{ij}

R_i : Suma de rangos asignados a los valores de la i-ésima muestra.

$$R_i = \sum_{j=1}^{n_i} R(x_{ij}) \quad \forall i$$

T : Estadístico de prueba Kruskal-Wallis

$$T = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(n+1)$$

Se rechaza la hipótesis nula si $T > \chi^2_{k-1,1-\alpha}$

Donde k-1 son los grados de libertad y 1-α es el nivel de confianza.
(o si VP < α se rechaza la hipótesis nula)

Ejemplo prueba K-W (4/4)

Linea 1		Linea 2		Linea 3	
Defectuosos	Rango	Defectuosos	Rango	Defectuosos	Rango
6	5	34	25	13	9.5
38	27	28	19	35	26
3	2	42	30	19	15
17	13	13	9.5	4	3
11	8	40	29	29	20
30	21	31	22	0	1
15	11	9	7	7	6
16	12	32	23	33	24
25	17	39	28	18	14
5	4	27	18	24	16
R_1	120	R_2	210.5	R_3	134.5

$$T = \frac{12}{30(31)}\left[\frac{(120)^2}{10} + \frac{(210.5)^2}{10} + \frac{(134.5)^2}{10}\right] - 3(31) = 6.099$$

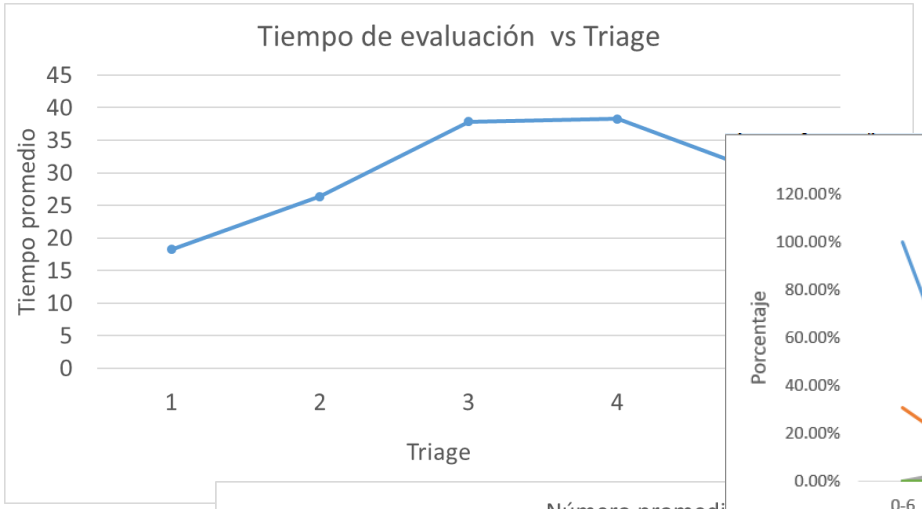
- $\chi^2_{0.05,2}$=5.9917 Se concluye con un nivel de signficancia de α=0.05 que al menos una de las líneas produce mayor numero de artículos defectuosos.

