# Prácticas de Estadística
# Ejercicio de la Práctica 1
# Estadística Descriptiva

true

25/11/2020

## Contents



## Presentación del ejercicio

La hoja de datos *airquality* de R contiene una serie de medidas sobre variables que desciben la calidad del aire en Nueva York entre mayo y septiembre de 1973. Concretamente, hay mediciones del nivel de ozono (en ppb), de la radiación solar (en lang), de la velocidad promedio del viento (en mph) y de la temperatura máxima diaria (en grados Farenhait). Podemos ver todos los detalles sobre estos datos si lanzamos a la consola de R *?airquality*. Además, tenemos el día y el mes en que se tomó cada medición.

El objetivo del ejercicio es realizar un análisis descriptivo de la hoja.

En esta ocasión no es necesario importar la hoja porque ya pertenece al entorno de trabajo que, por defecto, se incorpora al lanzar R.

Podemos ver un resumen inicial de todas las variables de esa hoja de datos a continuación:

```
summary(airquality)
```

```
##      Ozone           Solar.R           Wind             Temp      
##  Min.   :  1.00   Min.   :  7.0   Min.   : 1.700   Min.   :56.00  
##  1st Qu.: 18.00   1st Qu.:115.8   1st Qu.: 7.400   1st Qu.:72.00  
##  Median : 31.50   Median :205.0   Median : 9.700   Median :79.00  
##  Mean   : 42.13   Mean   :185.9   Mean   : 9.958   Mean   :77.88  
```

```
##  3rd Qu.: 63.25   3rd Qu.:258.8   3rd Qu.:11.500   3rd Qu.:85.00
##  Max.   :168.00   Max.   :334.0   Max.   :20.700   Max.   :97.00
##  NA's   :37       NA's   :7
##      Month            Day
##  Min.   :5.000   Min.   : 1.0
##  1st Qu.:6.000   1st Qu.: 8.0
##  Median :7.000   Median :16.0
##  Mean   :6.993   Mean   :15.8
##  3rd Qu.:8.000   3rd Qu.:23.0
##  Max.   :9.000   Max.   :31.0
##
```

# Un análisis estadístico descriptivo sobre indicadores climáticos del aire en el Nueva York de 1973

## Resumen

El informe recoge las principales características, desde el punto de vista descriptivo, de 4 indicadores relacionados con el clima y la calidad del aire en Nueva York, medidos entre mayo y septiembre de 1973: cantidad de ozono en el aire, nivel de radiación solar, velocidad del viento y temperatura máxima.

## Introducción

El clima es un fenómeno complejo sujeto a un alto nivel de incertidumbre y observable a través de múltiples indicadores; algunas de estas variables climáticas están relacionadas con la calidad del aire en las ciudades, como el nivel de ozono.

El punto de partida de este análisis lo constituye una hoja de datos que recopila 4 de esos indicadores en la ciudad de Nueva York a lo largo de los meses de mayo, junio, julio, agosto y septiembre del año 1973. El objetivo general del trabajo es el de proporcionar una visión global sobre el clima en Nueva York y sobre la calidad de su aire en lo que respecta al nivel de ozono. Como objetivos específicos, proponemos:

1. Realizar un análisis descriptivo básico que incluya la descripción de la distribución de frecuencias, medidas de posición, dispersión y forma, de los cuatro indicadores que ofrece la hoja: nivel de ozono, radiación solar, velocidad del viento y temperatura máxima diaria.
2. Detectar qué días del año resultaron atípicos en relación con cada uno de los cuatro indicadores.

## Métodología

Como hemos comentado, la hoja recoge mediciones recogidas entre mayo y septiembre de 1973, en la ciudad de Nueva York, de las siguientes variables:

1. Ozone: nivel medio de ozono en partes por billón desde las 13.00 hasta las 15.00 horas en la Isla de Roosevelt.
2. Solar.R: radiación solar en Langleys, en la frecuencia de banda de 4000???7700 Angstroms desde las 08.00 hasta las 12.00 horas en Central Park.
3. Wind: velocidad promedio del viento en millas a la hora desde las 7.00 hasta las 10.00 horas en el Aeropuerto de La Guardia.
4. Temp: temperatura máxima diaria en grados Fahrenheit en el aeropuerto de La Guardia.

Además, dos variables indican el día y el mes de cada medición tomada.

