Técnicas de Análisis de Datos

Actividad 1: Estadística Descriptiva

Máster en Gestión y Análisis de Grandes Volúmenes de Datos:

Big Data

29/11/2021

Cristina Varas Menadas

Índice

1. Pregunta 1

Se mide los niveles de una enzima que está presente en una muestra de agricultores expuestos a insecticidas agrícolas, obteniéndose los siguientes datos:

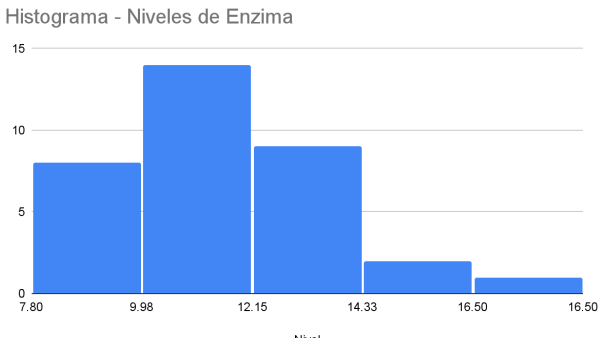
Individuo	Nivel	Individuo	Nivel	Individuo	Nivel
1	10'6	13	12'2	25	11'8
2	12'5	14	10'8	26	12'7
3	11'1	15	16'5	27	11'4
4	9'2	16	15'0	28	9'3
5	11'5	17	10'3	29	8'6
6	9'9	18	12'4	30	8'5
7	11'9	19	9'1	31	10'1
8	11'6	20	7'8	32	12'4
9	14'9	21	11'3	33	11'1
10	12'5	22	12'3	34	10'2
11	12'5	23	9'7		
12	12'3	24	12'0		

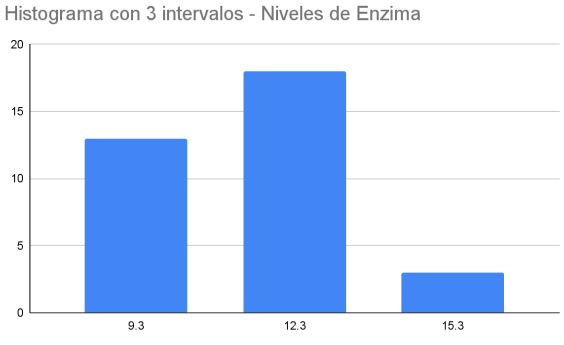
Los valores normales oscilan entre 8 y 18 U/ml (unidades por mililitro). Para esta muestra y utilizando Excel, Jasp y RStudio:

a) Represente un histograma.

Excel (GSheet)

Se ha dibujado el histograma sin especificar los intervalos y a continuación el histograma calculando sus intervalos.





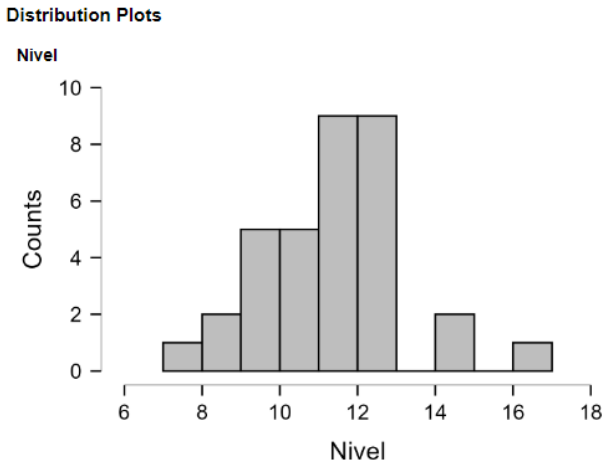
Para representar el histograma por intervalos se han hecho los siguientes cálculos:

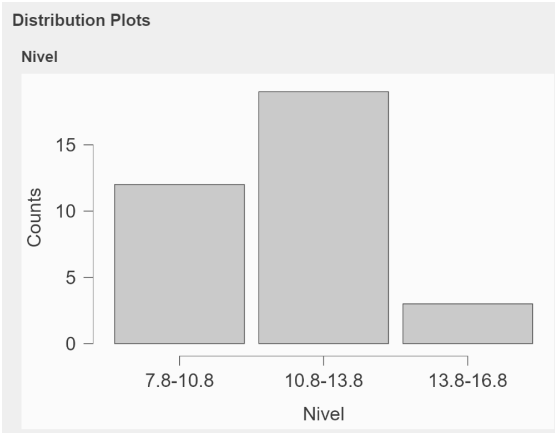
		Función utilizada
Recorrido de la variable	8.7	MAX(B2:B35) - MIN(B2:B35)
Número de intervalos (√N)	3	ROUND(1+(1.33*LOG10(COUNTA(A2:A35))))
Ancho de la clase	3	ROUND(E25/E26)

*Se pueden ver los cálculos en el archivo excel adjunto en la entrega.

Jasp

Histograma sin intervalos y con intervalos.

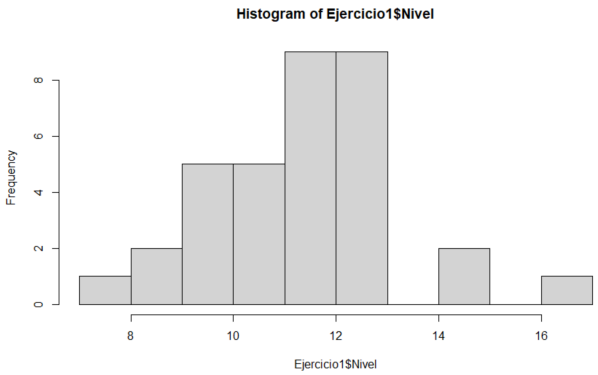




RStudio

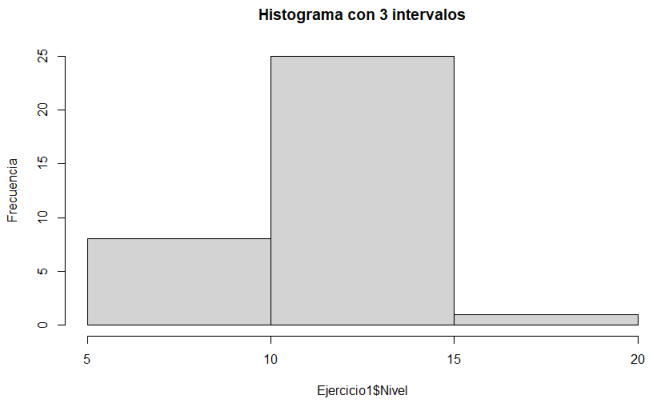
Histograma sin definir los intervalos:

```
hist(x = Ejercicio1$Nivel)
```



Histograma definiendo los intervalos:

```
hist(Ejercicio1$Nivel, breaks = 2, main = "Histograma con 3 intervalos",
ylab = "Frecuencia")
```



b) Determine los estadísticos media, mediana y desviación típica.

