

# Practica 2

Alejandro Cáceres  
UPC - Statistics 2019/2020

# Estadística descriptiva

- ▶ describir, organizar, representar y resumir un conjunto de datos
- ▶ tablas de frecuencias, representarlas gráficamente, calcular algunos estadísticos importantes (media aritmética, varianza y moda)
- ▶ interpretar todos estos resultados en R

# Report

- ▶ Para desarrollar código usamos un script (test.R)
- ▶ También podemos desarrollar código y hacer un report simultáneamente con Markdown

Abre un nuevo archivo markdown y guárdalo como 'practica.Rm'. En el boton knit selecciona knit to HTLM. Editalo.

# Datos

- ▶ La vez pasada habíamos visto los datos `airquality` como ejemplo de un `data.frame`
- ▶ carguémoslos desde un archivo de texto (`air.txt`)

## Tablas

Cargemos los datos `air.txt` de un archivo de texto (primero descarga los datos de ATENEA al escritorio)

```
> a <- read.table(file="air.txt")
```

`read.table` es una función de R con parámetros:  
`MiFuncion(Objeto, par1=x, par2=y, par3=y,...)`

Qué pasa si especificamos los parametros:  
`header=TRUE, dec=",", na.strings = "NA"`?

## Datos

Carga los datos con

```
> a <- read.table(file="air.txt",  
header=TRUE, dec=",", na.strings="NA")
```

Preguntas:

- ▶ Qué estructura de datos es a?
- ▶ Cuales son sus dimensiones?
- ▶ Cómo obtengo la variable Ozone?
- ▶ Qué estructura es la variable Ozone? es un vector?
- ▶ Quiero los diez primeros datos de Ozone.
- ▶ Qué hace `length(a$Ozone)`?

# Datos

- ▶ Datos categóricos: Discretos numerables, pueden ser (mes, día) y sin orden (sexo, color)

Exploremos la variable `Month` de `a`

# Tablas

Frecuencia absoluta ( $n_i$ ): Cuantas veces aparecen cada uno de los meses en la variable Month?

- ▶ Asígnale ( `<-` ) la variable 'ni' a la aplicación de la función `table` sobre `a$Month`
- ▶ Cuantas observaciones hubo el junio y agosto?



# Tablas

Frecuencia relativa ( $f_i$ ): Cuál es la proporción de veces que aparece cada mes, con respecto al número total de mediciones?

- ▶ Aplica la función `prop.table` sobre 'ni' (asígnale 'fi')
- ▶ Por qué `sum(fi)` da 1? (Qué es `sum`?)
- ▶ Cuál es la frecuencia relativa hasta julio?

nota: la frecuencia relativa es la frecuencia absoluta dividida por el número total de datos.

## Tablas

Frecuencia absoluta acumulada ( $N_i$ ): Forma un vector (de 5 componentes) que enumere el número de datos hasta el mes de mayo (mes 5), los datos hasta junio (mes 6), ... los datos hasta septiembre (mes 9)

- ▶ Aplica la función `cumsum` sobre 'ni' (asígnale 'Ni')
- ▶ Qué representa `Ni[-c(2,3,4,5)]`?

nota: Frecuencia absoluta acumulada es la suma de los distintos valores de la frecuencia absoluta tomando como referencia una categoría dada.

# Tablas

Frecuencia relativa acumulada ( $F_i$ ): Forma un vector (de 5 componentes) que indique la frecuencia relativa de datos hasta el mes de mayo (mes 5), la frecuencia relativa hasta junio (mes 6), ... la frecuencia relativa hasta septiembre (mes 9)

- ▶ Aplica la función `cumsum` sobre 'fi' (asígnale 'Fi')
- ▶ Qué representa `1-Fi[3]`?

nota: Frecuencia relativa acumulada es el resultado de dividir cada frecuencia absoluta acumulada por el número total de datos.

## Tablas

Pongamos todas las frecuencias en una matriz

```
> Tabla_Frec <- cbind(ni,Ni,fi,Fi)
```

```
# Se visualiza la tabla
```

```
> Tabla_Frec
```

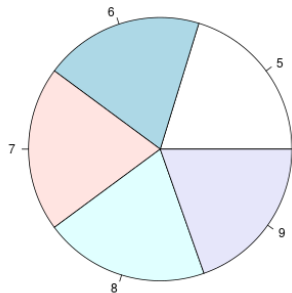
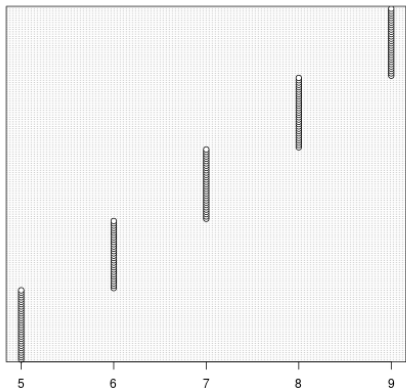
	ni	Ni	fi	Fi
5	31	31	0.2026144	0.2026144
6	30	61	0.1960784	0.3986928
7	31	92	0.2026144	0.6013072
8	31	123	0.2026144	0.8039216
9	30	153	0.1960784	1.0000000

Las frecuencias de los datos categóricos de pueden visualizar con diferentes gráficos

- ▶ Gráfico de sectores: función `pie` sobre `'ni'`  
Ilustra las frecuencias relativas de los datos mediante las porciones de un círculo.
- ▶ Gráfico de puntos: función `dotchart` sobre `'a$Month'`: el eje x representa el valor numérico de los datos

## Gráficos

Puedes obtener los siguientes gráficos?



