Descriptiva y regresión.

Grado en Biología sanitaria. Universidad de Alcalá. Curso 2017-18.

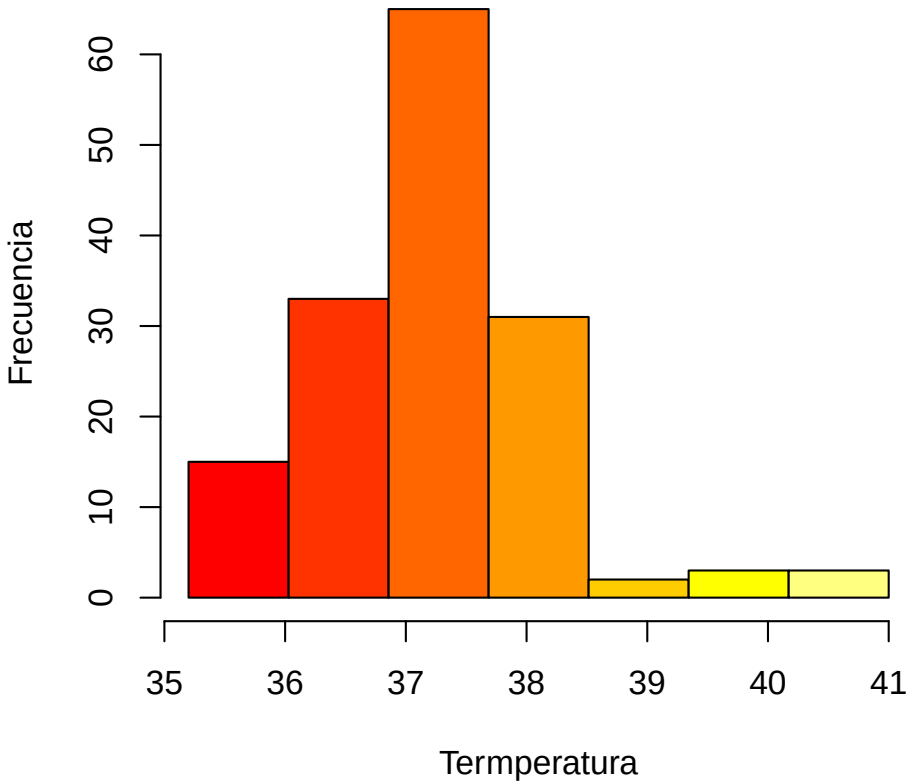
Sesión ~~desastre~~ de sastre

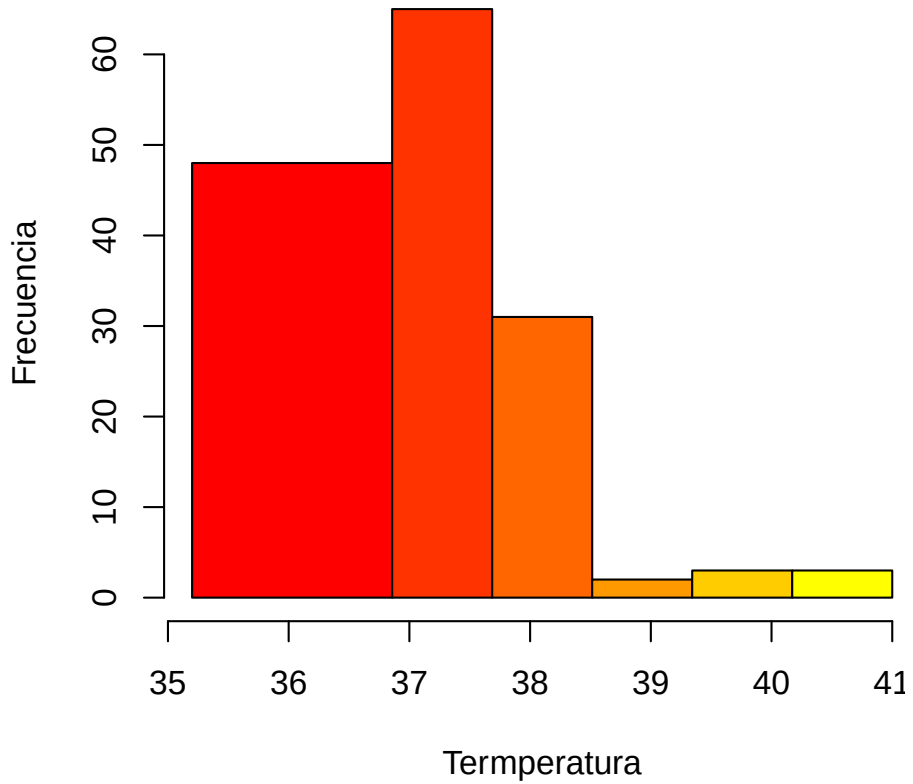
Autor: M. Marvá Ruiz. Actualizado: 2017-09-20

