COMENZAMOS EL ANÁLISIS DE DATOS

Llamamos

lote

a un conjunto de números similares, obtenidos de alguna manera, no hablamos de muestra aleatoria.

Ejemplos simples son

- los pesos de 21 estudiantes de un curso,
- el total de lluvia caída, en un lugar elegido, en cada año de los últimos 10 años,
- el total de ventas de este año de corredores de seguros de vida entre los 14 que más vendieron el año pasado,
- la cantidad de cortes de luz durante la última década en 11 circunscripciones de la Capital Federal,
- la cantidad de garrapatas halladas en cada una de 49 ratas.

Muchas veces interesa tener una descripción **global** de cada uno de estos lotes.

Este tipo de estructuras simples pueden tener características no fácilmente discernibles mirando los números.

LECTURA DE DATOS

Ejemplo 1. Consideremos los datos, que muestra la tabla 1. Se trata 59 puntos de fusión en ^{0}C obtenidos para distintas ceras naturales.

Tabla 1. Puntos de fusión de distintas ceras naturales

63.78	63.45	63.58	63.08	63.40	64.42	63.27	63.10
63.34	63.50	63.83	63.63	63.27	63.30	63.83	63.50
63.36	63.86	63.34	63.92	63.88	63.36	63.36	63.51
63.51	63.84	64.27	63.50	63.56	63.39	63.78	63.92
63.92	63.56	63.43	64.21	64.24	64.12	63.92	63.53
63.50	63.30	63.86	63.93	63.43	64.40	63.61	63.03
63.68	63.13	63.41	63.60	63.13	63.69	63.05	62.85
63.31	63.66	63.60					

Fueron obtenidos del estudio, realizado por White, Riethof y Kushnir (1960), con el objetivo de investigar métodos químicos para detectar la presencia de ceras sintéticas adicionadas a las ceras naturales de abeja.

El agregado de cera microcristalina eleva el punto de fusión de la cera de abeja.

Si todos los tipos de cera de abeja tuviesen el mismo punto de fusión, su determinación sería un procedimiento razonable para detectar diluciones.

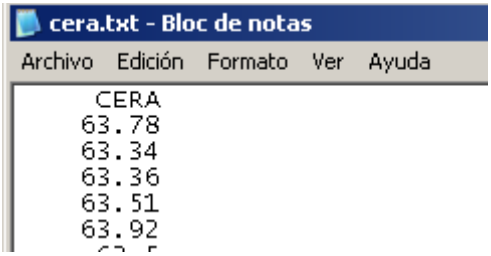
Sin embargo, el punto de fusión y otras propiedades químicas de la cera de abeja varían de una colmena a otra.

Los autores obtuvieron muestras de cera pura de abejas de 59 fuentes, midieron varias propiedades químicas y examinaron la variabilidad de las mediciones.

Mostraron que el agregado de 5% de cera microcristalina aumentaba el punto de fusión de la cera de abeja en .85 ^{0}C y que el agregado de 10% aumentaba el punto de fusión en 2.22 ^{0}C.

Desde la ventana de comandos o la ventana de escritura (script)

Los datos de la tabla, están en un archivo texto “cera.txt” que tiene



los **valores en columna** y en la **primera fila el nombre “CERA’**. A partir de ellos se generó **`Cera`** que es un objeto de R (un data frame) mediante la siguiente instrucción:

```
> Cera <-
read.table("E:\\diana\\cursos\\Análisis de
Datos\\Datan 2008\\Datos\\cera.txt",header=T)
```

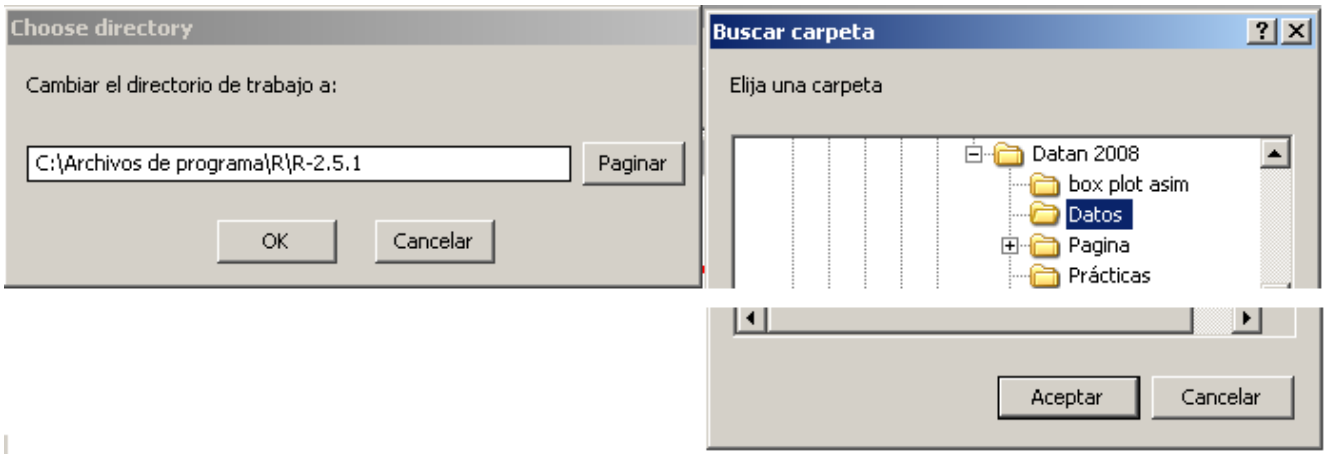
```
> Cera
    CERA
1  63.78
2  63.34
3  63.36
4  63.51
```

```
5  63.92
.........
```

¡Cuidado con los nombres!: el data frame **`Cera`** contiene una columna con nombre **`CERA`**

Veamos otra forma de leer el archivo "cera.txt" , esto es más conveniente cuando la dirección donde está el archivo es larga como ocurrió en este ejemplo.

