OCW-UPM Estadística para Ingeniería Civil y Medioambiental

Autores: E. M. García del Toro, C. Hermoso, E. J. Huertas

LABORATORIO DE ESTADÍSTICA CON MATLAB

PRÁCTICA 1 - Estadística descriptiva con MATLAB y diagramas de caja y bigotes (box-plot)

Parte A - Estadística Descriptiva con MATLAB

Una de las etapas más importantes en el proceso de investigación de cierto fenómeno se relaciona con la sistematización y análisis de la información y se denomina *análisis estadístico de la información*. Esta etapa comprende la recolección, análisis, interpretación y presentación de los datos que se poseen relativos a dicho fenómeno.

Dada una colección o serie de datos, se pueden representar gráficamente, calcular las medidas de tendencia central: media, mediana, moda, media aritmética, calcular las medidas de dispersión como: rango, varianza, desviación típica, coeficiente de variación de Pearson, etc, con el fin último de obtener información útil de los mencionados datos, y por ende comprender e identificar las leyes que guían o regulan los fenómenos estudiados.

Dada la siguiente serie de datos de precipitaciones anuales (annual rainfalls, en cm de pluviómetro), de cierta ciudad española durante los últimos 60 años

X=[40 33 28 25 11 21 22 17 22 19 17 16 28 26 20 15 21 20 19 24 10 29 23 34 24 33 26 14 13 18 28 23 28 21 29 24 11 31 25 18 25 26 20 34 22 30 27 32 35 39 18 29 16 37 28 29 10 34 29 38]

con ayuda de MATLAB, sin agrupar los datos:

a) Calcular la media aritmética, la mediana, y la moda.

b) Calcular el rango, la cuasivarianza, la varianza, la cuasidesviación típica, la desviación típica, y el coeficiente de variación de Pearson.

c) Calcular los cuartiles 1, 2 y 3, el rango intercuartílico, los percentiles 10, 25, 50 y 80, e interpretar estos resultados.

d) Estudiar la asimetría o sesgo de la distribución mediante el coeficiente de asimetría (skewness) y comparando media, mediana y moda.

e) Estudiar el apuntamiento o curtosis (kurtosis) de la distribución de datos.

A continuación, vamos a agrupar agrupar los datos anteriores en 6 clases mediante el comando tabulate. Matlab forma los intervalos o clases abiertos por la izquierda y cerrados por la derecha. Si ejecutamos `>>tabulate(X)`, podremos contar el número de ocurrencias en cada clase, para elaborar la siguiente tabla de frecuencias

clase i	clase 1	clase2	clase 3	clase 4	clase 5	clase 6
límites	[10,15]	(15,20]	(20,25]	(25,30]	(30,35]	(35,40]
n_i	6	10	14	17	8	5

y a continuación hacer varias representaciones gráficas de los datos agrupados.

Parte B - Diagrama de caja y bigotes (box-plot)

Los diagramas de caja (de caja y bigotes o box-plot) se utilizan para mostrar la distribución de una muestra (o muestras) de datos. Hay varias definiciones diferentes de la gráfica de caja y de los bigotes, y la presentación suele variar bastante en función del software que se utiliza para generarlos. La principal diferencia reside en la consideración de los valores atípicos, que pueden clasificarse como atípicos moderados (usualmente los que se encuentran fuera del intervalo $(Q_1 - 1.5 * IQR, Q_3 + 1.5 * IQR)$), o atípicos fuertes (los que se encuentran fuera del intervalo $(Q_1 - 3 * IQR, Q_3 + 3 * IQR)$).

Realice el correspondiente diagrama de caja y bigotes de los datos anteriores utilizando MATLAB y comente si hay datos que pudieran considerarse atípicos.

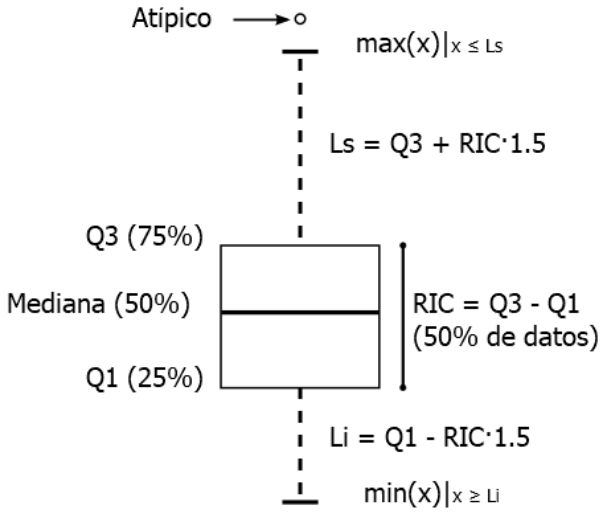

Figure 1: Diagrama de caja y bigotes (Fuente: Wikipedia [1])

Trabajo del alumno 1

Se seleccionan tres tipos diferentes de cable de acero trenzado, y se determina el límite de fatiga, en MPa, para cada muestra con los siguientes resultados:

tipo1=[350 350 350 358 370 370 370 371 371 372 372 384 391 391 392]
tipo2=[350 354 359 363 365 368 369 371 373 374 376 380 383 388 392]
tipo3=[350 361 362 364 364 365 366 371 377 377 377 379 380 380 392]

con la ayuda de MATLAB:

1. Analice de forma descriptiva los datos y obtenga conclusiones e información útil de los datos.
2. Compare semejanzas y diferencias entres los tres tipos de cable de acero.
3. ¿Hay ensayos cuyo resultado pudiera considerarse atípico? ¿Cuales?

Trabajo del alumno 2

Los siguientes datos son los caudales máximos anuales en m^3/s en el Río Colorado en Black Canyon (EEUU) durante el período de 52 años desde 1878 a 1929 (ver [1], problema 1.3)

Y=[1980 1130 3120 2120 1700 2550 8500 3260 3960 2270 1700 1570 2830 2120 2410 2550 1980 2120 2410 2410 1420 1980 2690 3260 1840 2410 1840 3120 3290 3170 1980 4960 2120 2550 4250 1980 4670 1700 2410 4550 2690 2270 5660 5950 3400 3120 2070 1470 2410 3310 3230 3090]

Para ambos problemas, con la ayuda de MATLAB:

1. Analice de forma descriptiva los datos y obtenga conclusiones e información útil sobre los caudales del Rio Colorado en el periodo señalado.
2. ¿Hubo años en que se registraron caudales atípicos? ¿Cuáles?

Referencias:

[1] N. T. Kottegoda, R. Rosso, *Applied Statistics for Civil and Environmental Engineers*, 2nd. Ed., Aug 2008, Wiley-Blackwell. ISBN: 978-1-405-17917-1

[2] **https://es.wikipedia.org/wiki/Diagrama_de_caja**