Excel (GSheet)

Media: `=AVERAGE(B2:B3)`

Mediana: `=MEDIAN(B2:B35)`

Desviación típica: `=STDEV(B2:B35)`

Media	11.35294118
Mediana	11.45
Desviación típica	1.874588487

Jasp

Descriptive Statistics

	Nivel
Valid	34
Missing	0
Mode	12.500
Median	11.450
Mean	11.353
Std. Deviation	1.875
Variance	3.514
Minimum	7.800
Maximum	16.500

RStudio

```
> mean(Ejercicio1$Nivel, na.rm = TRUE)
[1] 11.35294
> median(Ejercicio1$Nivel, na.rm = TRUE)
[1] 11.45
> sd(Ejercicio1$Nivel, na.rm = TRUE)
[1] 1.874588
```

También es posible calcular todos estos estadísticos con la siguiente línea de código:

```
> summary(Ejercicio1$Nivel)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   7.80   10.12   11.45   11.35   12.38   16.50
```

c) Realice análisis de los resultados obtenidos.

En el histograma representado con las diferentes herramientas podemos observar una representación gráfica del grupo de datos estadísticos sobre los niveles de enzimas agrupados en intervalos numéricos.

Podemos ver que en el intervalo de 10-15 se encuentra la frecuencia más alta de niveles de enzimas, más concretamente entre las enzimas de niveles 11-13. Se puede decir que la gráfica forma cierta similitud a una campana de Gauss, acumulándo los niveles en valores centrales y siendo más bajos en los extremos.

La media, siendo el valor promedio de un conjunto de datos numéricos, es 11,35. Es acertado, pues como decía, la mayoría de los datos están en el intervalo de 11 a 13.

La mediana, siendo el valor que ocupa el lugar central de todos los datos cuando estos están ordenados de menor a mayor, es 11,45. Los valores van de 8 a 17 aproximadamente, por lo que es correcto que 11,45 sea el valor central.

La desviación típica, siendo la separación existente entre cualquier valor de la serie y la media aritmética de todos los datos de dicha serie, es 1,87.

2. Pregunta 2

Actualmente, se está estudiando en las distintas comunidades autónomas el número de hijos por familia para estudiar la natalidad. Uno de los trabajadores que está haciendo las encuestas, recoge los datos de su barrio donde hay 100 familias. Ha obtenido los siguientes datos que aparecen en la tabla siguiente:

1	3	3	0	4	3	1	4	0	0
2	1	0	3	1	2	1	4	1	2
3	3	4	2	0	4	3	0	2	3
1	3	4	2	2	4	4	4	2	1
4	2	1	1	0	1	1	2	3	0
3	3	3	1	1	3	3	0	2	3
4	3	0	3	1	2	2	1	2	3
3	2	1	3	1	3	4	4	4	1
3	0	3	1	0	4	3	2	3	2
1	2	0	2	0	0	2	2	3	4

Para esta muestra y utilizando Excel, Jasp y RStudio:

a) Represente la tabla de frecuencias correspondiente.

Excel

```
FREQUENCY(A1:J10,A13:A17)
```

Número de hijos	Número de familias (frecuencia absoluta)	Frecuencia relativa	Frecuencia absoluta acumulada	Frecuencia relativa acumulada
0	15	0.15	15	0.15
1	21	0.21	36	0.36
2	21	0.21	57	0.57
3	27	0.27	84	0.84
4	16	0.16	100	1
	0	1		
Total	100			0

A continuación se presenta la tabla de frecuencias por intervalos.

Cálculo del número de intervalos:

		Función utilizada
Recorrido de la variable	4	MAX(A1:J10) - MIN(A1:J10)
Número de intervalos (√N)	4	ROUND(1+(1.33*LOG10(COUNTA(A1:J10))))
Ancho de la clase	1	ROUND(B26/B27)

	Intervalos						
Intervalos	min	max	Marca de la clase, ci	Frecuencia absoluta	Frecuencia relativa	Frecuencia absoluta acumulada	Frecuencia relativa acumulada
[0, 1]	0	1	0.5	36	0.36	36	0.36
(1, 2]	1	2	1.5	21	0.21	57	0.57
(2, 3]	2	3	2.5	27	0.27	84	0.84
(3, 4]	3	4	3.5	16	0.16	100	1
				0			
Totales				100	1		2.77

Realmente la tabla resultante con intervalos aporta la misma información que la tabla con el análisis en intervalos, pues los datos son enteros entre 1 y 4.

RStudio

```
tabla_frecuencias ← as.data.frame(table(cyl = Ejercicio2))
```

```
> transform(tabla_frecuencias,
+           FreqRelativa = round(prop.table(tabla_frecuencias$Freq),3),
+           FreqAbsolutaAcumulada = cumsum(tabla_frecuencias$Freq),
+           FreqRelativaAcumulada = round(cumsum(prop.table(tabla_frecuencias$Freq)),3))
  cyl Freq FreqRelativa FreqAbsolutaAcumulada FreqRelativaAcumulada
1   0   15         0.15                    15                  0.15
2   1   21         0.21                    36                  0.36
3   2   21         0.21                    57                  0.57
4   3   27         0.27                    84                  0.84
5   4   16         0.16                   100                  1.00
```

Jasp

Frequencies for V1

V1	Frequency	Percent	Valid Percent	Cumulative Percent
0	15	15.152	15.152	15.152
1	20	20.202	20.202	35.354
2	21	21.212	21.212	56.566
3	27	27.273	27.273	83.838
4	16	16.162	16.162	100.000
Missing	0	0.000		
Total	99	100.000		

b) Determine los estadísticos media y la mediana.

Excel

Media	2.08
Mediana	2

RStudio

```
> media <- mean(Ejercicio2$V1, na.rm = TRUE)
> media
[1] 2.08
> mediana <- median(Ejercicio2$V1, na.rm = TRUE)
> mediana
[1] 2
```

Jasp

Descriptive Statistics

	V1
Valid	89
Missing	0
Mode	3.000
Median	2.000
Mean	2.079
Std. Deviation	1.316
Minimum	0.000
Maximum	4.000

c) Realice análisis de los resultados obtenidos. (tanto con los datos absolutos como con los datos acumulados).

Se puede observar que es mayor el número de familias que tienen 3 hijos. Sin embargo, la media y la mediana se colocan en torno a los 2 hijos por tratarse también muy común encontrar familias con 2 hijos.