En R cambiando el directorio de trabajo, al directorio donde se encuentra el archivo con los datos

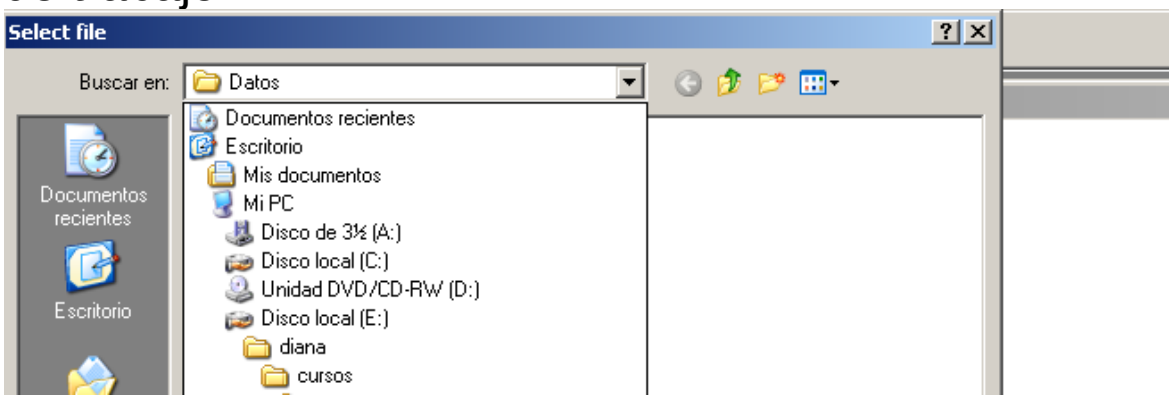



```
> cera <- read.table ("cera.txt")
```

En esta asignación hemos creado el marco de datos (data frame) con nombre **`cera`** que contiene los mismos datos y estructura que **`Cera`**

Si los datos están en una subcarpeta es mejor utilizar la función **`file.choose ()`** para seleccionar el archivo

La función **`file.choose ()`** abre una ventana que permite seleccionar el archivo sin previamente cambiar el directorio de trabajo



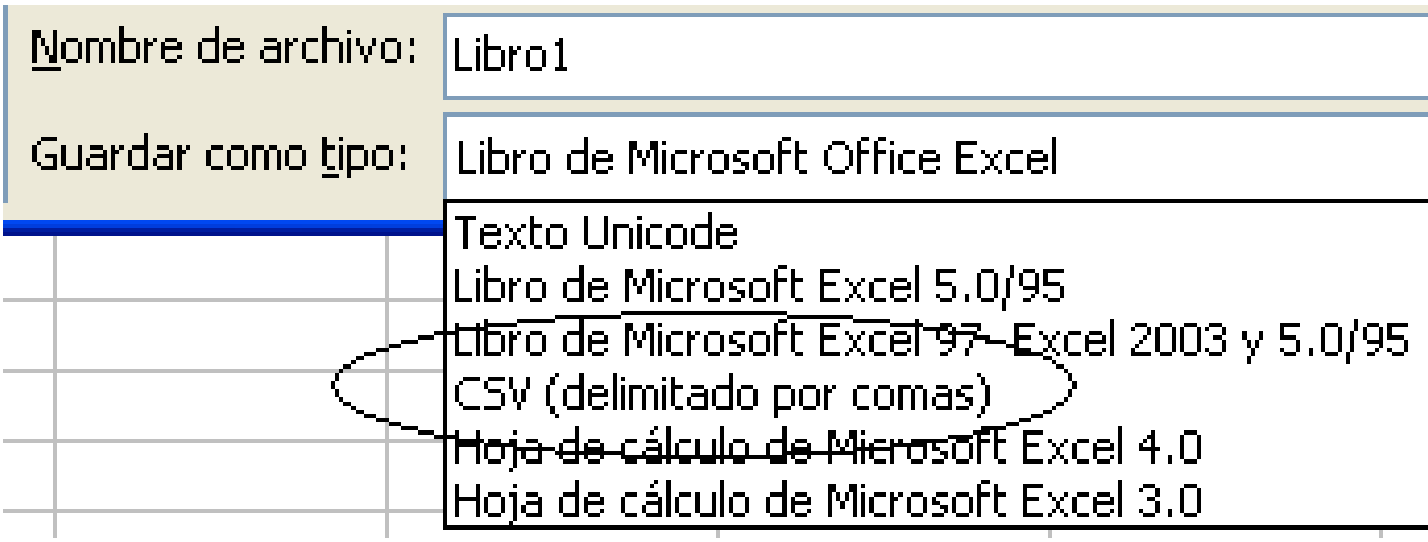

```
cera2 <- read.table(file.choose(),header=T)
```

ó

```
path <- file.choose()
cera2 <- read.table( path, header=T)
```

Lectura de datos provenientes de una planilla excel

1) Exportar desde el **excel** los datos en formato ”csv”, utilizando la opción **guardar como**



2) leer los datos

```
datos <-
read.csv2(file.choose())#excel en
castellano
datos <- read.csv
(file.choose())#excel en inglés
```

vea como le quedaron las primeras 10 filas

```
datos[1:10,]
```

Escritura de datos

1) A un archivo texto.

write.table (datos[1:10,],file= file.choose())# elijo nombre.txt

2) A un archivo csv

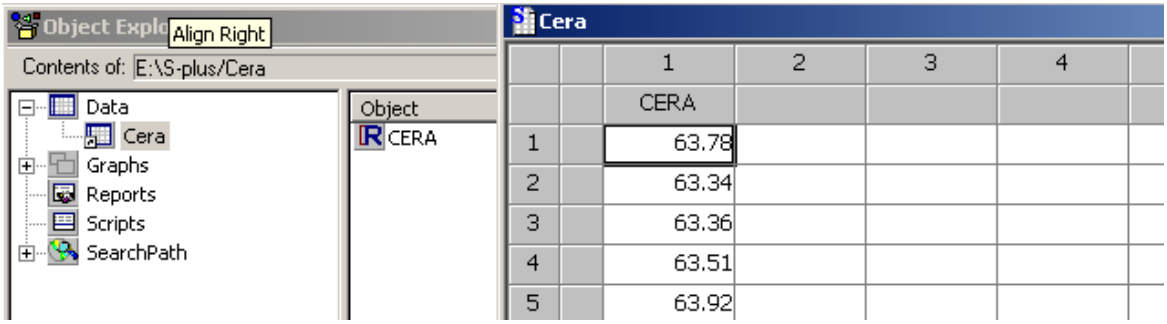
write.csv (datos[1:10,],file= file.choose())# elijo nombre.csv para excel en inglés

write.csv2 (datos[1:10,],file= file.choose())# elijo nombre.csv para excel en castellano

En S-plus siguiendo menues:

File -> Import Data -> From File -> Browse -> Seleccionar -> OK

-> File Format (ASCII..txt) ; . create new data set



Histogramas

El histograma es un método largamente utilizado para presentar los datos. Muestra la forma de la distribución de los datos de la misma manera que la función de densidad muestra las probabilidades.