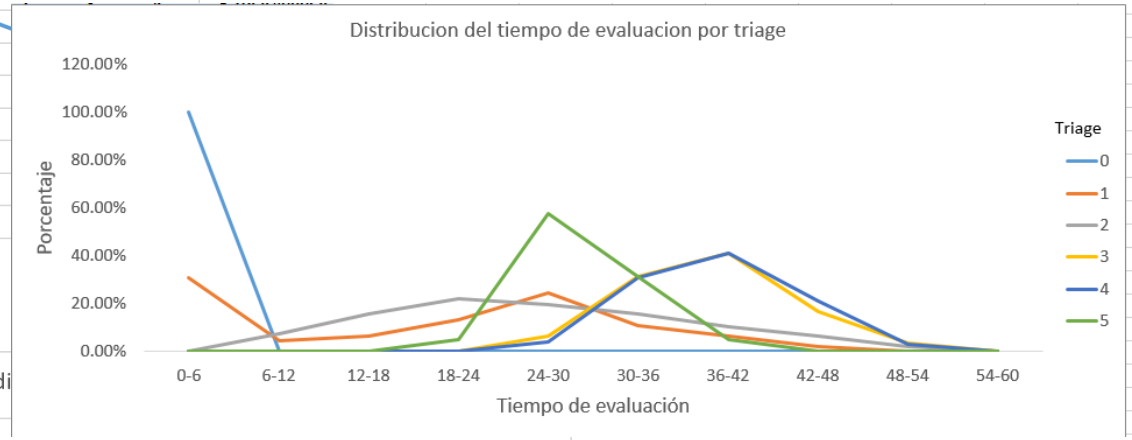
Ejemplo

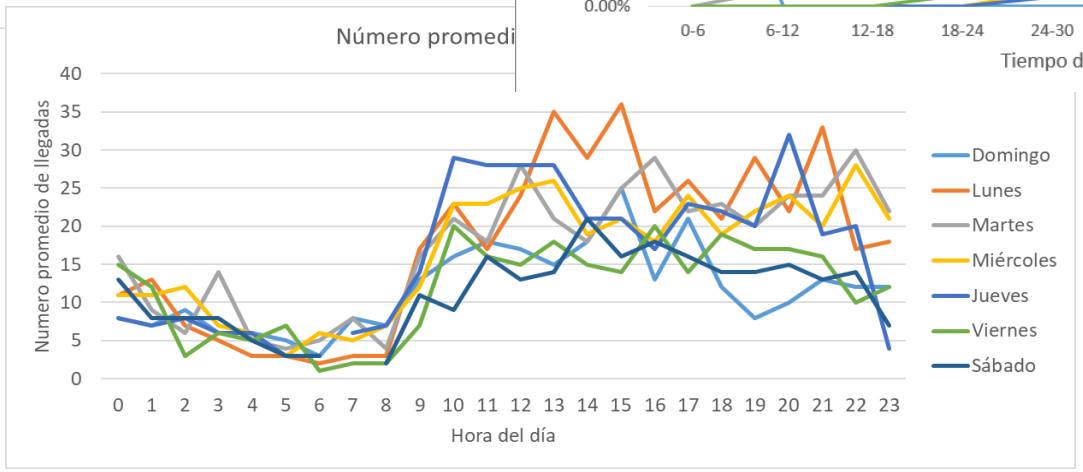
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	Semana	Día	Hora de llegada	Minuto de llegada	Id	Tiempo entre llegadas (minutos)	TriageNum	Modo de Llegada	Tiempo en el sistema (minutos)	Tiempo de Registro (minutos)	Tiempo de Triage (minutos)	Tiempo Evaluación (minutos)	Tiempo Diagnóstico (minutos)	Tiempo Prueba de Sangre (minutos)	Tiempo RayosX (minutos)	Doctor Evaluación	Doctor Diagnóstico
2	0	Lunes	0	36.13068	Patient 1	36.1307	2	walkIn	36.131	4.988	6.851	17.649	6.642	0.000	0.000	doctor 7	doctor 7
3	0	Lunes	1	42.9348852	Patient 2	66.8042	3	Ambulance	65.912	0.000	0.000	39.732	5.813	20.367	0.000	doctor 1	doctor 1
4	0	Lunes	1	52.4408262	Patient 5	9.5059	2	Ambulance	44.195	0.000	0.000	29.579	14.616	0.000	0.000	doctor 1	doctor 1
5	0	Lunes	1	59.2308933	Patient 3	6.7901	2	walkIn	74.516	4.017	6.376	19.628	4.088	0.000	40.408	doctor 1	doctor 1
6	0	Lunes	2	7.44431113	Patient 4	8.2134	2	walkIn	75.52	5.182	6.635	8.875	7.129	0.000	47.699	doctor 6	doctor 1
7	0	Lunes	2	32.2802492	Patient 6	24.8359	2	walkIn	79.366	5.040	6.360	40.682	0.376	0.000	26.908	doctor 1	doctor 1
8	0	Lunes	2	41.3982407	Patient 8	9.1180	1	Ambulance	53.446	0.000	0.000	13.814	1.055	0.000	38.577	doctor 6	doctor 1
9	0	Lunes	2	58.9729852	Patient 7	17.5747	3	Ambulance	90.915	0.000	0.000	33.844	8.049	0.000	49.022	doctor 7	doctor 1
10	0	Lunes	5	11.4318759	Patient 10	132.4589	3	Ambulance	77.626	0.000	0.000	34.853	1.654	0.000	41.119	doctor 1	doctor 1
11	0	Lunes	5	18.2742103	Patient 9	6.8423	2	walkIn	85.892	6.026	6.727	27.160	7.553	0.000	38.426	doctor 6	doctor 1
12	0	Lunes	6	16.2677339	Patient 11	57.9935	2	walkIn	43.684	4.678	6.515	18.173	14.317	0.000	0.000	doctor 6	doctor 1
13	0	Lunes	7	1.95792408	Patient 12	45.6902	2	walkIn	61.37	6.090	5.096	21.304	8.939	19.942	0.000	doctor 1	doctor 1
14	0	Lunes	8	17.4089876	Patient 13	75.4511	2	walkIn	36.847	6.115	7.623	11.913	11.197	0.000	0.000	doctor 2	doctor 2
15	0	Lunes	9	17.6958039	Patient 14	60.2868	2	walkIn	95.287	7.197	6.015	24.943	2.049	0.000	50.390	doctor 1	doctor 2
16	0	Lunes	9	33.7034427	Patient 17	16.0076	2	walkIn	69.711	4.534	6.648	11.284	10.463	0.000	36.782	doctor 7	doctor 2
17	0	Lunes	9	49.3696961	Patient 15	15.6663	2	Ambulance	107.711	0.000	0.000	33.434	27.293	0.000	46.984	doctor 6	doctor 2
18	0	Lunes	9	50.6044146	Patient 19	1.2347	2	walkIn	54.055	5.158	6.173	13.801	1.235	23.747	0.000	doctor 7	doctor 3
19	0	Lunes	9	53.1471747	Patient 22	2.5428	2	walkIn	29.792	4.001	5.253	9.642	7.513	0.000	0.000	doctor 3	doctor 3
20	0	Lunes	10	12.2081131	Patient 20	19.0609	2	walkIn	67.912	5.225	4.442	49.845	8.400	0.000	0.000	doctor 2	doctor 2