3. Pregunta 3

60; 66; 77; 70; 66; 68; 57; 70; 66; 52; 75; 65; 69; 71; 58; 66; 67; 74; 61;
63; 69; 80; 59; 66; 70; 67; 78; 75; 64; 71; 81; 62; 64; 69; 68; 72; 83; 56;
65; 74; 67; 54; 65; 65; 69; 61; 67; 73; 57; 62; 67; 68; 63; 67; 71; 68; 76;
61; 62; 63; 76; 61; 67; 67; 64; 72; 64; 73; 79; 58; 67; 71; 68; 59; 69; 70;
66; 62; 63; 66;

Calcule:

a) Obtenga la tabla de frecuencias correspondiente según los criterios correspondientes. Haga análisis de resultados (tanto absoluto como acumulado).(con Excel y RStudio)

Excel

Tabla de frecuencias sin intervalos:

Peso	Número de personas (frecuencia absoluta)	Frecuencia relativa	Frecuencia acumulada	Frecuencia relativa acumulada
60	1	0.0125	1	0.0125
66	1	0.0125	2	0.025
77	1	0.0125	3	0.0375
70	2	0.025	5	0.0625
68	2	0.025	7	0.0875
57	2	0.025	9	0.1125
52	1	0.0125	10	0.125
75	4	0.05	14	0.175
65	4	0.05	18	0.225
69	4	0.05	22	0.275
71	4	0.05	26	0.325
58	4	0.05	30	0.375
67	7	0.0875	37	0.4625
74	9	0.1125	46	0.575
61	5	0.0625	51	0.6375
63	5	0.0625	56	0.7
80	4	0.05	60	0.75
59	4	0.05	64	0.8

78	2	0.025	66	0.825
64	2	0.025	68	0.85
81	2	0.025	70	0.875
62	2	0.025	10	0.125
72	2	0.025	74	0.925
83	1	0.0125	75	0.9375
56	1	0.0125	76	0.95
54	1	0.0125	77	0.9625
73	1	0.0125	78	0.975
76	1	0.0125	79	0.9875
79	1	0.0125	80	1
	0			
1970	80			

Tabla de frecuencias con intervalos:

	Intervalos						
Intervalos	min	max	Marca de la clase, ci	Frecuencia absoluta	Frecuencia relativa	Frecuencia absoluta acumulada	Frecuencia relativa acumulada
[52, 59]	52	59	55.5	9	0.1125	9	0.1125
(59, 67]	59	67	63	37	0.4625	46	0.575
(67, 75]	67	75	71	26	0.325	72	0.9
(75, 83]	75	83	79	8	0.1	80	1
				0			
Totales				80	1		

RStudio

Tabla de frecuencias sin intervalos:

```
tabla_frecuencias2 ← as.data.frame(table(cyl = Ejercicio3))
```

```
> transform(tabla_frecuencias2,
+           FreqRelativa = round(prop.table(tabla_frecuencias2$Freq),3),
+           FreqAbsolutaAcumulada = cumsum(tabla_frecuencias2$Freq),
+           FreqRelativaAcumulada = round(cumsum(prop.table(tabla_frecuencias2$Freq)),3))
   cyl Freq FreqRelativa FreqAbsolutaAcumulada FreqRelativaAcumulada
1   52    1        0.012                     1                 0.012
2   54    1        0.012                     2                 0.025
3   56    1        0.012                     3                 0.037
4   57    2        0.025                     5                 0.062
5   58    2        0.025                     7                 0.086
6   59    2        0.025                     9                 0.111
7   60    1        0.012                    10                 0.123
8   61    4        0.049                    14                 0.173
9   62    4        0.049                    18                 0.222
10  63    4        0.049                    22                 0.272
11  64    5        0.062                    27                 0.333
12  65    3        0.037                    30                 0.370
13  66    7        0.086                    37                 0.457
14  67    9        0.111                    46                 0.568
15  68    5        0.062                    51                 0.630
16  69    5        0.062                    56                 0.691
17  70    4        0.049                    60                 0.741
18  71    4        0.049                    64                 0.790
19  72    2        0.025                    66                 0.815
20  73    3        0.037                    69                 0.852
21  74    2        0.025                    71                 0.877
22  75    2        0.025                    73                 0.901
23  76    2        0.025                    75                 0.926
24  77    1        0.012                    76                 0.938
25  78    1        0.012                    77                 0.951
26  79    1        0.012                    78                 0.963
27  80    1        0.012                    79                 0.975
28  81    1        0.012                    80                 0.988
29  83    1        0.012                    81                 1.000
```

Tabla de frecuencias por intervalos:

```
> # Tabla de frecuencias con 3 intervalos
> niveles <- Ejercicio3$V1
> frecuencias <- as.data.frame(table(clases=factor(cut(niveles, breaks = 4))))
>
> # Tabla de frecuencias completa con 3 intervalos
> tabla_frecuencias_intervalos = transform(frecuencias,
+                                          FrecAcumulada = cumsum(Freq),
+                                          FrecRelativa = round(Freq/10, 3),
+                                          FrecRelativaAcumulada = round(cumsum(Freq/10), 3))
```

```
> frecuencias
       clases Freq
1   (52,59.8]    9
2 (59.8,67.5]   37
3 (67.5,75.2]   27
4   (75.2,83]    8
> tabla_frecuencias_intervalos
       clases Freq FrecAcumulada FrecRelativa FrecRelativaAcumulada
1   (52,59.8]    9             9          0.9                    0.9
2 (59.8,67.5]   37            46          3.7                    4.6
3 (67.5,75.2]   27            73          2.7                    7.3
4   (75.2,83]    8            81          0.8                    8.1
```

Jasp

Tabla de frecuencias sin intervalos:

Frequency Tables

Frequencies for V60

V60	Frequency	Percent	Valid Percent	Cumulative Percent
52	1	1.250	1.250	1.250
54	1	1.250	1.250	2.500
56	1	1.250	1.250	3.750
57	2	2.500	2.500	6.250
58	2	2.500	2.500	8.750
59	2	2.500	2.500	11.250
61	4	5.000	5.000	16.250
62	4	5.000	5.000	21.250
63	4	5.000	5.000	26.250
64	5	6.250	6.250	32.500
65	3	3.750	3.750	36.250
66	7	8.750	8.750	45.000
67	9	11.250	11.250	56.250
68	5	6.250	6.250	62.500
69	5	6.250	6.250	68.750
70	4	5.000	5.000	73.750
71	4	5.000	5.000	78.750
72	2	2.500	2.500	81.250
73	3	3.750	3.750	85.000
74	2	2.500	2.500	87.500
75	2	2.500	2.500	90.000
76	2	2.500	2.500	92.500
77	1	1.250	1.250	93.750
78	1	1.250	1.250	95.000
79	1	1.250	1.250	96.250
80	1	1.250	1.250	97.500
81	1	1.250	1.250	98.750
83	1	1.250	1.250	100.000
Missing	0	0.000		
Total	80	100.000		

Tabla de frecuencias con intervalos:

Frequency Tables ▼

Frequencies for V60 ▼

V60	Frequency	Percent	Valid Percent	Cumulative Percent
52-59	9	11.250	11.250	11.250
59-67	36	45.000	45.000	56.250
67-75	27	33.750	33.750	90.000
75-83	8	10.000	10.000	100.000
Missing	0	0.000		
Total	80	100.000		

b) Calcule el porcentaje de personas de peso menor que 65 Kg.