El rango de los valores de los datos es dividido en **intervalos-bins** y se grafica la cantidad o proporción de observaciones que caen dentro de cada intervalo.

La figura 1 muestra tres histogramas de los puntos de fusión de ceras puras correspondientes a los datos de la tabla 1 que fueron obtenidos mediante las siguientes instrucciones:

```
        par(mfrow=c(3,1)) # ventana gráfica
dividida en tres filas y una columna

        hist(Cera$CERA,nclass=25)
        hist(Cera$CERA)
        hist(Cera$CERA,nclass=4)
```

El parámetro **`nclass`** da una cantidad “sugerida” de clases para la función **`hist.`**

Obtenemos intervalos con longitud creciente. Si el ancho de los intervalos es muy pequeño el histograma resultante es muy irregular; si es muy grande la forma está sobresuavizada y oscurecida.

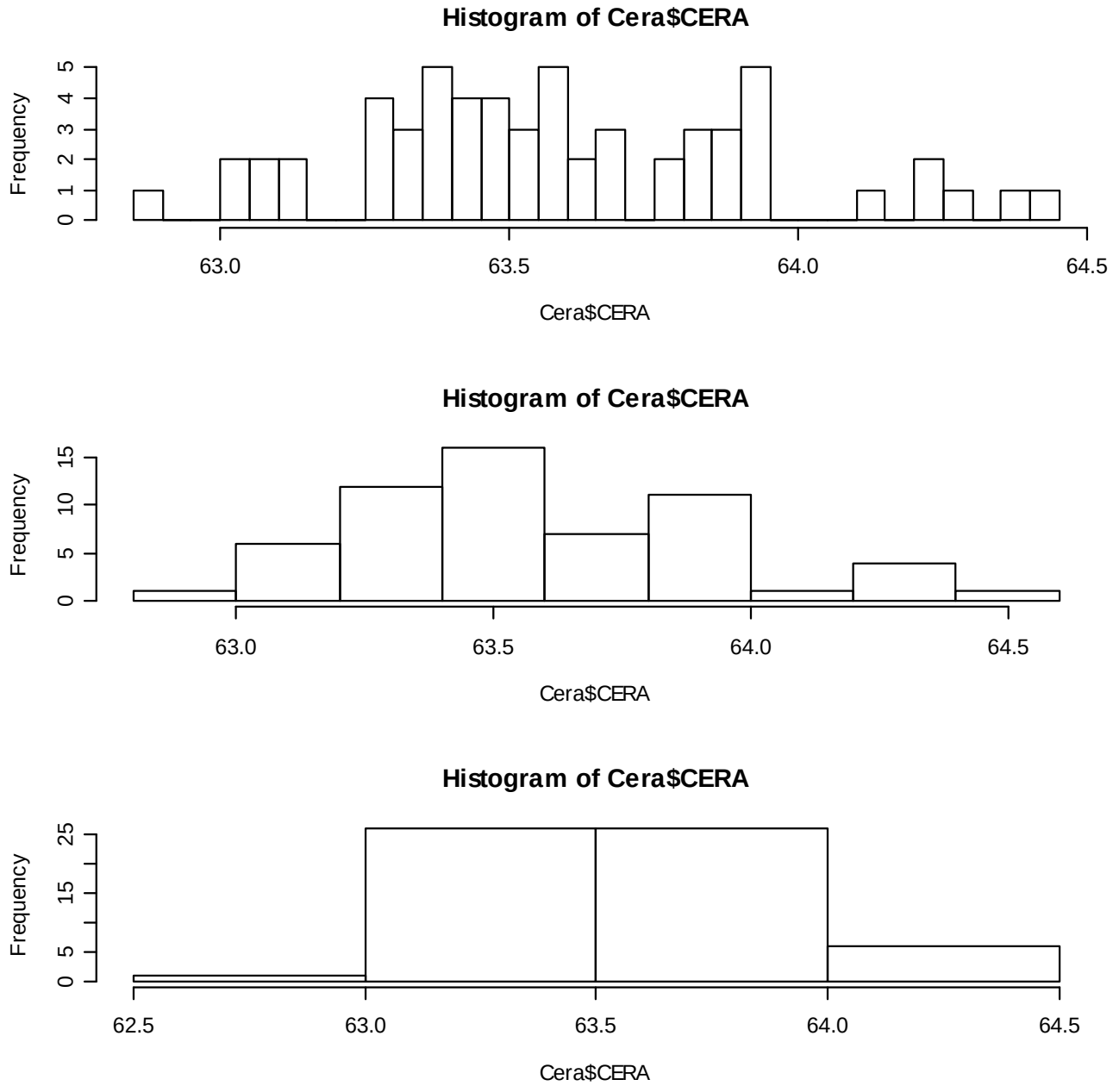


Figura 1. tres histogramas de los puntos de fusión de ceras naturales

Del primero de los histogramas, podemos ver que el agregado de 5% de cera microcristalina (aumenta el punto de fusión en .85 ^{0}C) puede ser muy difícil de detectar especialmente si fue realizado en ceras con bajo punto de fusión, pero el agregado de 10% (aumenta el punto de fusión en 2.22 ^{0}C) sería detectable.

Los histogramas son histogramas de frecuencias absolutas y tienen distintas escalas en los ejes.

Si los intervalos no tienen la misma longitud el histograma resultante puede ser engañoso.