Cuál es de sectores y Cuál es de puntos?

# Datos

- ▶ Datos numéricos: Pueden ser discretos (años) o continuos (Ozono)

Exploremos la variable `Ozone` de `a`

## Estadísticos descriptivos

Los datos continuos generalmente son sumarizados con estadísticos descriptivos

- ▶ media: el promedio de los datos  $\bar{x} = \frac{1}{n} \sum_{i=1..n} x_i$
- ▶ mínimo: el valor mínimo de los datos
- ▶ máximo: el valor máximo de los datos
- ▶ mediana: el valor que divide los datos en dos partes iguales (el 50% de los datos son menores que la mediana)

Cuales son la media, mínimo, máximo y mediana de los datos de Ozono?

Funciones: mean, min, max, median  
(parámetro na.rm=TRUE)



## Gráficos

Los datos continuos también pueden ser visualizados en gráficos de tallos y hojas (stem):

```
> stem(a$Ozone)
```

```
The decimal point is 1 digit(s) to the right of the |
```

```
0 | 1467778999
1 | 01112233334444666688889
2 | 0000111123333334478889
3 | 001222455667799
4 | 01444556789
5 | 0299
6 | 134456
7 | 13367889
8 | 024559
9 | 1677
10 | 8
11 | 058
12 | 2
13 | 5
14 |
15 |
16 | 8
```

Los números del gráfico corresponden a los datos de ozono ordenados de mínimo a máximo (1,4,6,7,7,7,8,9,9,9,10,11,11,11,12,13 ... 135,168)

## Según el gráfico de tallos y hojas

```
> stem(a$Ozone)
```

The decimal point is 1 digit(s) to the right of the |

```
0 | 1467778999
1 | 01112233334444666688889
2 | 0000111123333334478889
3 | 001222455667799
4 | 01444556789
5 | 0299
6 | 134456
7 | 13367889
8 | 024559
9 | 1677
10 | 8
11 | 058
12 | 2
13 | 5
14 |
15 |
16 | 8
```

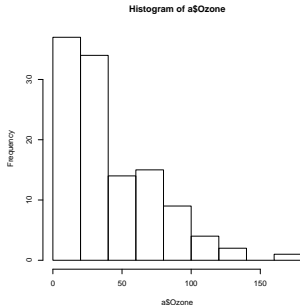
Cuál es la decena que mas observaciones tiene?

## Gráficos

En el gráfico de tallos y hojas los datos están agrupados por decenas, pero se pueden agrupar y visualizar usando subintervalos (bins) flexibles

```
> hist(a$Ozone, br=10)
```

Que hace el parámetro br de hist?



# Estadísticos descriptivos

Los gráficos muestran que la dispersión de los datos es importante. Esta se puede medir con

- ▶ error standard (corregido):

$$s = \left[ \frac{1}{n-1} \sum_{i=1..n} (x_i - \bar{x})^2 \right]^{1/2}$$

distancia cuadrática alrededor de la **media**

Cuál es el error standard de Ozone?

Función: `sd` (parámetro `na.rm=TRUE`)

Nota:  $s^2$  es la varianza muestral

## Estadísticos descriptivos

- ▶ Rango intercuartílico: los valores que están por encima del 25% de los datos y por debajo del 75% (dispersión alrededor de la **mediana**)

```
> quantile(a$Ozone, probs=XX, na.rm=TRUE)
```

probs = 0.25 (primer cuartil 25%)

probs = 0.75 (tercer cuartil 75%),

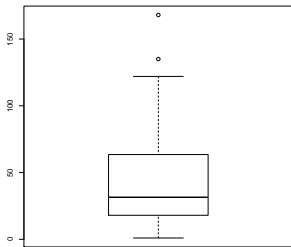
La diferencia entre el primer y tercer cuartil es el rango intercuartílico. Cuál es para ozono?

Coindice con `IQR(a$Ozone, na.rm=TRUE)`?

## Gráficos

**Box plot:** Ilustra los cuartiles (caja), la mediana (línea), el 5% y el 95% de los datos (líneas horizontales) y los valores atípicos (los puntos mas allá de las líneas)

```
> boxplot(a$Ozone)
```



# Gráficos

Para obtener los datos atípicos

```
> bp <- boxplot(a$ozone)
```

```
> class(bp)
```

```
[1] "list"
```

```
> names(bp)
```

```
[1] "stats" "n"      "conf"   "out"    "group"
"names"
```

```
> bp$out
```

Es el valor máximo de ozono atípico?