Excel

% personas con <65 kg `=COUNTIF(A1:A80, "<=65")= 30`

c) ¿Cuántas personas tienen peso mayor o igual que 70 Kg? pero menor que 85?

Excel

Número de personas con un peso entre 70 (incluido) y 85 `= COUNTIFS(A1:A80, ">=70", A1:A80, "<85")= 25`

Número de personas con <65 kg	30
Número de personas con un peso entre 70 (incluido) y 85	25

d) Calcule los estadísticos con RStudio y compárelos con los obtenidos vía Excel.

RStudio

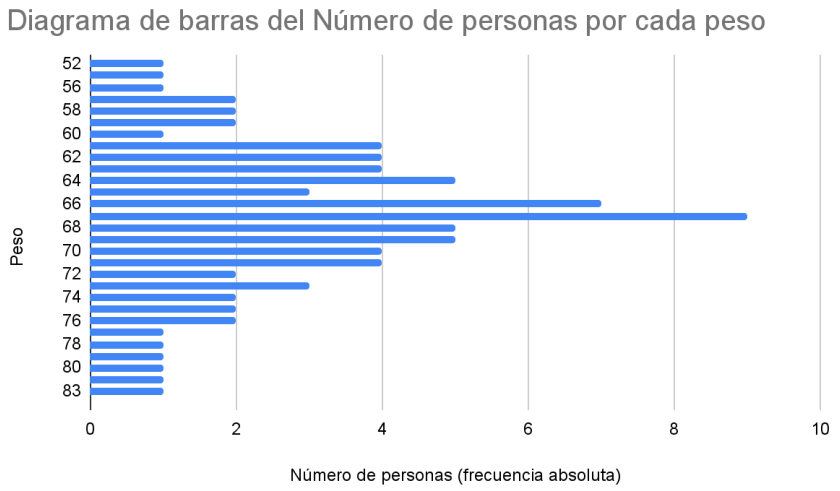
```
> Ejercicio3[Ejercicio3$V1 <= 65,]
 [1] 60 57 52 65 58 61 63 59 64 62 64 56 65 54 65 61 57 62 63 61 62 63 61 64 64 64 58 59 62 63

> Ejercicio3[Ejercicio3$V1 >= 70 & Ejercicio3$V1 < 85,]
 [1] 77 70 70 75 71 74 80 70 78 75 71 81 72 83 74 73 71 76 76 72 73 73 79 71 70
```

```
> grupo_pesos1 <- Ejercicio3[Ejercicio3$V1 <= 65,]
> grupo_pesos2 <- Ejercicio3[Ejercicio3$V1 >= 70 & Ejercicio3$V1 < 85,]
> length(grupo_pesos1)
[1] 30
> length(grupo_pesos2)
[1] 25
```

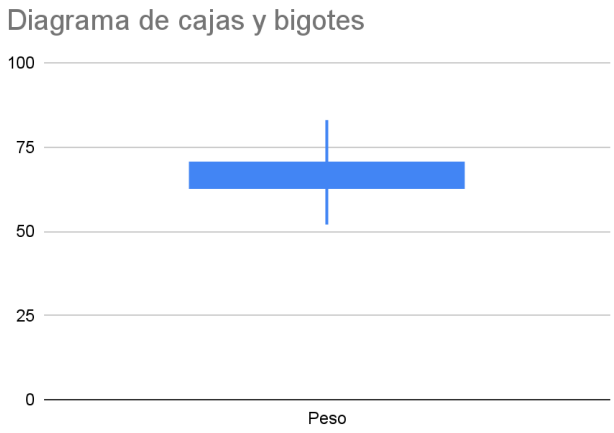
Se puede ver que la forma de ejecución entre ambas herramientas es diferente y la codificación es algo más compleja en R, pero el resultado es exactamente el mismo.

e) Represente un diagrama de barras y uno de cajas y bigotes. Realice el análisis correspondiente (Excel, Jasp y RStudio)

Excel

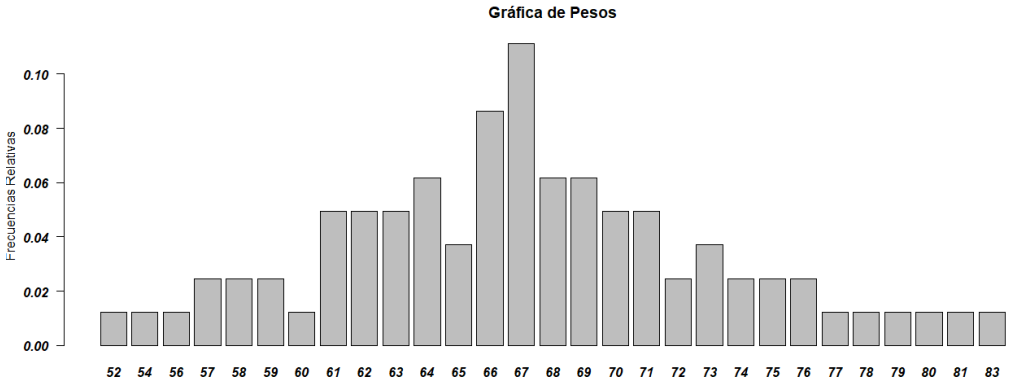


Se ha presentado el diagrama de barras en horizontal para comprender mejor la herramienta y probar diferentes formas y composiciones con la orientación, los títulos, etc.

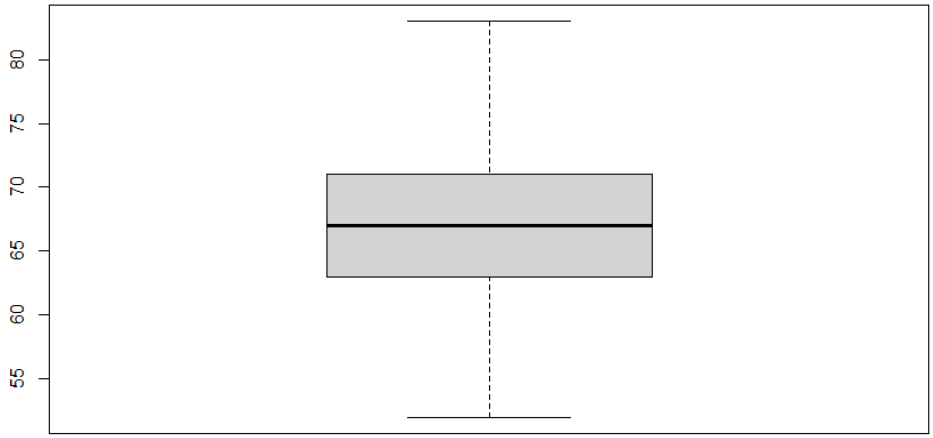


RStudio

```
barplot(prop.table(table(Ejercicio3$V1)) ,main="Gráfica de Pesos",
        ylab ="Frecuencias Relativas",las=1,font.axis=4)
```