Por lo tanto se recomienda graficar la **proporción de observaciones dividida la longitud del intervalo**;

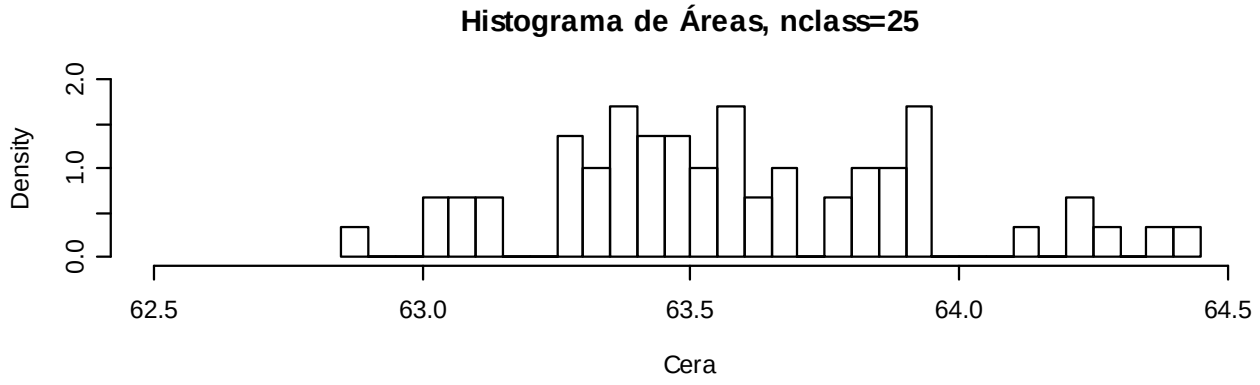
En este caso, el área total bajo el histograma es 1 y la altura indica la ***densidad*** *relativa* de los datos sobre el eje horizontal.

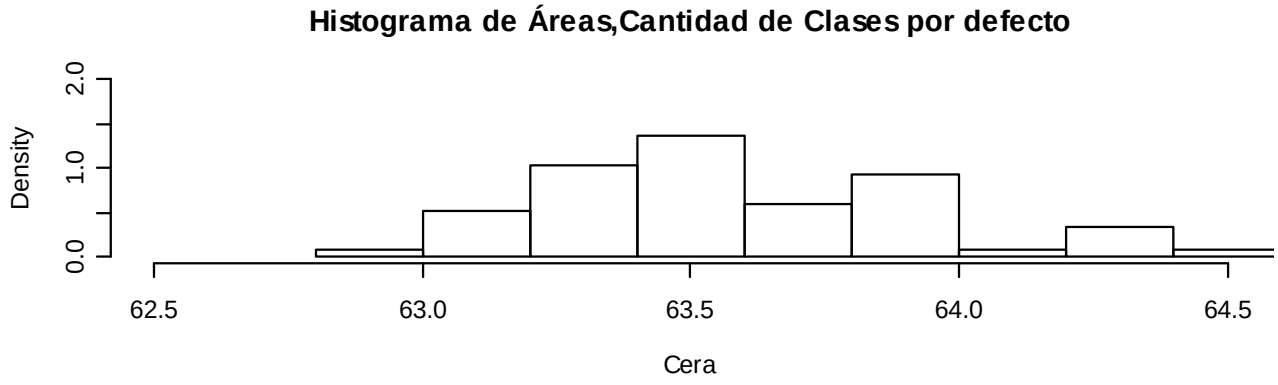
Fijamos las escalas, construimos un histograma de áreas y cambiamos las etiquetas del eje x:

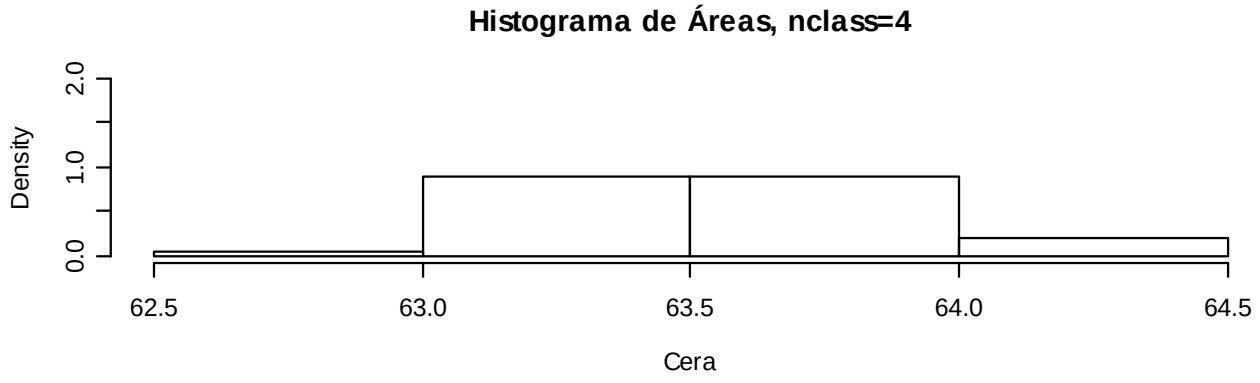
```
hist(Cera$CERA,nclass=25,probability=T,
    main="Histograma de Áreas, nclass=25",
    xlab= "Cera"
       ,xlim=c(62.5,64.5),ylim=c(0,2))


hist(Cera$CERA,probability=T,
  main="Histograma de Áreas,Cantidad de
  Clases por defecto",
xlab= "Cera"
       ,xlim=c(62.5,64.5),ylim=c(0,2))
```

```
hist(Cera$CERA,nclass=4,probability=T,
    main="Histograma de Áreas, nclass=4",
 xlab= "Cera"
      ,xlim=c(62.5,64.5),ylim=c(0,2))
```







La elección del ancho de los intervalos, o la determinación de la cantidad de los mismos, es generalmente realizada en forma subjetiva en un intento de obtener un balance entre un histograma muy irregular y uno muy suavizado.