Histogramas con clases de distinta longitud

Conocemos la temperatura de 157 pacientes: 35.2, 35.4, 35.4, 35.5, 35.5, 35.5...

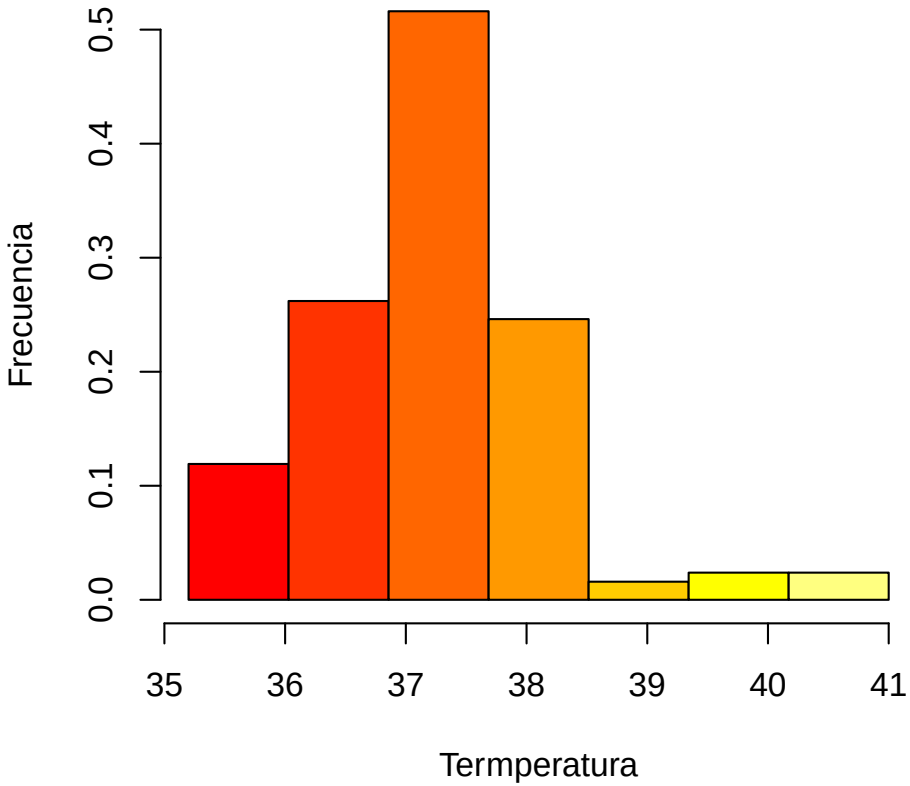
Histogramas con clases de distinta longitud