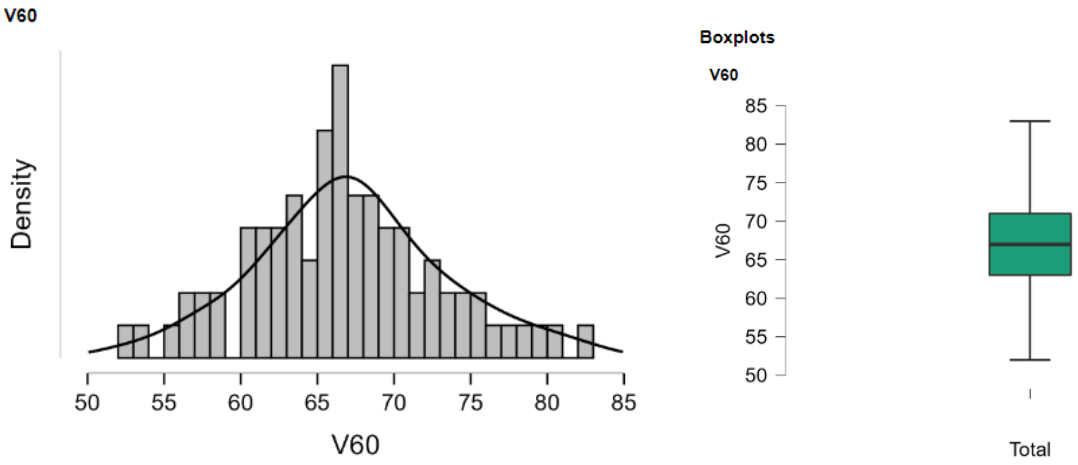



```
> boxplot(Ejercicio3$V1)
```



Jasp

Para formar los gráficos en Jasp es tan sencillo como seleccionar el tipo de gráfico deseado:



Al igual que en Excel y Rstudio, vemos las mismas representaciones del diagrama de barras y el diagrama de cajas y bigotes.

Un diagrama de barras es un gráfico utilizado para representar datos de variables cualitativas o discretas. Está formado por barras rectangulares cuya altura es proporcional a la frecuencia de la variable. En nuestro ejemplo podemos ver que los valores centrales de peso son más comunes, mientras que es menos común encontrar personas con muy bajo peso o peso más alto.

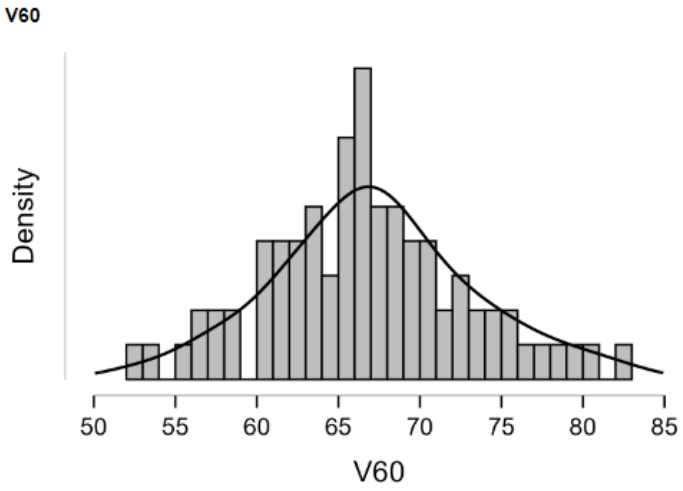
Un diagrama de cajas y bigotes describe varias características importantes, como son la dispersión y la simetría. Representa tres cuartiles y los valores mínimo y máximo de los datos sobre un rectángulo.

Vemos en el diagrama de cajas y bigotes la siguiente información:

- La línea más gruesa en el centro representa la **mediana**, que se encuentra en torno a 67 y la cual separa la mitad superior de los datos de la mitad inferior.
- El p**rimer y el tercer cuartil** delimitan los límites inferior(Q1) y superior (Q3). Los cuartiles dividen la muestra en 4 partes iguales. El primer cuartil representa el valor hasta el primer 25% de la muestra, el segundo cuartil el valor hasta el 50% de la muestra y el tercer cuartil el valor hasta el 75% de la muestra. Teniendo esto en cuenta, en nuestro diagrama vemos que los valores del primer y tercer cuartil son aproximadamente 62 y 71 aproximadamente. Esto quiere decir que el 50% central de la variable oscila entre las puntuaciones 62 y 71.
- La **línea vertical** nos muestra entre qué puntuación se mueve la variable. En sus extremos identifica el **valor máximo y el valor mínimo** de la variable, lo que nos permite identificar casos con valores atípicos (valores con gran distancia a la media. En nuestro caso el rango de valores va de 52 a 83, siendo estos el mínimo y el máximo de la muestra.
- Fuera de esta línea vertical, en nuestro diagrama, no encontramos **valores extremos**, los cuales se dibujarían como casos puntuales fuera del intervalo.

f) Los datos obedecen a una curva de normalidad?. Obténgala y realice el análisis correspondiente. (Jasp)

Jasp



La normalidad significa que el conjunto de datos puede estar bien modelado por una distribución normal. Se puede concluir que los datos obedecen a una curva de normalidad por los siguientes motivos:

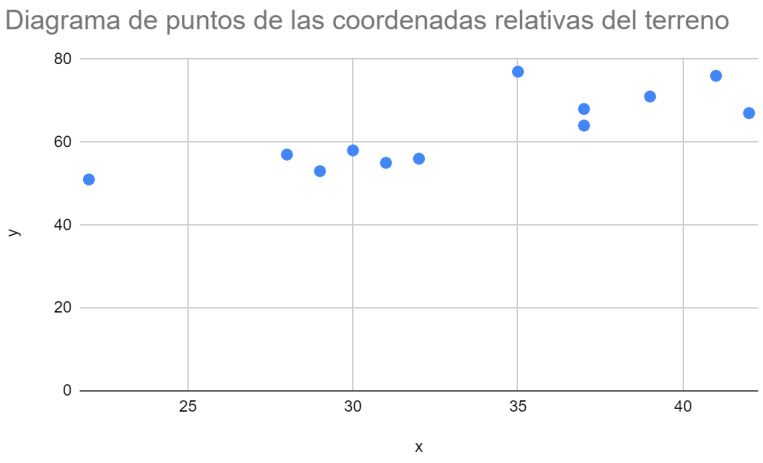
- Visualmente, la curva normal tiene forma de campana y es simétrica.
- Su distribución muestra una elevación en el centro con colas que bajan por ambos lados.
- Las tres medidas de centralización (media, mediana y moda) coinciden en el punto superior de la curva (media = 67.0, mediana = 67.175, moda = 67)
- Cumple la regla empírica:
    - Aproximadamente el 68% de sus valores se encuentran a no más de una desviación estándar respecto de la media.
    - Aproximadamente el 95% de sus valores se encuentran a no más de dos desviaciones estándares respecto de la media.
    - Casi todos sus valores se encuentran a no más de tres desviaciones estándares respecto de la media.