Reglas para la cantidad de intervalos

La tabla siguiente muestra el número de intervalos sugeridos por tres reglas para valores elegidos de la cantidad n de datos entre 10 y 300. Las reglas proponen: tome la parte entera de

i) $10 \log_{10} n$, Dixon y Kronmal(1965)
ii) $2\, n^{1/2}$, Velleman (1976)
iii) $1 + \log_2 n$, Sturges (1926).

Tabla 1

	Regla (Parte entera de)		
n	Dixon y Kronmal $10 \log_{10} n$	Velleman $2\, n^{1/2}$	Sturges $1 + \log_2 n$
10	**10.0**	**6.3**	**4.3**
20	**13.0**	**8.9**	**5.3**
30	**14.7**	**10.9**	**5.9**
40	**16.0**	**12.6**	**6.3**
50	**16.9**	**14.1**	**6.6**
75	**18.7**	**17.3**	**7.2**
100	**20.0**	**20.0**	**7.6**
150	**21.7**	**24.4**	**8.2**
200	**23.0**	**28.2**	**8.6**
300	**24.7**	**34.6**	**9.2**

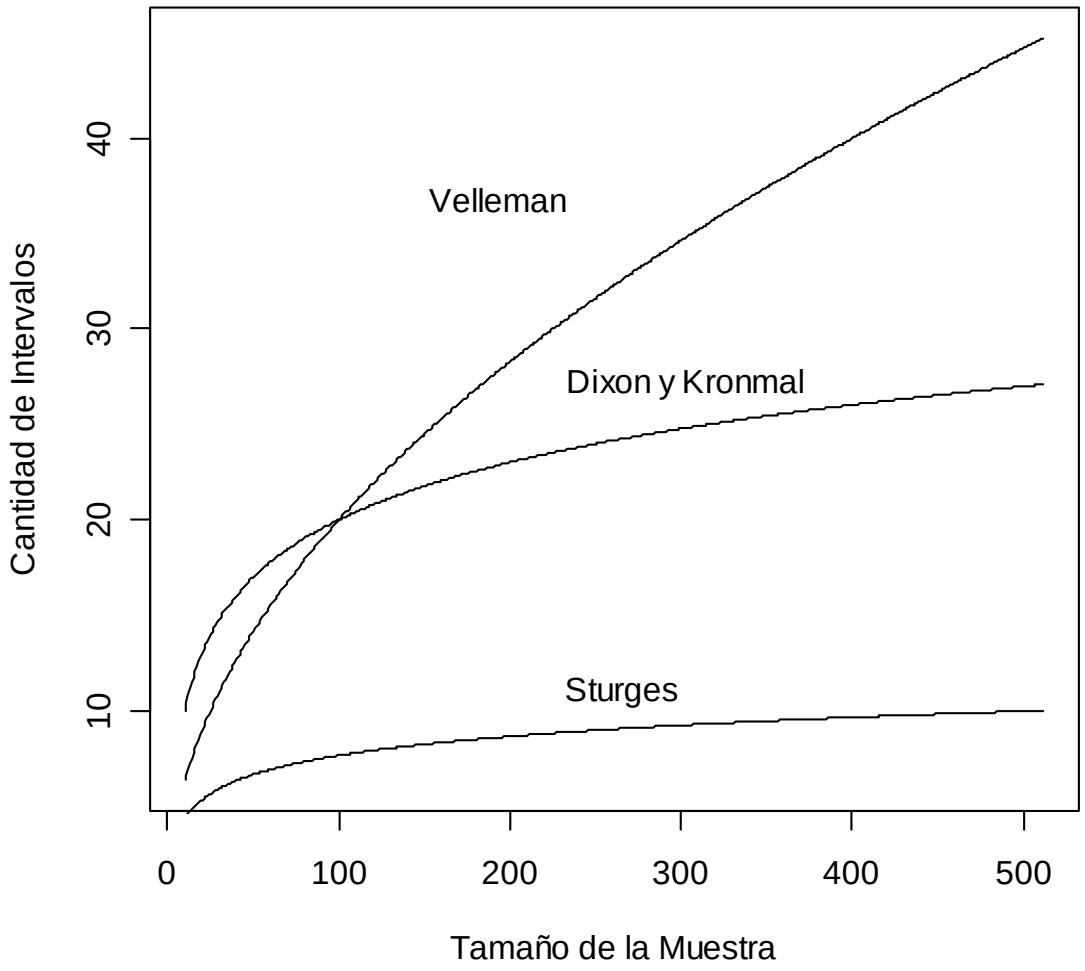
```
 16    12.0       8.0          5
 32    15.1      11.3          6
 64    18.1      16.0          7
128    21.1      22.6          8
256    24.1      32.0          9
512    27.1      45.3         10
```

La primera, L = [10 $\log_{10} n$], da una cota superior (si n<100) para la cantidad de intervalos que es generalmente bastante efectiva en la práctica (en lo que sigue [ ] indica parte entera).

Sin embargo, puede interesar tener una menor cantidad de intervalos cuando *n* es pequeño (digamos 50 o menos). Para ese caso Velleman sugirió utilizar L = [2 $n^{1/2}$].

```
ejex<- seq(10,512)
velle <- 2*sqrt(ejex)
sturges <- 1+logb(ejex,2)
dixkron <- 10*log10(ejex)
par(csi=0.24)
plot(ejex,velle,lty=1, type="l",
     xlab="Tamaño de la Muestra",
     ylab="Cantidad de Intervalos")
lines(ejex,sturges,lty=1, type="l")
lines(ejex,dixkron,lty=1, type="l")

leg1<- c("Velleman","Dixon y Kronmal",
  "Sturges")
text(locator(1),leg1[1])
text(locator(1),leg1[2])
text(locator(1),leg1[3])
```



Reglas para la longitud del intervalo

Por la regla de Sturges la longitud resulta

$$\text{rango(datos)} / (1 + \log_2 n)$$

y esto es frecuentemente muy grande.

Los valores atípicos, outliers, pueden agrandar dramáticamente el rango y así aumentar el tamaño de los intervalos.