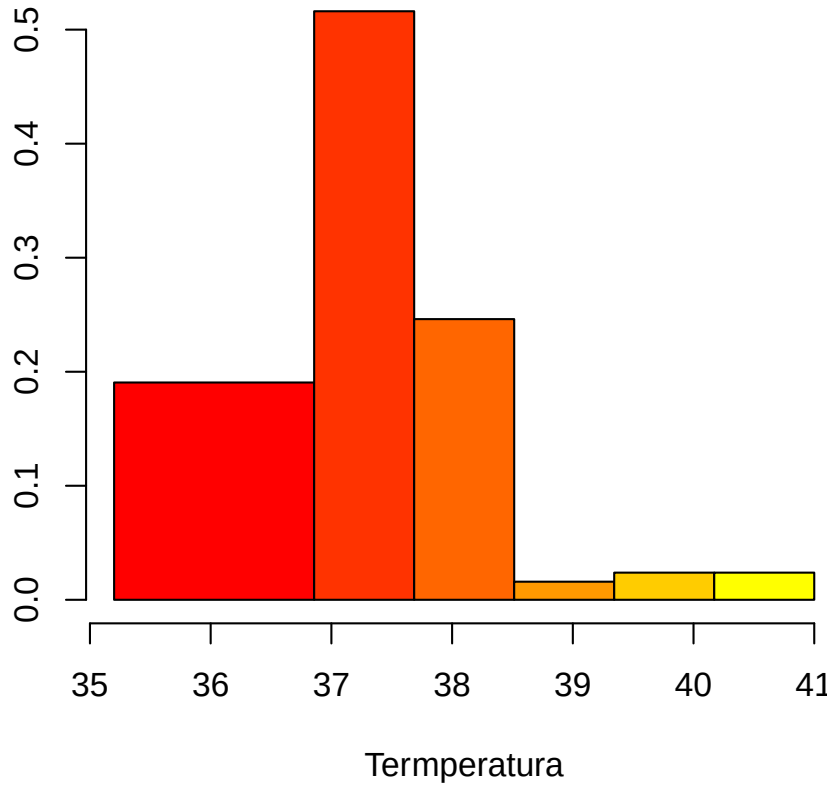




Para que tenga sentido el gráfico de la derecha, ¿qué unidades usamos en el eje vertical?

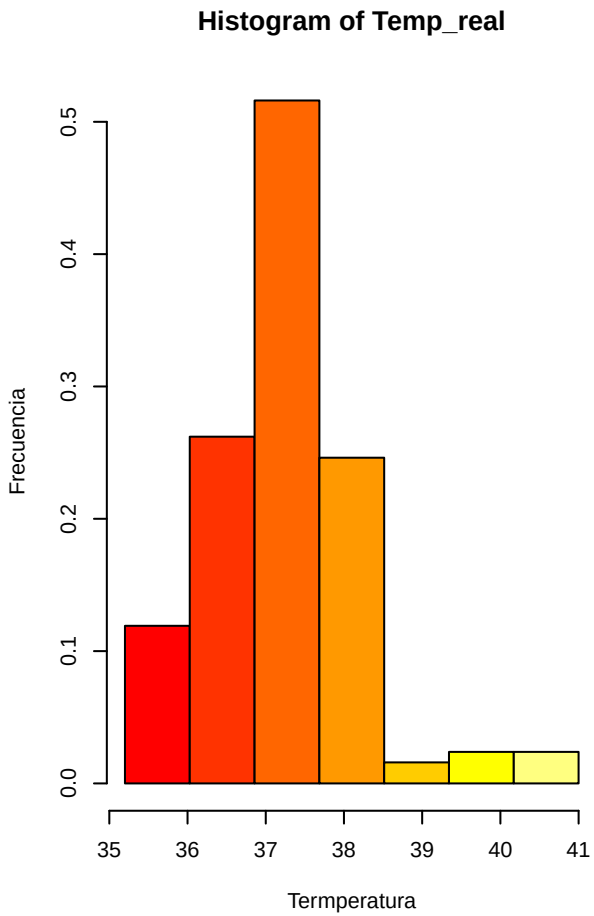
Histogramas con clases de distinta longitud