4. Pregunta 4

Sea una parcela o porción de terreno, en la cual se han tomado las coordenadas Relativas:

x	y
37	64
39	71
29	53
42	67
31	55
30	58
35	77
28	57
32	56
22	51
41	76
37	68

a) Representación de la nube de puntos



b) La covarianza. Análisis de resultados

La covarianza es un valor que indica el grado de variación conjunta de dos variables aleatorias respecto a sus medias. Es el dato básico para determinar si existe una dependencia entre ambas variables y además es el dato necesario para estimar otros parámetros básicos, como el coeficiente de correlación lineal o la recta de regresión.

```
Covariancia = COVAR(A2:A13,B2:B13)
```

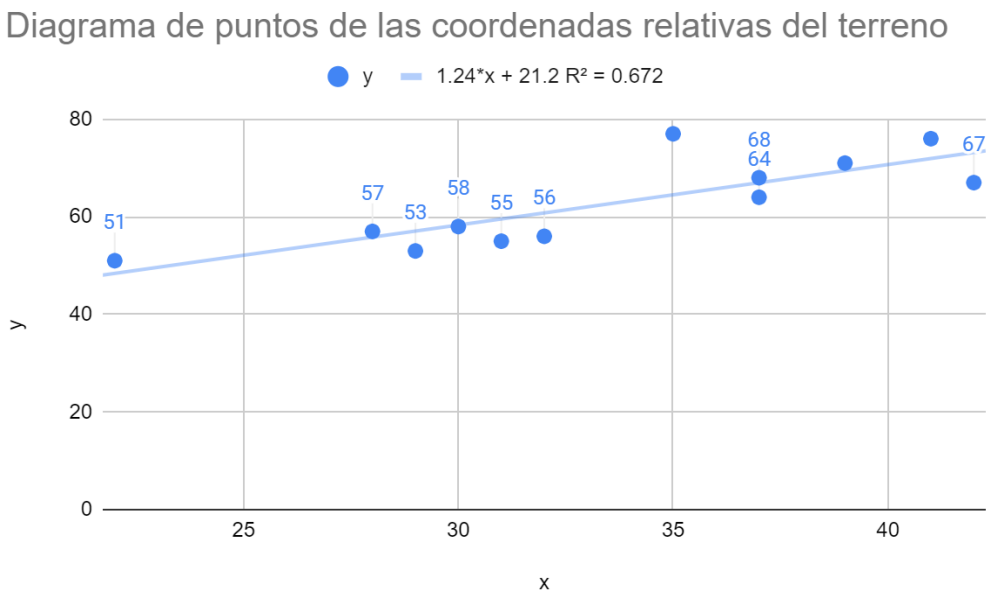
40.14583333

c) La recta de regresión y el coeficiente de determinación.

En el siguiente gráfico podemos ver representada la recta de regresión y el coeficiente de determinación, siendo estos:

Recta de regresión: $1,24x + 21,2$

Coeficiente de determinación: $R = \sqrt{0,672} = 0,81975$



d) Realice análisis de resultados.

• Escoja la aplicación informática que considere. (Excel o RStudio)

Se ha escogido excel para realizar el análisis de resultados:

	Pendiente	Intersección
Coeficiente	0.5423585702	-0.4496669481
Standard error	0.1198720706	7.592344683
Coeficiente de determinación	0.6718180618	3.572610047
Grados de libertad	20.47090298	10
Regresión SS, Residual SS	261.2812412	127.6354255

El coeficiente de determinación, también llamado R-cuadrado es una estadística que indica qué tan cerca están de la línea de regresión ajustada. Es la proporción de la varianza en la variable de respuesta que se puede explicar por la variable explicativa.

El R-cuadrado siempre está entre 0 y 100%:

- 0% indica que el modelo no explica ninguna porción de la variabilidad de los datos de respuesta en torno a su media.
- 100% indica que el modelo explica toda la variabilidad de los datos de respuesta en torno a su media.

En este ejemplo, aproximadamente el 67,73% de la variación en las coordenadas “y” de la parcela se puede explicar por las coordenadas “x” de la parcela. Por lo tanto, tenemos un 67% y podemos determinar que tenemos un porcentaje de R-cuadrado bastante alto,ajustándose bastante bien el modelo a los datos.

5. Pregunta 5

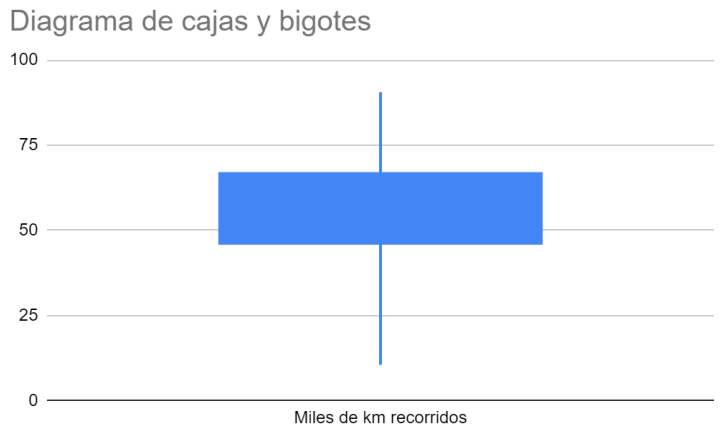
Un fabricante de neumáticos ha recabado una muestra de datos en diferentes concesionarios, sobre la cantidad de miles de kilómetros recorridos por un modelo concreto de esos neumáticos hasta que se ha producido un pinchazo o un reventón del neumático. Los concesionarios la han proporcionado los siguientes datos:

miles de kms de un modelo de neumáticos							
61.979	43.068	41.539	62.215	51.269	82.919	34.182	37.654
51.179	74.582	58.708	48.035	67.124	41.830	61.030	58.267
74.239	60.727	56.155	86.070	90.565	53.751	76.580	68.629
48.240	57.884	55.257	84.656	48.662	10.504	60.951	38.420
79.426	67.662	53.324	49.011	29.480	41.128	30.252	33.412
47.012	71.360	78.635	41.715	72.635	41.463	48.996	48.172
55.643	55.912	46.681	66.519	59.168	66.313	35.884	28.625
84.588	40.709	50.238	61.390	85.720	45.313	46.724	61.752
63.692	70.003	65.996	55.989	49.677	46.502	67.467	64.398
44.411	41.886	34.754	59.888	59.449	67.632	89.116	69.483
48.698	65.854	75.850	36.949	75.548	69.010	61.477	65.585
52.452	50.432	37.748	51.831	73.808	61.065	35.807	57.277
80.502	35.342	44.719	37.402				

a) Represente un diagrama de cajas y bigotes empleando Excel, jasp y rstudio calculando los estadísticos correspondientes para el análisis adecuado de los resultados de esta muestra.