Dos propuestas implementadas en S-plus

i) $h_n = 3.49\, s\, n^{-1/3}$ Scott (1979),

ii) $h_n = 2\, R\, n^{-1/3}$ Freedman & Diaconis (1981),

R es el rango intercuartil.

ii) da intervalos un poco más pequeños que i)

En el caso de datos gaussianos estándar, (*s* = 1 y *R* = 1.349), para todos los tamaños muestrales tenemos que la relación entre los anchos de los intervalos es 3.49 / 2.698.

Las opciones de cantidad de clases de la función **`hist`** incluyen las reglas de Scott y Friedman

en R hist(Cera$CERA,nclass=nclass.FD)
Las opciones son
nclass.Sturges
nclass.scott
nclass.FD(x)

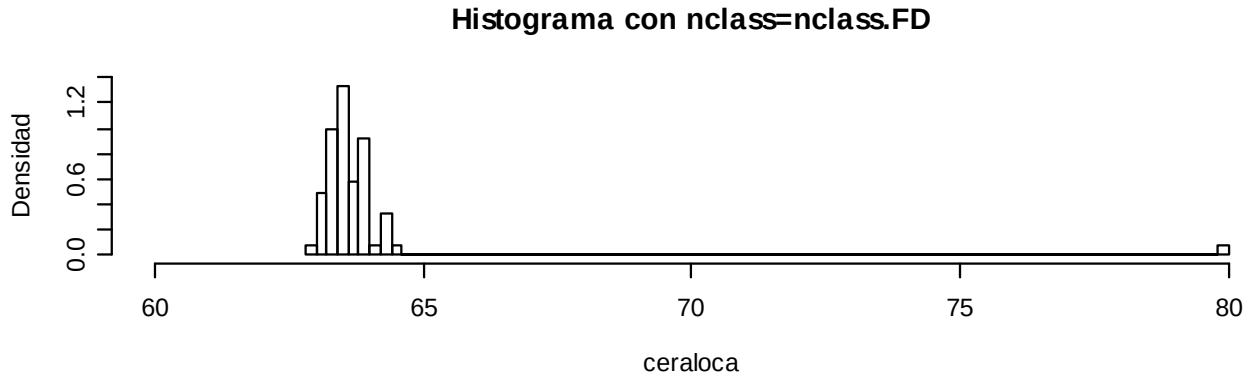
en S-plus

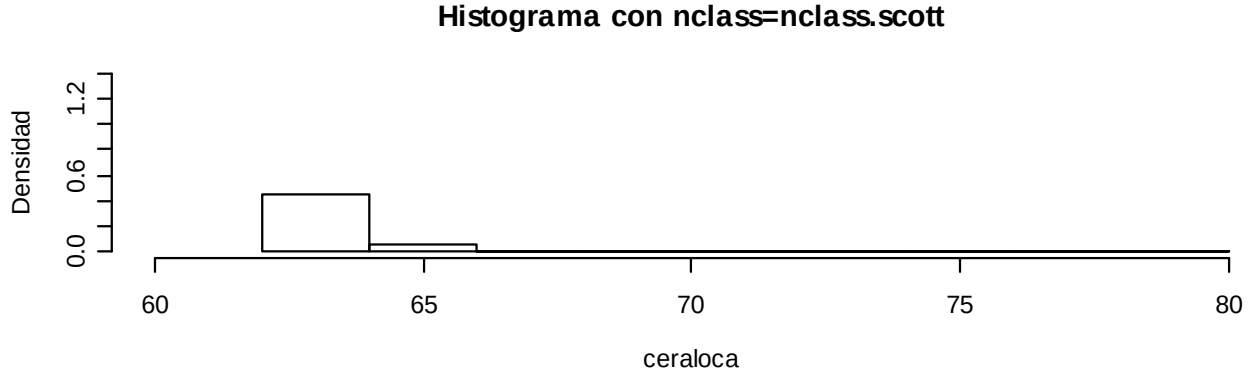
```
hist(x, nclass=<<see below>>, .....)
```

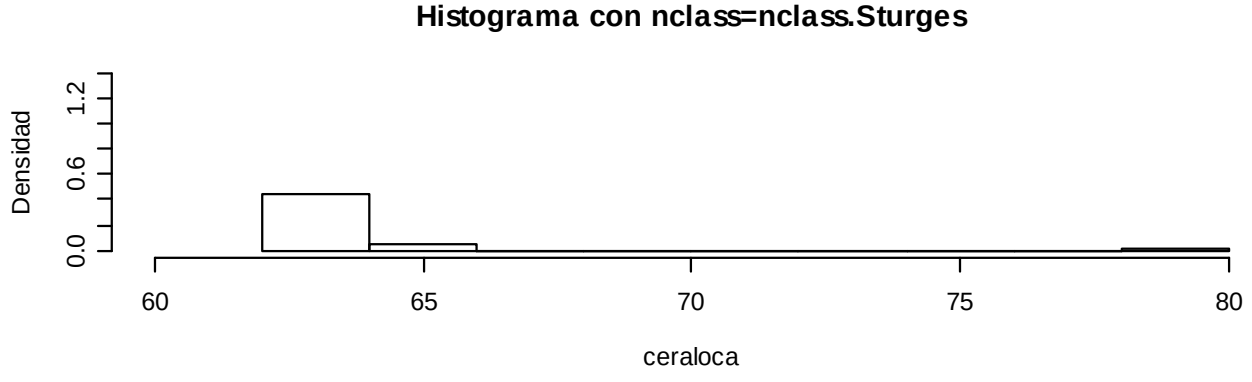
nclass

recommendation for the number of classes (i.e., bars) the histogram should have. This may be an integer, a function to apply to x which returns an integer, or a character string specifying which built-in method to use. Available methods for calculating the number of classes are Sturges (`sturges`), Freedman-Diaconis (`fd`), and Scott (`scott`).

Las figuras siguientes muestran el comportamiento que tienen las distintas reglas para la longitud de los intervalos cuando agregamos un valor atípico al conjunto de datos cera.







Vemos que tanto la regla de Sturges (valor por defecto) como la de Scott producen, con el agregado de un valor atípico, intervalos de clase con longitud demasiado grande que oculta la estructura de los datos.