- ¿Cómo se distribuye el tiempo entre llegadas? (¿se ve afectado por otra variable?, ej: día, modo de llegada, triage)
- ¿Cómo se distribuye el tiempo de evaluación? (triage, día, modo de llegada?)

Análisis exploratorio de los datos

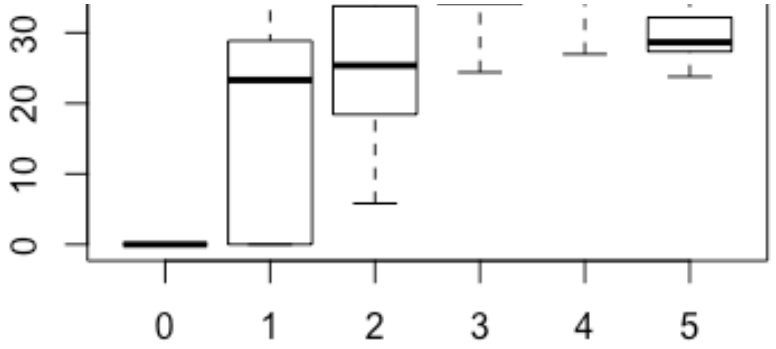






Homogeneidad

```
1  library(readxl)
2  hosp <- read_excel("C:/Users/pescuder/Desktop/DatosHospitalDescriptiva.xlsx")
3  View(hosp)
4  tev = hosp$`Tiempo Evaluación  (minutos)`
5  triage = hosp$TriageNum
6  boxplot(tev ~ triage, data = hosp)
7  kruskal.test(tev ~ triage, data = hosp)
8
```




```
        Kruskal-Wallis rank sum test

data:  tev by triage
Kruskal-Wallis chi-squared = 786.37, df = 5, p-value < 2.2e-16
```

?