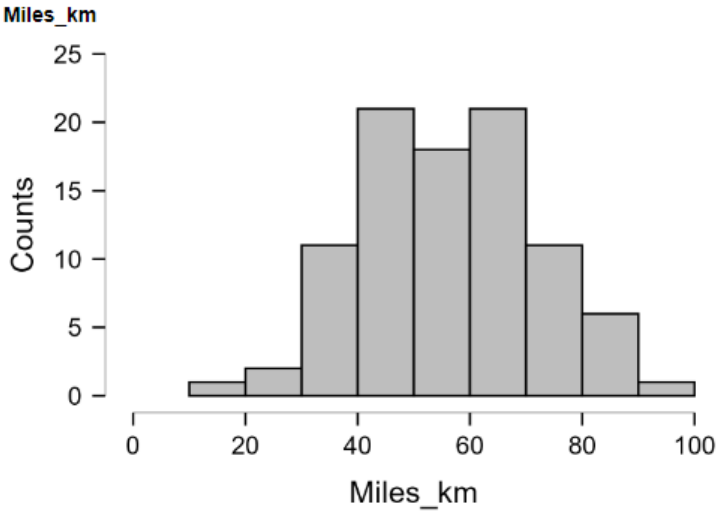
Excel

Mínimo	Cuartil 1	Cuartil 3	Máximo
10.504	46.20475	66.67025	90.565



Jasp

Distribution Plots



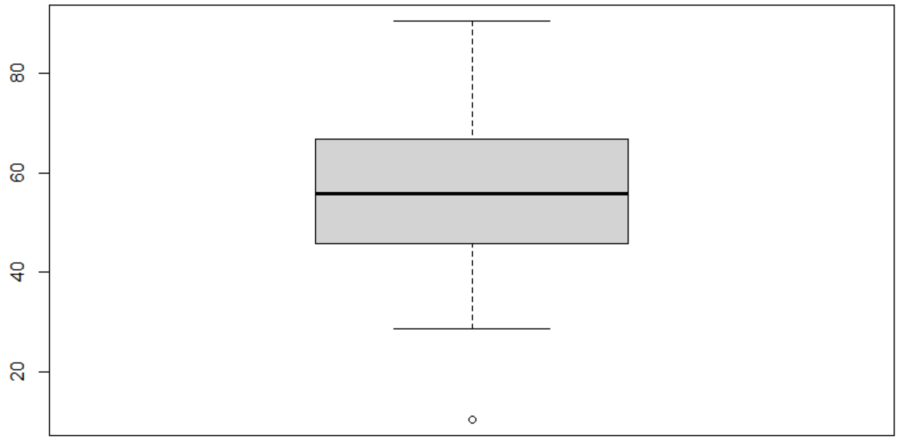
Descriptive Statistics

	Miles_km
Valid	92
Missing	0
Mode [a]	10.504
Median	55.950
Mean	56.226
Std. Deviation	15.643
Minimum	10.504
Maximum	90.565
25th percentile	46.205
50th percentile	55.950
75th percentile	66.670

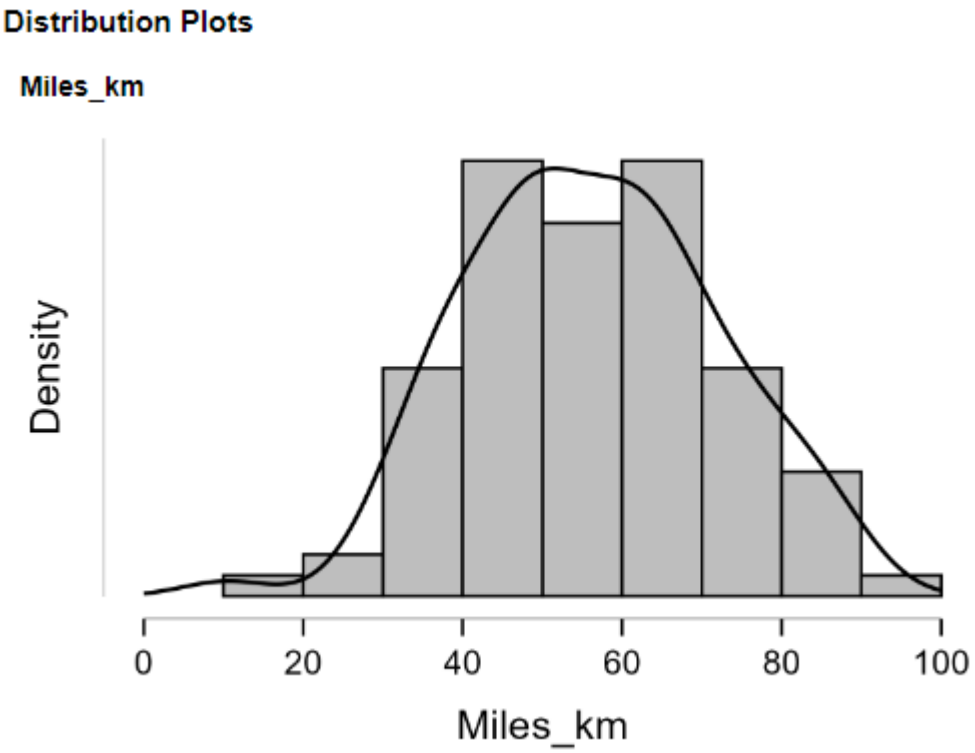
[a] More than one mode exists, only the first is reported

RStudio

```
> boxplot(Ejercicio5$Miles_km)
```



b) Realice una prueba de normalidad numérica y gráficamente para comentar si se pueden extrapolar estos resultados a la población de ese tipo de neumáticos. Comente resultados.