Las siguientes instrucciones permiten obtener los histogramas anteriores:

```
ceraloca <- c(Cera$CERA,80)
par(mfrow=c(3,1))
hist(ceraloca,nclass=nclass.FD,xlim=c(60,80),ylim=c(0,1.4),
  main="Histograma con nclass=nclass.FD", probability=T,
  ylab="Densidad")
hist(ceraloca,nclass=nclass.scott,xlim=c(60,80),ylim=c(0,1.4),
  main="Histograma con nclass=nclass.scott",probability=T,
  ylab="Densidad")
hist(ceraloca,nclass=nclass.Sturges,xlim=c(60,80),ylim=c(0,1.4),
  main="Histograma     con     nclass=nclass.Sturges",probability=T,
ylab="Densidad")
```

El aspecto más interesante de las dos reglas para el ancho del intervalo, h_n, quizás sea que ambas dependen fundamentalmente de

$$n^{-1/3}.$$

Si transformáramos ese ancho de intervalo en un número de clases sugeridas tendríamos un comportamiento de

$$n^{1/3},$$

una forma funcional entre

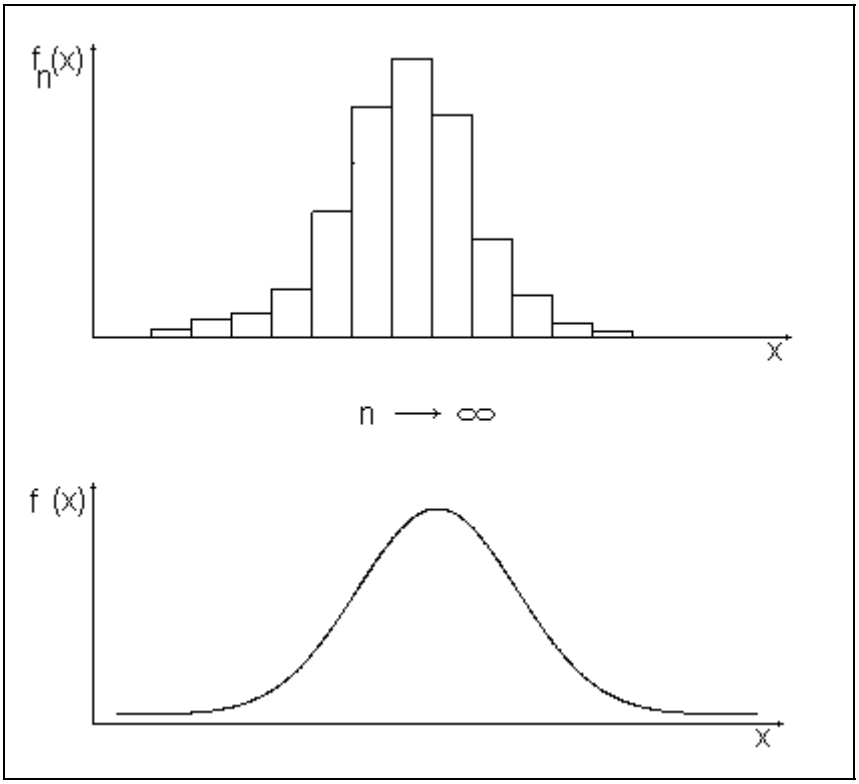
$$\log(n) \text{ y } n^{1/2}.$$

La función histograma estándar **`hist(x,..)`** por defecto toma la fórmula de Sturges (1926),$1 + \log_2 n$.

Histogramas de Áreas - Curvas de densidad.

En un histograma de áreas, la altura de cada rectángulo de clase es la proporción de puntos que fueron observados en el intervalo de clase dividida por la longitud del intervalo, esto es la densidad de puntos por unidad y el área total es 1.

Si los valores que obtenemos provienen de una variable continua, y aumentamos el tamaño de la muestra, (siempre que los datos estén elegidos en forma adecuada si (muestra aleatoria)), podremos reducir la longitud de los intervalos de clase. De esta manera, a medida que aumenta la cantidad de datos las alturas de los intervalos va formando una curva cada vez más suave. La curva límite es llamada curva de densidad de la variable correspondiente.



Estimación de una curva de densidad.

Los histogramas se utilizan frecuentemente para describir datos para los cuales no se ha realizado ningún tipo de supuesto. Si los datos se modelan como una muestra aleatoria proveniente de alguna distribución continua. el histograma de áreas puede ser considerado como un estimador de la función de densidad de probabilidad.

Sesgo y Varianza del Histograma como estimador de una densidad.

En lo que sigue supondremos que X1, ..., Xn son v.a.i.i.d con función de densidad de probabilidad desconocida f.

La construcción de un histograma se caracteriza por los siguientes pasos:

- Dividir la recta real en intervalos o bins

$$Bj = [x0 + (j-1)h, x0 + jh) \text{ , con } j \in Z$$

donde x0 es el origen del histograma y h > 0 es la longitud de los intervalos, también llamada ancho de banda (bandwidth).

- Contar cuantos datos caen en cada intervalo

Más formalmente podemos escribir

Histograma
$\hat{f}_h(x) = (nh)^{-1} \sum_{i=1}^{n} \sum_{j} I(X_i \in B_j) I(x \in B_j)$

Observación: se han modificado las frecuencias de manera que el área de los rectángulos del histograma sea 1.

Sesgo

Sin pérdida de generalidad supondremos x0 = 0, y que x $\in$ Bj. O sea a partir de ahora j = j (x) y por lo tanto

$$\hat{f}_h(x) = (nh)^{-1} \sum_{i=1}^{n} I(X_i \in B_j)$$

Como los Xi están idénticamente distribuidos

$$E(\hat{f}_h(x)) = (nh)^{-1} \sum_{i=1}^{n} E\left[I(X_i \in B_j)\right] =$$

$$= (nh)^{-1} nE\left[I(X \in B_j)\right]$$

$$= h^{-1} \int_{(j-1)h}^{jh} f(u)du$$

En general el último término no coincidirá con $f(x)$ salvo que $f(x)$ = cte en B_j.