Tarea

Leer:

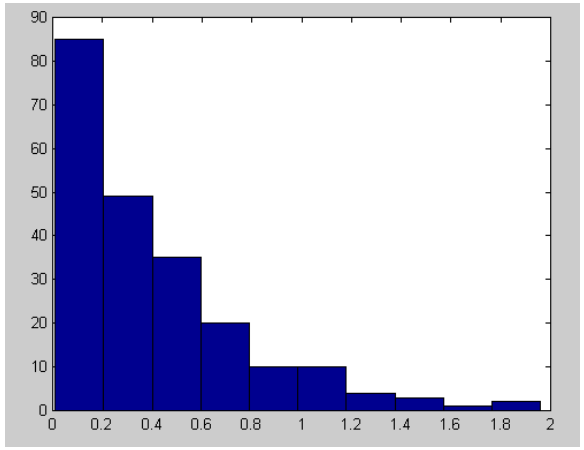
- Law, A. M. (2013, December). A tutorial on how to select simulation input probability distributions. In *2013 Winter Simulations Conference (WSC)* (pp. 306-320). IEEE. https://www.informs-sim.org/wsc16papers/012.pdf

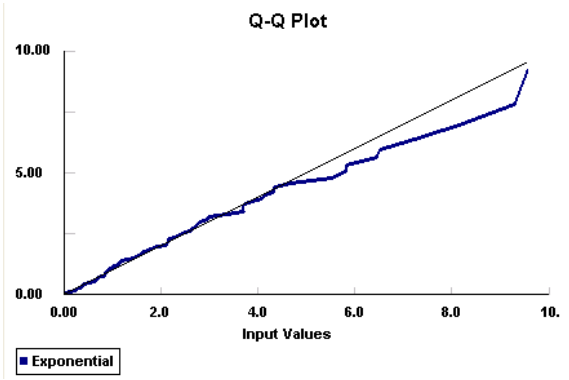
AJUSTE DE DISTRIBUCIONES TEÓRICAS

- Razones para ajustar o para no ajustar distribuciones teóricas a los datos de entrada:
  - Probar supuestos.
  - Para no usar directamente los datos históricos (son pocos, independencia entre replicaciones, valores raros que se perderían). Para validar si es posible y algunas veces recomendable.
  - Es recomendable usar datos de distribuciones empíricas solo en modelos iniciales o generales, o cuando no es posible ajustar distribución alguna.
  - Es mas fácil cambiar una distribución teórica
  - Es más fácil de manejar una distribución teórica dentro del modelo de simulación.
  - Truncar las distribuciones teóricas es recomendable algunas veces

Pasos para el ajuste de distribuciones teóricas

- Paso 1: Hipótesis sobre el tipo de distribución muy útil un histograma y la utilización de algunos resultados teóricos o prácticos disponibles en la literatura.





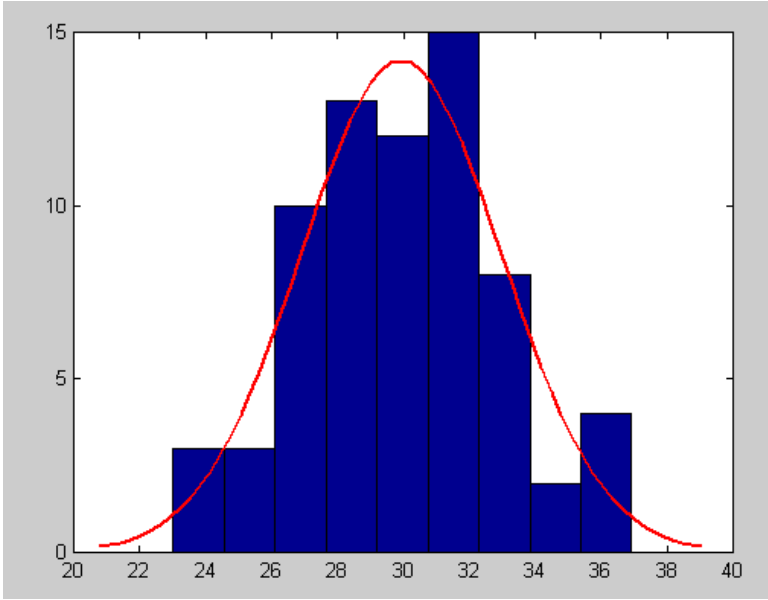
- Paso 2: Estimación de los parámetros de la distribución.
- Paso 3: Prueba de Bondad de Ajuste.