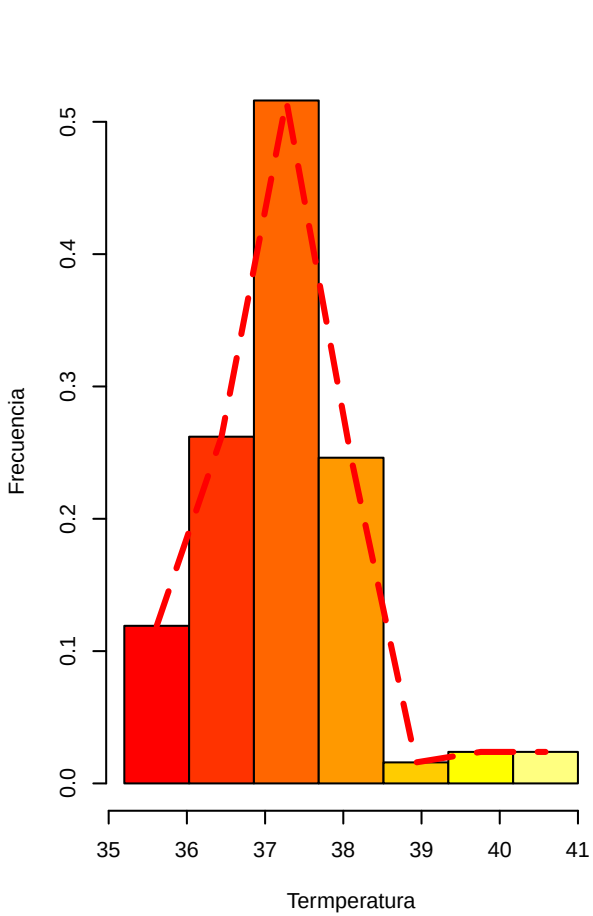


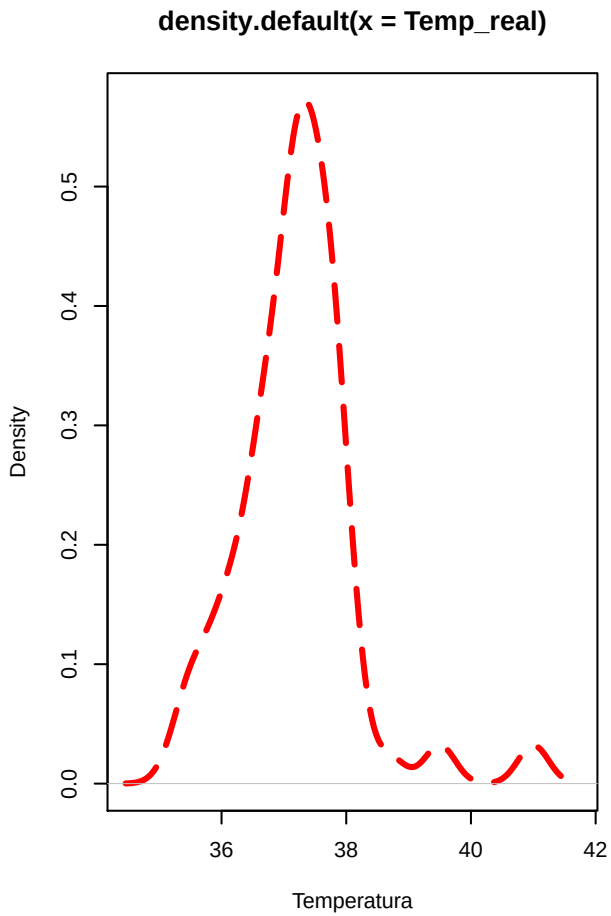


Usar las frecuencias relativas para que el área total sea siempre 1

Histograma, polígono de frecuencias y gráfico de densidad







Tablas de frecuencias acumuladas y relativas.