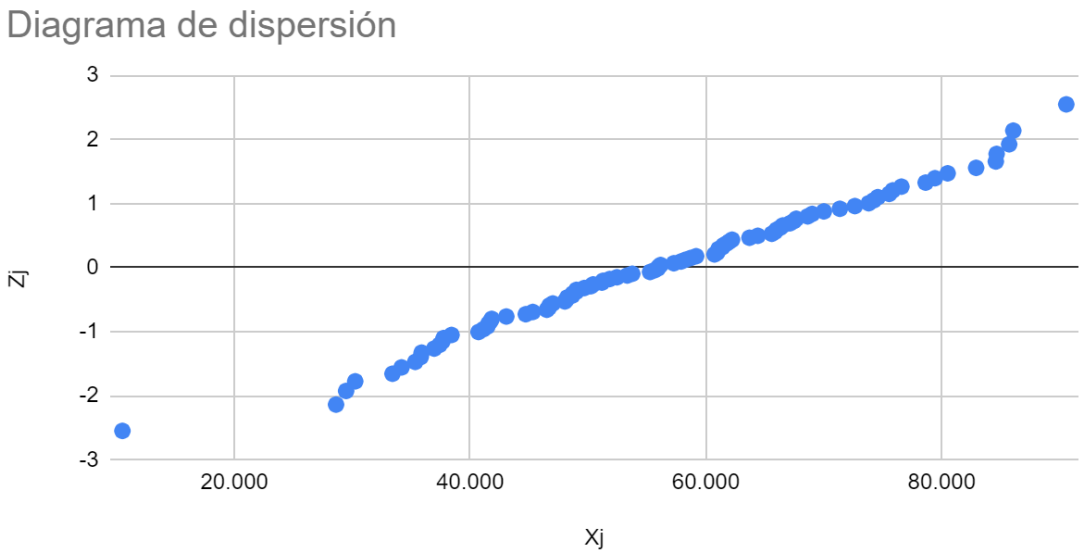
Para hacer la prueba de normalidad, para analizarlo de forma numérica se ha decidido dibujar un diagrama de dispersión en excel, además de este histograma (el cual visualmente, nos puede indicar una distribución normal). Para ello, por cada valor se ha calculado la probabilidad asociada a la función de distribución normal y su correspondiente función de distribución normal.

	Total		
n	92	Probabilidad asociada a la función de distribución normal	Función de distribución normal (NORMSINV)
j	Xj	(j-0,5)/n	Zj
1	10.504	0.005434782609	-2.54686443
2	28.625	0.01630434783	-2.136868412
3	29.480	0.02717391304	-1.924053913
4	30.252	0.03804347826	-1.773856098
5	33.412	0.04891304348	-1.655485342
6	34.182	0.0597826087	-1.556601121
7	35.342	0.07065217391	-1.470951079
8	35.807	0.08152173913	-1.394908353
9	35.884	0.09239130435	-1.32617237

10	36.949	0.1032608696	-1.263187679
11	37.402	0.1141304348	-1.204850891
12	37.654	0.125	-1.150349381
13	37.748	0.1358695652	-1.099066341
14	38.420	0.1467391304	-1.050521829
15	40.709	0.1576086957	-1.004334539
16	41.128	0.1684782609	-0.9601961155
17	41.463	0.1793478261	-0.9178533405
18	41.539	0.1902173913	-0.8770954738
19	41.715	0.2010869565	-0.8377450369
20	41.830	0.2119565217	-0.799650976
21	43.068	0.222826087	-0.762683498
22	44.719	0.2336956522	-0.7267301097
23	45.313	0.2445652174	-0.6916925367
24	46.502	0.2554347826	-0.6574842964
25	46.681	0.2663043478	-0.6240287634
26	46.724	0.277173913	-0.5912576127
27	47.012	0.2880434783	-0.5591095521
28	48.035	0.2989130435	-0.5275292833
29	48.172	0.3097826087	-0.4964666418
30	48.240	0.3206521739	-0.465875879
31	48.662	0.3315217391	-0.4357150594
32	48.698	0.3423913043	-0.4059455499
33	48.996	0.3532608696	-0.3765315843
34	49.011	0.3641304348	-0.3474398899
35	49.677	0.375	-0.3186393636
36	50.238	0.3858695652	-0.290100791
37	50.432	0.3967391304	-0.2617965992
38	51.179	0.4076086957	-0.2337006392
39	51.269	0.4184782609	-0.2057879915
40	51.831	0.4293478261	-0.1780347919
41	52.452	0.4402173913	-0.1504180738
42	53.324	0.4510869565	-0.1229156243
43	53.751	0.4619565217	-0.09550585055
44	55.257	0.472826087	-0.06816765595
45	55.643	0.4836956522	-0.04088032299
46	55.912	0.4945652174	-0.01362340117
47	55.989	0.5054347826	0.01362340117

48	56.155	0.5163043478	0.04088032299
49	57.277	0.527173913	0.06816765595
50	57.884	0.5380434783	0.09550585055
51	58.267	0.5489130435	0.1229156243
52	58.708	0.5597826087	0.1504180738
53	59.168	0.5706521739	0.1780347919
54	60.727	0.5815217391	0.2057879915
55	60.951	0.5923913043	0.2337006392
56	61.030	0.6032608696	0.2617965992
57	61.065	0.6141304348	0.290100791
58	61.390	0.625	0.3186393636
59	61.477	0.6358695652	0.3474398899
60	61.752	0.6467391304	0.3765315843
61	61.979	0.6576086957	0.4059455499
62	62.215	0.6684782609	0.4357150594
63	63.692	0.6793478261	0.465875879
64	64.398	0.6902173913	0.4964666418
65	65.585	0.7010869565	0.5275292833
66	65.854	0.7119565217	0.5591095521
67	65.996	0.722826087	0.5912576127
68	66.313	0.7336956522	0.6240287634
69	66.519	0.7445652174	0.6574842964
70	67.124	0.7554347826	0.6916925367
71	67.467	0.7663043478	0.7267301097
72	67.662	0.777173913	0.762683498
73	68.629	0.7880434783	0.799650976
74	69.010	0.7989130435	0.8377450369
75	70.003	0.8097826087	0.8770954738
76	71.360	0.8206521739	0.9178533405
77	72.635	0.8315217391	0.9601961155
78	73.808	0.8423913043	1.004334539
79	74.239	0.8532608696	1.050521829
80	74.582	0.8641304348	1.099066341
81	75.548	0.875	1.150349381
82	75.850	0.8858695652	1.204850891
83	76.580	0.8967391304	1.263187679
84	78.635	0.9076086957	1.32617237
85	79.426	0.9184782609	1.394908353

86	80.502	0.9293478261	1.470951079
87	82.919	0.9402173913	1.556601121
88	84.588	0.9510869565	1.655485342
89	84.656	0.9619565217	1.773856098
90	85.720	0.972826087	1.924053913
91	86.070	0.9836956522	2.136868412
92	90.565	0.9945652174	2.54686443



Como vemos en la gráfica, los valores siguen una tendencia de línea recta. Esto indica que los datos están siguiendo una distribución normal. Encontramos dos datos en los extremos, pero que realmente no cambian la tendencia de la recta. Que los datos sigan una distribución normal nos permite poder analizar los datos mediante pruebas estadísticas convencionales.

Volviendo al diagrama de cajas y bigotes del apartado anterior, vemos también que los cuartiles no están muy separados, lo que indica que los datos siguen una tendencia y no están muy dispersos. Esta es otra evidencia para determinar la normalidad.

Teniendo en cuenta todo lo expuesto anteriormente, se podría considerar que se puede extrapolar los resultados a la población de ese tipo de neumáticos.