Pruebas de Bondad de Ajuste

- Con ***n*** muy pequeño o moderadamente pequeño la potencia de la prueba es poca. Con ***n*** muy grande es muy probable que se rechace la hipótesis nula.
- No rechazar una distribución no significa aceptar la hipótesis como verdadera. Proceso sistemático para detectar diferencias marcadas.
- Para evaluar la bondad del ajuste de las distribuciones teóricas a los datos, se usan generalmente algunos Test como el Test Chi-Cuadrado y el Test Kolmogorov Smirnov, pero es mas usado el Valor P.
- Un Valor P pequeño indica que podría rechazarse Ho, inversamente valores P grandes indican que la Hipotesis nula no debería ser aceptada. Cuando se compara el ajuste de dos distribuciones diferentes, la distribución con el mayor Valor P es la que muestra el mejor ajuste.
- En términos generales: el Valor P o nivel de significancia alcanzado es el nivel mínimo de significancia para el cual los datos observados indican que se debe rechazar la hipótesis nula. Si el valor p es menor que el nivel de confianza (α), H0 se rechaza. Si el valor p es mayor que el nivel de confianza, H0 no se rechaza

Pruebas de Bondad de Ajuste: Prueba Chi Cuadrado (1/2)

- Puede verse como la superposición de un histograma y la distribución que se quiere ajustar.
- Datos continuos o discretos.
- Cualquier distribución.
- Se requiere agrupación de los datos en intervalos (tamaño y número de intervalos).
- Valida para muestras grandes.



Pruebas de Bondad de Ajuste: Prueba Chi Cuadrado (1/2)

H_0:la distribución de los datos generados es idéntica a la distribución teórica f

- Pasos:
  1. Se organizan las *n* observaciones en *k* intervalos de clase.
  2. Se calcula el siguiente Test:

$$\chi_0^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

- Donde O_i es la frecuencia observada en el i-ésimo intervalo, E_i es la frecuencia esperada en ese intervalo y se calcula como np_i , donde p_i es la probabilidad asociada con el i-ésimo intervalo dependiendo de la distribución planteada en la hipótesis inicial.

  3. Se rechaza la hipótesis nula si $\chi_0^2 > \chi_{k-1,1-\alpha}^2$

Tiempo entre llegada (min)
0.36
4.53
7.33
2.45
11.41
0.80
12.71
16.07
0.44
2.60
0.34
3.92
0.13
4.70
0.83
1.12
0.55
5.83
11.61
2.27
3.70
5.87
5.05
4.07
7.43
0.55
2.15
3.71
10.64
13.20
0.21
9.69
0.05

Promedio	5.000
Desviación estándar	4.446
Mín	0.05
Max	17.28

H0: La variable Tiempo entre llegadas se distribuye exponencialmente con $\lambda = 1/5$

$$\int_a^b \frac{1}{5}\exp\left(-\frac{x}{5}\right)dx = e^{-a/5} - e^{-b/5}$$

Intervalo	Lim Inf	Lim Sup	Frecuencia Observada (O_i)	Frecuencia relativa Observada	Exponencial $f(x) = \lambda e^{-\lambda x}$	Frecuencia Esperada (E_i)	$(O_i-E_i)^2/E_i$
1	0	2.5	21	0.420	0.393	19.673	0.089
2	2.50	5.0	10	0.200	0.239	11.933	0.313
3	5.00	7.5	9	0.180	0.145	7.237	0.429
4	7.50	10.0	3	0.060	0.088	4.390	0.440
5	10.00	12.5	3	0.060	0.053	2.663	0.043
6	12.50	15.0	2	0.040	0.032	1.615	0.092
7	15.00	17.5	2	0.040	0.020	0.979	
		Total	50		0.969803	48.490	1.406
		λ	1/5				