En primer lugar, constataremos que se han recogido medidas todos los días del período de observación. A continuación, para cada una de las cuatro variables de interés, obtendremos:

1. Una representación de su distribución de frecuencias. Dado que se trata de variables continuas, la herramienta será el histograma.
2. Medida de posición: media y cuartiles.
3. Medida de dispersión: coeficiente de variación.
4. Medida de forma: coeficiente de asimetría de Fisher.
5. Diagrama de caja donde puedan identificarse los días atípicos en cuanto al valor de cada variable.

## Resultados

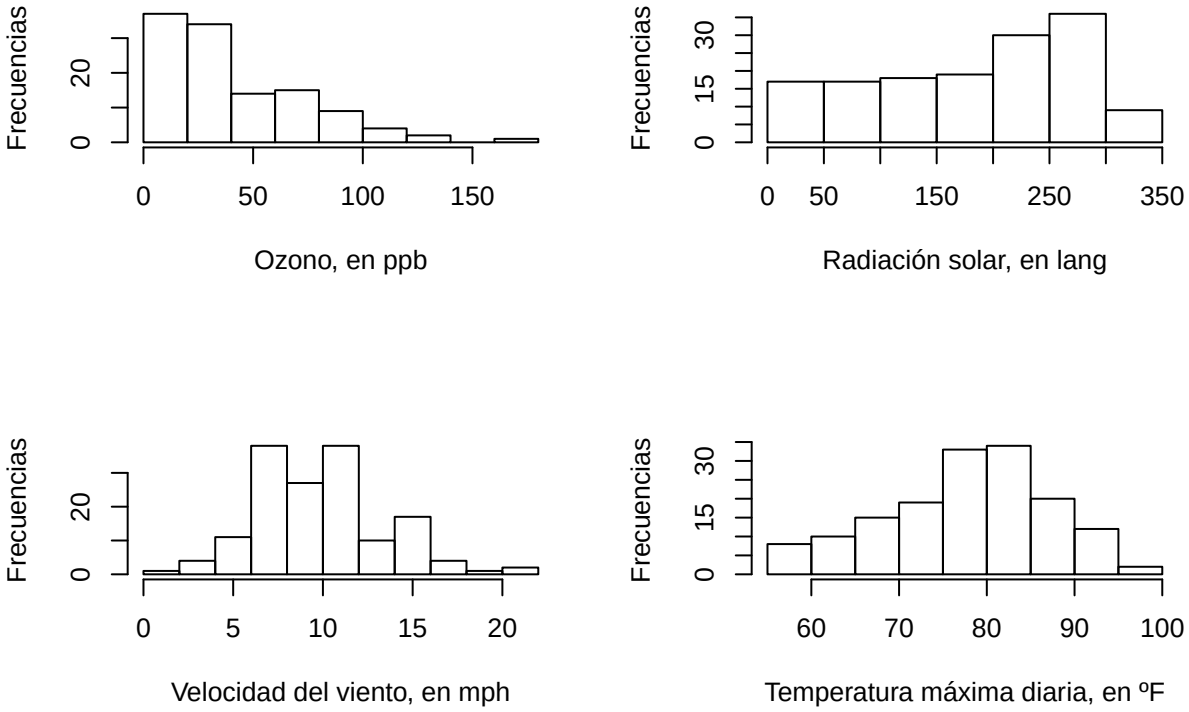
En primer lugar, a modo de comprobación, obtenemos una tabla de frecuencias absolutas de la variable *mes* para comprobar que se han tomado tantas medidas de cada variable como días tienen los meses entre mayo y septiembre:

```
##
##  5  6  7  8  9
## 31 30 31 31 30
```

Todo parece indicar que sí se tomaron medidas cada día de cada mes

Las 4 figuras que se muestran a continuación correponden a los histogramas de las variables objeto del análisis.

```
par(mfrow = c(2, 2))
hist(airquality$Ozone, ylab = "Frecuencias", xlab = "Ozono, en ppb", main = "")
hist(airquality$Solar.R, ylab = "Frecuencias", xlab = "Radiación solar, en lang", main = "")
hist(airquality$Wind, ylab = "Frecuencias", xlab = "Velocidad del viento, en mph", main = "")
hist(airquality$Temp, ylab = "Frecuencias", xlab = "Temperatura máxima diaria, en ºF", main = "")
```




Por su parte, la tabla siguiente contiene las medidas descriptivas mencionadas.

```
##                Ozono Radiación.solar Velocidad.del.viento Temperatura
## Media     42.1293103     185.9315068            9.9575163  77.8823529
## P25       18.0000000     115.7500000            7.4000000  72.0000000
## Me        31.5000000     205.0000000            9.7000000  79.0000000
## P75       63.2500000     258.7500000           11.5000000  85.0000000
## CV         0.7830151       0.4843634            0.3538032   0.1215329
## Coef.asim  1.2098656      -0.4192893            0.3410275  -0.3705073
```