Vamos a trabajar con los datos del fichero PIMA

```
library(MASS)
library(knitr)
kable(head(Pima.tr))
```

npreg	glu	bp	skin	bmi	ped	age	type
5	86	68	28	30.2	0.364	24	No
7	195	70	33	25.1	0.163	55	Yes
5	77	82	41	35.8	0.156	35	No
0	165	76	43	47.9	0.259	26	No
0	107	60	25	26.4	0.133	23	No
5	97	76	27	35.6	0.378	52	Yes

Y para calcular, por ejemplo, el percentil 60 de la variable `npreg`, usamos `quantile`.

```
quantile(Pima.tr$npreg, probs = 0.6)
```

```
## 60%
##   3
```

¿Cómo lo calcularías "a mano"?

Frecuencias acumuladas y relativas acumuladas.

- Dada la tabla de frecuencias absolutas de una variable cuantitativa discreta

Valor	x_1	x_2	$\cdots$	x_k
Frecuencia	f_1	f_2	$\cdots$	f_k

  las **frecuencias acumuladas** se definen así:

$$F_1 = f_1, \quad F_2 = F_1 + f_2, \quad F_3 = F_2 + f_3, \ \ldots, F_k = F_{k-1} + f_k.$$

  La frecuencia acumulada de x_k indica *cuántos* valores de la muestra son $\leq x_k$.

- Las **frecuencias relativas acumuladas** se definen dividiendo las acumuladas por n (total datos de la muestra).

$$F_1' = \frac{F_1}{n}, \quad F_2' = \frac{F_2}{n}, \ \ldots, F_k' = \frac{F_k}{n}.$$

  La frecuencia relativa acumulada de x_k sirve para calcular *qué porcentaje* de valores de la muestra son $\leq x_k$.

Ejemplo de cálculo:

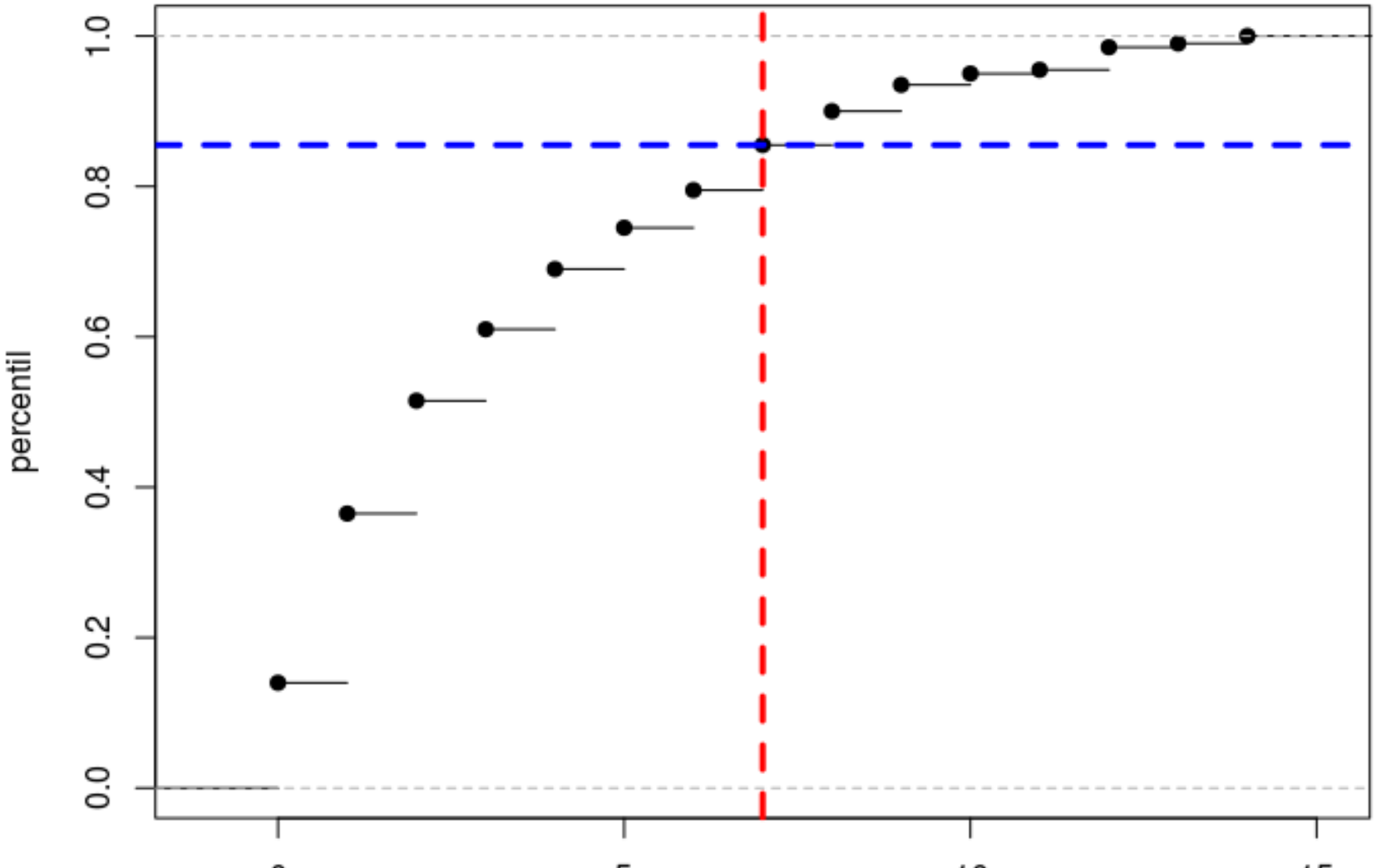
En la práctica 1 aprenderemos a calcular con R las tabla de frecuencias absolutas, relativas, acumuladas y relatativas acumuladas de la variable de la variable `npreg` en el fichero `Pima.tr`