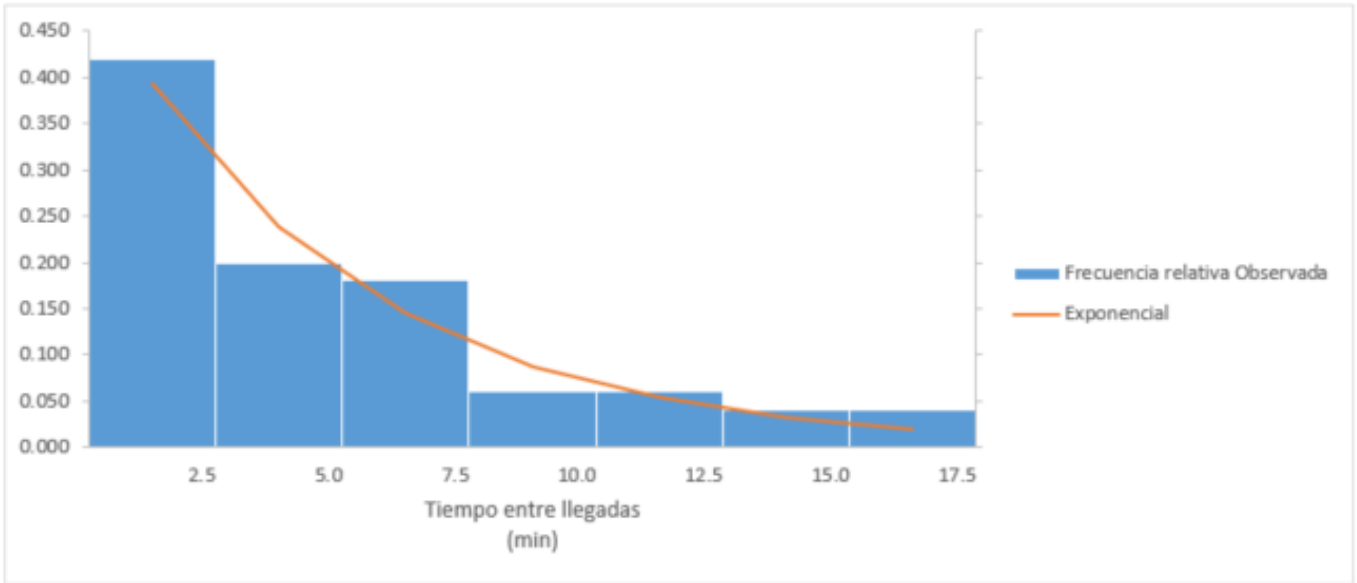
k= 7

α= 0.05

$\chi^2_{k-1,1-\alpha}$= 12.59159

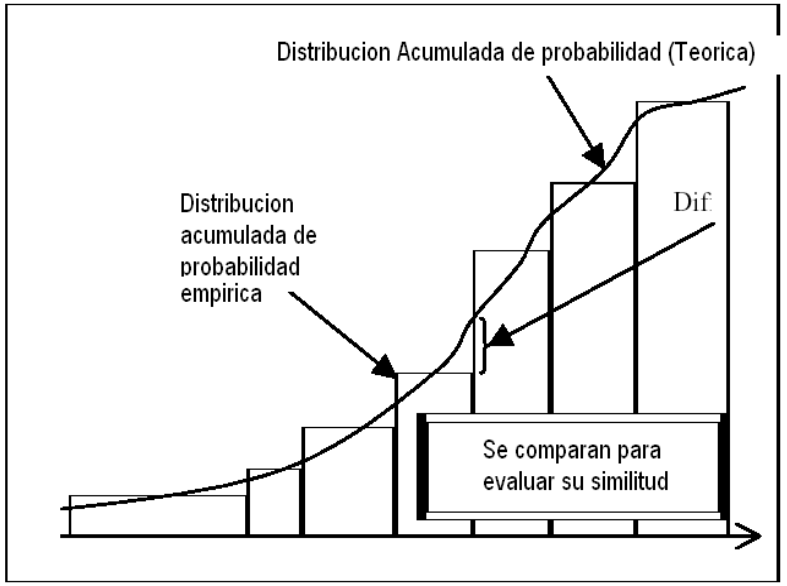
χ^2_0= 1.406

Se rechaza la hipótesis nula si $\chi^2_0 > \chi^2_{k-1,1-\alpha}$



Pruebas de Bondad de Ajuste Prueba Kolmogorov-Smirnov (K-S) (1/2):