## Análisis

Del análisis de las distribuciones de frecuencias podemos destacar los siguientes aspectos:

- En el ozono la mayoría de los días hay niveles bajos, pero por alguna razón hay unos pocos días en que se dispara.
- En la radiación solar parece pasar justo lo contrario: lo más frecuente son días con valores de radiación entre 200 y 300 lang, pero hay algunos días que destacan por su baja radiación. Quizá simplemente son días nublados.
- En la velocidad del tiempo y la temperatura máxima diaria, sin embargo, parece que hay unos valores centrales más frecuentes y luego días de observaciones menores o mayores con las mismas frecuencias.

Estas valoraciones tienen que ver con la forma de las distribuciones de frecuencias, que analizaremos de forma cuantitativa, mediante el coeficiente de asimetría, a continuación.

### Ozono

En este enlace se considera que niveles de ozono por encima de 100 ppb son peligrosos para grupos de riesgo. El nivel medio observado en nuestros datos está por debajo de los 100, así como el percentil 75, lo que indica que no más del 25 por ciento de los días se padecieron niveles problemáticos de ozono. De hecho, el número de días por encima de 100 ppb fue de 7.

El coeficiente de variación refleja una dispersión moderada, es decir, indica cierta variabilidad de los niveles de ozono a lo largo del período de observación. Por su parte, el coeficiente de asimetría confirma lo que observábamos en el histograma: una fuerte asimetría a la derecha como consecuencia de la existencia de elevados niveles de ozono en algunos días en particular.

### Radiación solar

En Wikipedia hemos encontrado que:

*La insolación anual en la parte alta de la atmósfera a diferentes latitudes es: Para el polo la insolación anual es 133,2 kilolangleys/año. En el ecuador asciende a 320,9 kilolangleys/año, donde el kilolangley=1000 langleys.*

Nuestros datos se refieren a un período de 4 horas. Tomemos como referencia el ecuador, que recibe una radiación promedio en 4 horas de 219.7945205 Langleys. Por tanto, aunque la media de nuestros datos está por debajo, no ocurre así con la mediana: ésta se aproxima bastante al promedio por lo que podemos confirmar que al menos la mitad de los días observados tuvieron una radiación por encima del promedio del ecuador. La dispersión, según muestra el coeficiente de variación es moderada-baja, indicando no muchas diferencias en la radiación a lo largo de todo el período de observación. Finalmente, el coeficiente de asimetría es sólo ligeramente negativo, indicando simetría de la distribución de frecuencias: las diferencias observadas con respecto a la radiación media se dan casi en el mismo sentido a la izquierda y a la derecha de ésta.

### Velocidad del viento

Vamos a valorar las medidas de posición en la escala de Beaufort:

- La velocidad media del viento se considera brisa suave.
- El percentil 25 se considera brisa ligera. Por tanto el 25% de los días tuvieron vientos inferiores o iguales a brisa ligera.
- La mediana es considerada brisa suave, así que el 50% de los días hubo, como máximo, una brisa moderada.
- El percentil 75 es considerado igualmente brisa suave. Por tanto, no más del 75% de los días hubo vientos por encima de una brisa suave.

La cercanía entre los percentiles 75 y 25 ya indica no excesiva variabilidad en la velocidad del viento, conclusión que se ve ratificada por el coeficiente de variación, del 35.38%. Finalmente, el coeficiente de asimetría indica que la distribución de frecuencias es simétrica a izquierda y derecha de la media.

### Temperatura máxima diaria

Para comparar las medidas descriptivas con algún valor de referencia, vamos a considerar el promedio de los meses de mayo a septiembre en Nueva York que aparecen en este enlace, que es de

```
(71+79+84+83+75)/5
```

```
## [1] 78.4
```

y se refieren al período 1981 a 2010:

- El valor medio de nuestros datos (corresponden a 1973) está ligeramente por debajo, pero no así la mediana.
- Como cabe esperar, el percentil 25 está por debajo del valor medio entre 1981 y 2010, y el percentil 75 por encima.

Por otra parte, el coeficiente de variación es el menor de todos los observados en las variables analizadas, e indica que el período de mayo a septiembre de 1973 fue moderadamente estable en cuanto a las temperaturas máximas diarias. Finalmente, el coeficiente de asimetría también es cercano a cero, indicando simetría de las frecuencias de valores a izquierda y derecha de la media.