Valor	Frec.Absoluta	Frec.Relativa	Frec.Acumulada	Frec.Rel.Acumulada
0	28	0.140	28	0.140
1	45	0.225	73	0.365
2	30	0.150	103	0.515
3	19	0.095	122	0.610
4	16	0.080	138	0.690
5	11	0.055	149	0.745
6	10	0.050	159	0.795
7	12	0.060	171	0.855
8	9	0.045	180	0.900
9	7	0.035	187	0.935
10	3	0.015	190	0.950
11	1	0.005	191	0.955
12	6	0.030	197	0.985
13	1	0.005	198	0.990
14	2	0.010	200	1.000

¿Ves la relación entre frecuencias relativas acumuladas y percentiles?

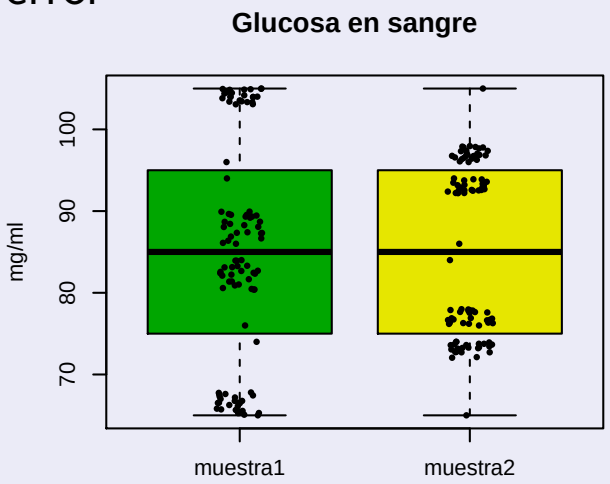
Gráfico de frecuencias relativas acumuladas

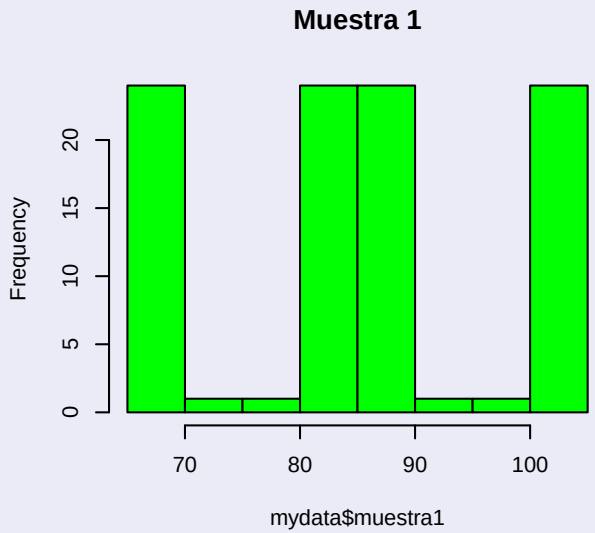
o función empírica de frecuencias/distribución