El *sesgo del histograma en x*, está definido por:

$$E(\hat{f}_h(x)) - f(x)$$

Luego

$$Sesgo(\hat{f}_h(x)) = h^{-1}\int_{B_j} f(u)du - f(x)$$

Veamos con más detalle cómo depende el sesgo de h y de la densidad.

Consideremos en particular el caso de *f* lineal en el intervalo B_j que estamos considerando:

$$f(x) = a + cx \quad x \in B_j \quad a, c \in R$$

Es fácil ver que en este caso

$$Sesgo(\hat{f}_h(x)) = h^{-1}\int_{B_j} (f(u) - f(x))du$$

$$= c((j - 1/2)h - x)$$

Podemos escribir *c* = $f'((j-1/2)h)$,

es decir, como la derivada de la densidad en el punto medio (j-1/2)h del intervalo B_j. Por lo tanto

$$Sesgo(\hat{f}_h(x)) = ((j-1/2)h - x) f'((j-1/2)h)$$
$$= O(h)O(1)$$
$$= O(h)$$

En general *f* no es lineal pero un desarrollo de Taylor de primer orden de *f* reduce el problema al caso lineal. El sesgo del histograma está dado por

Sesgo del Histograma
$Sesgo(\hat{f}_h(x)) = ((j-1/2)h - x) f'((j-1/2)h)$ $+ o(h) \qquad h \to 0$

Del primer sumando resulta que el sesgo del histograma decrece en el orden de O(h) a medida que la longitud de los intervalos, h, decrece. El *o(h)* viene del desarrollo de Taylor.

Varianza

Estudiaremos a continuación la estabilidad de la estimación, medida por la varianza.

Como las Xi son iid la función indicadora,

$$I(X_i \in B_j)$$

es una variable Bernouilli con varianza *p (1-p)* donde

$$p = P(X_i \in B_j) = \int_{B_j} f(u)du$$

Por lo tanto

$$Var(\hat{f}_h(x)) = Var((nh)^{-1}\sum_{i=1}^{n} I(X_i \in B_j))$$

$$= (nh)^{-2}\sum_{i=1}^{n} Var(I(X_i \in B_j))$$

$$= n^{-1}h^{-2}(\int_{B_j} f(u)du)(1-\int_{B_j} f(u)du)$$

$$= (nh)^{-1}(h^{-1}\int_{B_j} f(u)du)(1-O(h))$$

$$= (nh)^{-1}(f(x)+o(1)),\ h \to 0,\ nh \to \infty$$

Varianza del histograma
$Var(\hat{f}_h(x)) = (nh)^{-1} f(x) + o((nh)^{-1}),\ \ nh \to \infty$

Observamos que la varianza es proporcional a *f(x)* y decrece cuando *nh* aumenta. Para un tamaño *n* fijo reducir la varianza implica aumentar la longitud *h* del intervalo. Esto contradice el objetivo de reducir el sesgo mediante la reducción de la longitud h del intervalo, en lo que se suele llamar “variance-bias trade-off”.

Error Cuadrático Medio

El error cuadrático medio es una medida habitualmente utilizada en estadística para evaluar la precisión de un estimador.

$$
\begin{aligned}
&ECM(\hat{f}_h(x) = E\left[(\hat{f}_h(x) - f(x))^2\right] \\
&= E\left[(\hat{f}_h(x) - E((\hat{f}_h(x)) + E((\hat{f}_h(x)) - f(x))^2\right] \\
&= E\left[(\hat{f}_h(x) - E((\hat{f}_h(x)))^2\right] \\
&+ 2E\left[(\hat{f}_h(x) - E((\hat{f}_h(x))\right]\left[E((\hat{f}_h(x)) - f(x)\right]^2 \\
&+ \left[E((\hat{f}_h(x)) - f(x)\right]^2 \\
&= Var\left[\hat{f}_h(x)\right] + \left[Sesgo(\hat{f}_h(x))\right]^2
\end{aligned}
$$

Es la suma de la varianza y el sesgo al cuadrado.

De las expresiones anteriores de sesgo y varianza resulta

Error Cuadrático Medio del Histograma
$$ECM(\hat{f}_h(x) = Var\left[\hat{f}_h(x)\right] + \left[Sesgo(\hat{f}_h(x))\right]^2$$ $$= \frac{1}{nh} f(x) + ((j - 1/2)h - x)^2\left[f'((j - 1/2)h)\right]^2$$ $$+ o(h) + o(\frac{1}{nh})$$

Minimizar el ECM con respecto a h establece un compromiso en el dilema que presenta el “variance-bias trade-off”, es decir

un compromiso en un sobresuavizado (si h es demasiado grande) o un sub suavizado (si h es demasiado chica).

Si $h \to 0, nh \to \infty$ el $ECM(\hat{f}_h(x) \to 0$, por lo tanto el histograma es un estimador consistente de f(x).

Necesitamos que h decrezca para reducir el sesgo pero tiene que haber suficientes observaciones en cada intervalo para mantener baja la varianza.

En la práctica es difícil hallar un h óptimo a partir de la expresión del ECM ya que interviene la densidad desconocida, tanto en la componente de la varianza como en la del sesgo.

En vez de evaluar el error cuadrático en un valor fijo x podríamos estar interesados en una medida de la calidad d el histograma en su totalidad. Es por eso que se define el Error Cuadrático Medio Integrado:

$$ECMI = \int_{-\infty}^{\infty} ECM(\hat{f}_h(x))dx$$

$$ECMI = (nh)^{-1} + h^2/12 \int (f'(x))^2 dx + o(h^2) + o((nh)^{-1})$$

Minimizando los dos primeros sumandos del *ECMI* para el histograma con respecto a h, obtenemos la longitud óptima para el intervalo

$$h_0 = \left(\frac{6}{n \int (f'(x))^2 dx} \right)^{1/3}$$

Observemos que esta solución no es de utilidad práctica ya que depende de la suavidad de la función de densidad y que

$$h_0 \propto n^{-1/3}$$

es coherente con las reglas de Scott y Friedman.

Más detalles en W. HÄRDLE (1990) **Smoothing Techniques with Implementation in S**. Berlin : Springer-Verlag, 1990. (Springer Series in Statistics).