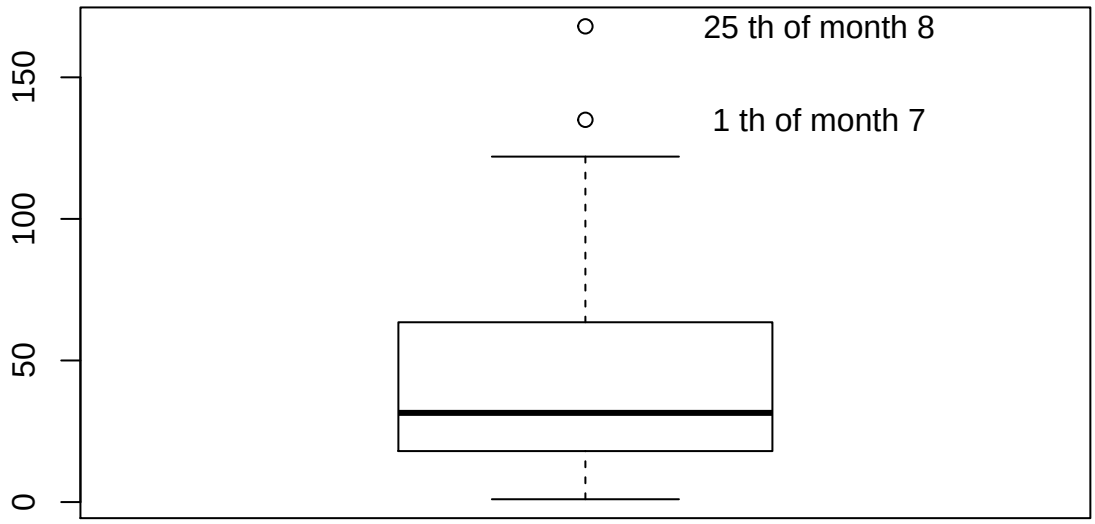
### Análisis de la presencia de datos atípicos

Ahora vamos a identificar datos atípicos en cada una de las variables observadas. En caso de existir datos atípicos, vamos a identificar el día y el mes en que se dio.

### Ozono

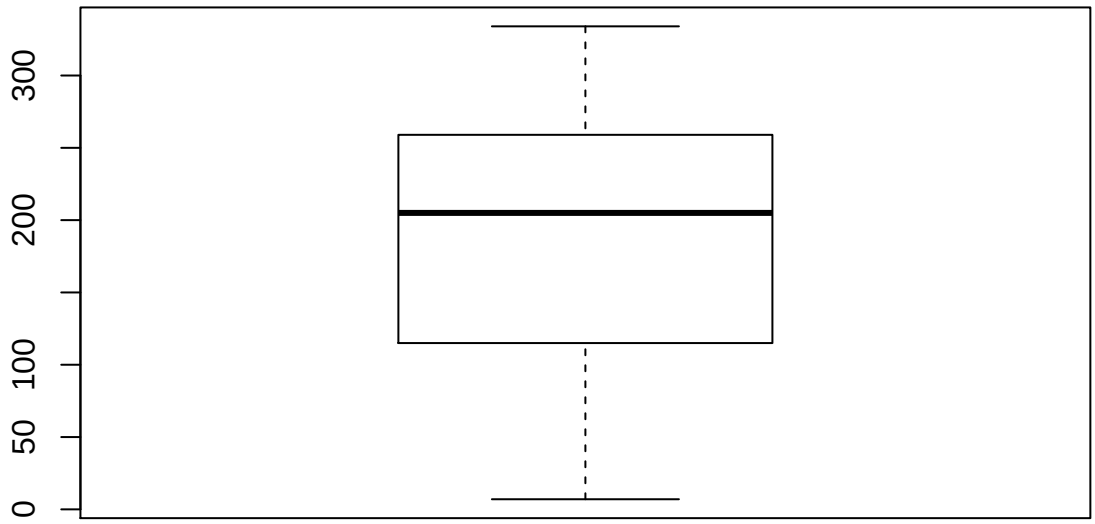
Mostramos a continuación el diagrama de caja con la identificación de los valores atípicos:

```
datos <- airquality[!is.na(airquality$Ozone), ]# Eliminamos los que no tienen valor observado
bp <- boxplot(datos$Ozone)
filtro <- datos$Ozone > bp$stats[5] | datos$Ozone < bp$stats[1]
x <- rep(1.25, sum(filtro))
y <- datos$Ozone[filtro]
id <- paste(datos$Day[filtro], "th of month", datos$Month[filtro])
text(x, y, id)
```