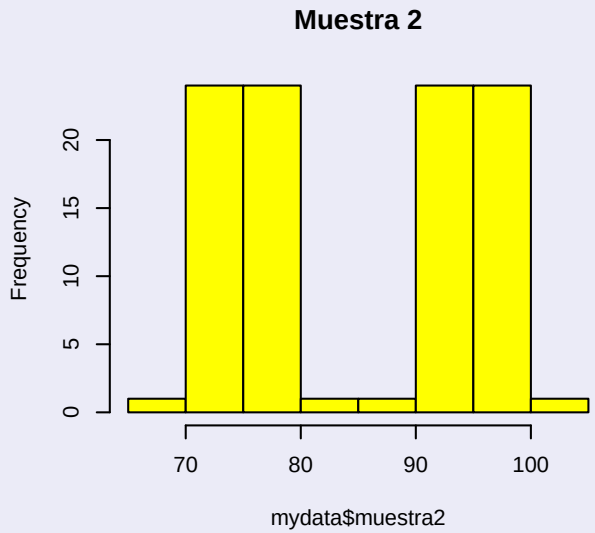


Más allá del `boxplot`.

El boxplot "oculta" información sobre la distribución de los datos, lo que puede inducir a error

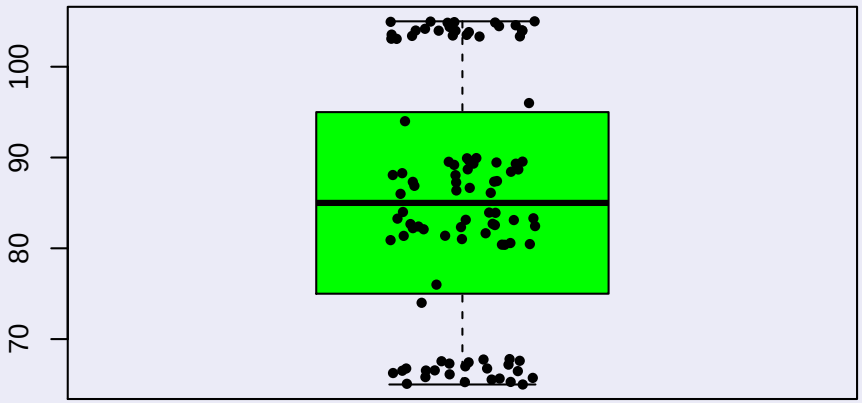


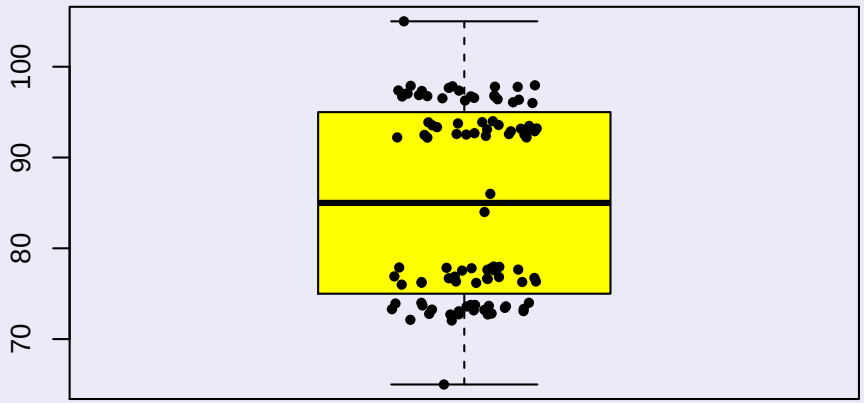


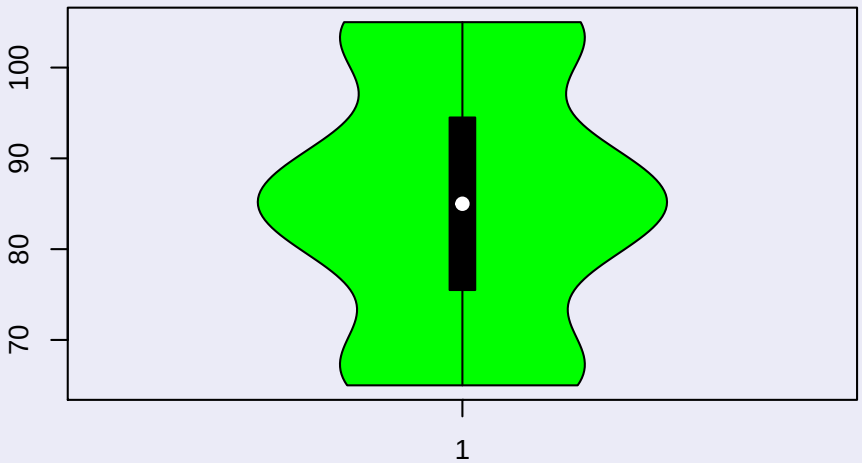


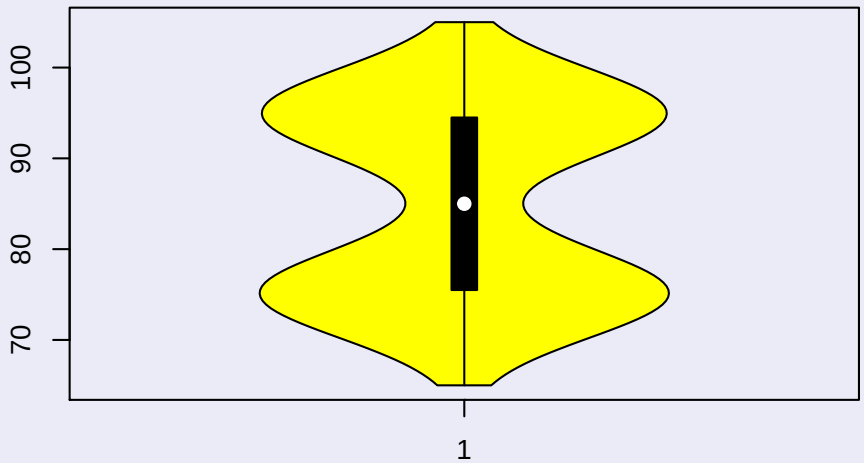
Solución: los gráficos de violín.

Una alternativa al boxplot que ha ganado popularidad recientemente es el *violinplot*:









Este tipo de gráficos combina un boxplot con una de densidad de la muestra.

VARIOS

- Medidas relativas
- **Grupos trabajo**
- **Cuestionarios**

VARIOS

- Dos estimaciones del `bmi` medio de las indias de la tabla PIMA con 2 muestras de tamaño 30

```
set.seed(20170920)
m1=sample(Pima.tr$bmi, size = 30, replace = T)
m2=sample(Pima.tr$bmi, size = 30, replace = T)
mean(m1)
```

```
## [1] 31.99333
```

```
mean(m2)
```

```
## [1] 34.44333
```

- Dos rectas de regresión a partir de dos muestras diferentes (GeoGebra)
- **Leer fichero de datos sobre osteoporosis (práctica 1)**. Copia y pega el siguiente código en tu script

```
# descomenta y ejecuta para limpiar la memoria (de R)
# rm(list = ls())
os="http://www3.uah.es/marcos_marva/sanitaria1718/datos/osteoporosis.csv"
osteo = read.table(file = os, sep = "\t", header = TRUE, dec = ",")
```