- Puede verse como la comparación entre la Distribución Acumulada de Probabilidad Obtenida con los datos y la de la distribución teórica considerada.
- No requieren agrupación. No requiere intervalos.
- Datos continuos.
- Valida para cualquier tamaño de muestra.
- Para darle mejor peso a los valores de las diferencias (en especial de las colas) se usa la prueba A-D (ver ayuda en el Stat Fit)



Pruebas de Bondad de Ajuste Prueba Kolmogorov-Smirnov (K-S) (1/2):

- Esta basado en la Máxima desviación entre la distribución acumulada *F(x)* y por la distribución *S(x)* así:

$$D = Max\ (D^{+}, D^{-})$$

$$D^{+} = Max(\ i/n - F(x)),\ i=1,..,n$$

$$D^{-} = Max(F(x)-\ (i-1)/n\ ),\ i=1,..,n$$

Donde D es el estadístico KS, y *x* es el valor del i-ésimo punto.

Tarea

- Generar 100 datos de una distribución Exponencial y hacer las pruebas de bondad de ajuste Chi-Cuadrado y Kolmogorov-Smirnov (paso a paso, sin usar software de ajuste de distribuciones)
- Consultar Prueba de bondad de ajuste Anderson y Darling

Selección de distribuciones en ausencia de datos históricos

- Usar distribuciones Beta, Triangular o Uniforme
  - Beta: es difícil de estimar el valor de sus parámetros, usar una distribución sesgada a la derecha (muy común en la práctica).
  - Triangular: estimación del mínimo, máximo y moda.
  - Uniforme: si solo se sabe el rango en el que puede variar la variable aleatoria.

    Consulta: Utilización de las distribuciones Triangular y Beta en ausencia de datos históricos (porque usarlas, como estimar los parámetros de la distribución triangular).

- Usar resultados conocidos de la literatura y de la naturaleza del proceso.
  - Tiempos entre llegadas a un sistema de servicio (Exp)
  - Tiempos de Atención de médicos (Lognormal)
  - Tiempo de Ejecución de una etapa (Análisis PERT) (Beta)
  - Ciclos de vida de productos (Log-logistics)
  - Ver Law (2007) para más ejemplos

Consulta

- ¿Qué hacer cuando no hay datos existentes?
  - Kuhl, M. E., Steiger, N. M., Lada, E. K., Wagner, M. A., & Wilson, J. R. (2006, December). Introduction to modeling and generating probabilistic input processes for simulation. In *Proceedings of the 2006 Winter simulation Conference* (pp. 19-35). IEEE. https://www.informs-sim.org/wsc08papers/008.pdf (Leer sección: *Fitting Beta Distributions to Data or Subjective Information*)
  - Law, A. M. (2013, December). A tutorial on how to select simulation input probability distributions. In *2013 Winter Simulations Conference (WSC)* (pp. 306-320). IEEE. https://www.informs-sim.org/wsc16papers/012.pdf
  - Biller, B., & Gunes, C. (2010, December). Introduction to simulation input modeling. In *Proceedings of the 2010 Winter Simulation Conference* (pp. 49-58). IEEE.https://www.informs-sim.org/wsc10papers/006.pdf
  - Banks, Jerry, ed. *Handbook of simulation*. New York: Wiley, 1998. (Biblioteca Digital EAFIT Capítulo 3, section 3.7)

- Material de apoyo para esta unidad:
  - Robinson, Stewart. Simulation: the practice of model development and use. Palgrave Macmillan, 2014 (capítulo 7).
  - Law, Averill M. "A tutorial on how to select simulation input probability distributions." *Simulation Conference (WSC), Proceedings of the 2013 Winter Simulation Conference*. IEEE, 2013.
  - Banks, Jerry, ed. *Handbook of simulation*. New York: Wiley, 1998. (Biblioteca Digital EAFIT Capítulo 3)
  - Law, A.M. “Simulation Modeling and Analysis”, Fourth Edition, New York:McGraw-Hill, 2007 (Capítulo 6- en Biblioteca).