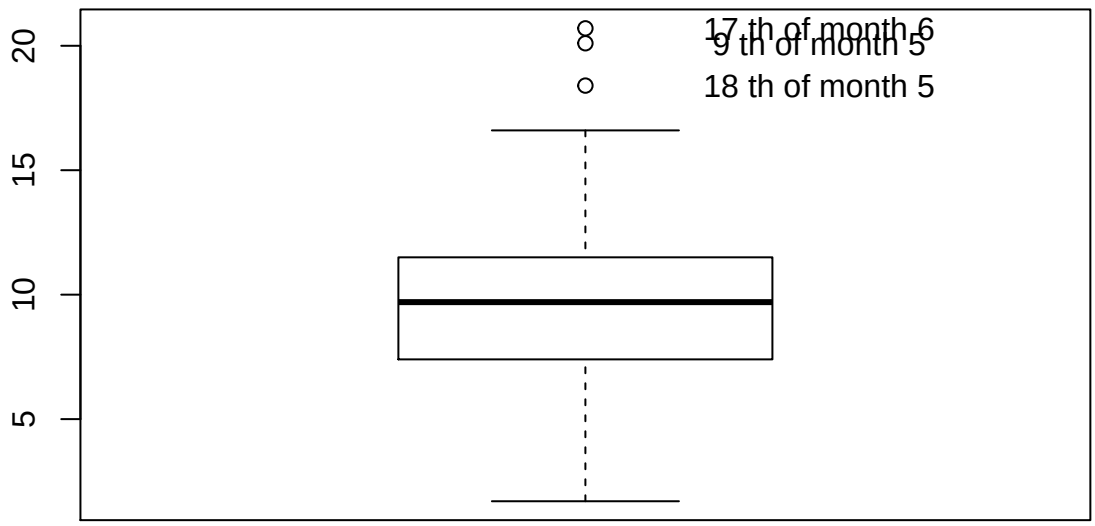

**Radiación solar**

```
datos <- airquality[!is.na(airquality$Solar.R), ]# Eliminamos los que no tienen valor observado
bp <- boxplot(datos$Solar.R)
```



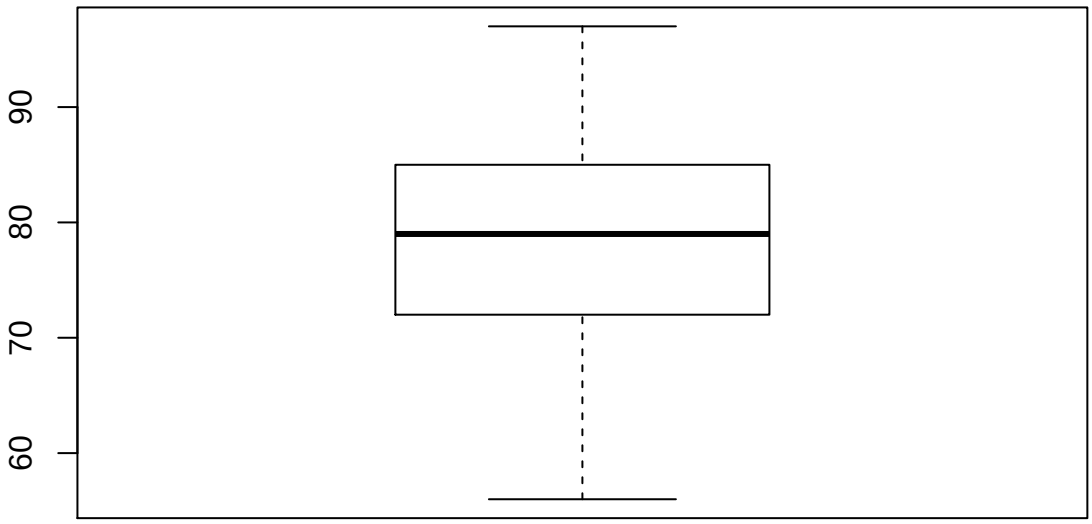

**Velocidad promedio del viento**

```
datos <- airquality[!is.na(airquality$Wind), ]# Eliminamos los que no tienen valor observado
bp <- boxplot(datos$Wind)
filtro <- datos$Wind > bp$stats[5] | datos$Wind < bp$stats[1]
x <- rep(1.25, sum(filtro))
y <- datos$Wind[filtro]
id <- paste(datos$Day[filtro], "th of month", datos$Month[filtro])
text(x, y, id)
```




**Temperatura máxima diaria**

```
datos <- airquality[!is.na(airquality$Temp), ]# Eliminamos los que no tienen valor observado
bp <- boxplot(datos$Temp)
```




**Conclusiones**

(Destaca las conclusiones más relevantes)