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 sensación **global** de cada uno de estos lotes.

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

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** y se grafica la cantidad o proporción de observaciones que caen dentro de cada intervalo. 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**;

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

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.

LECTURA DE DATOS

Desde la ventana de comandos o la ventana de escritura

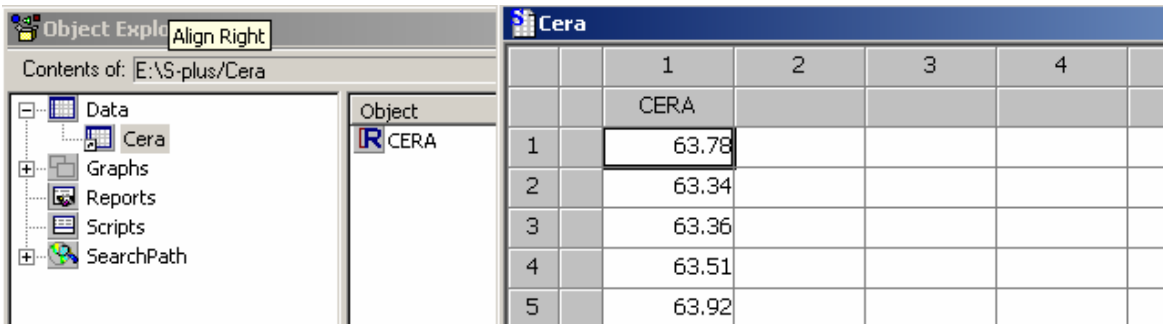
Los datos de la tabla, que están en el directorio raíz en un archivo texto “cera.txt” tiene los **valores en columna** y en la **primera fila el nombre “cera’**, fueron importados mediante la siguiente instrucción, generando un data frame:

```
> Cera <- read.table("c:\\cera.txt",header=T)
```

Si la dirección donde está el archivo es larga, es decir que el archivo está en subcarpetas, es más fácil leerlo **siguiendo menúes:**

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

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



En R

Si los datos están en el directorio raíz:

```
Cera <- read.table("c:\\cera.TXT",header=T)
```

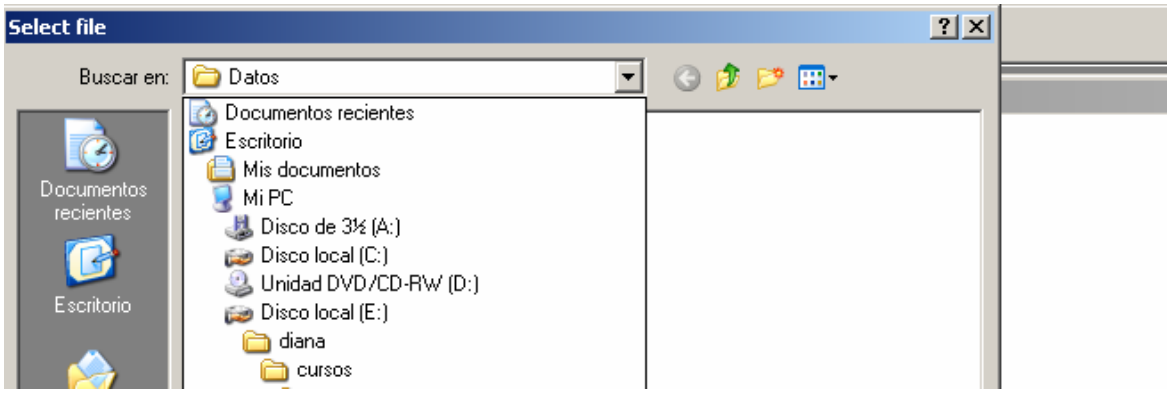
Si los datos están en una subcarpeta es mejor utilizar la función **`file.choose ()`** para seleccionar el archivo

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

ó

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

La función **`file.choose ()`** abre la siguiente ventana que permite seleccionar el archivo.



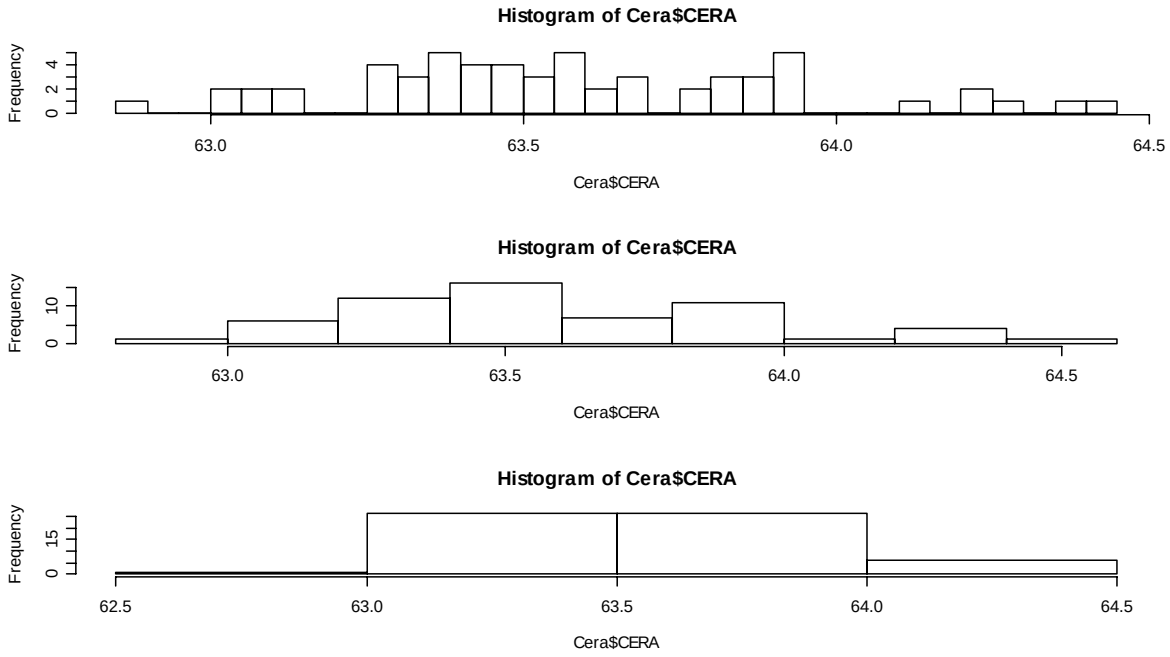


Figura 1. tres histogramas de los puntos de fusión de ceras puras

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))
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.

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.

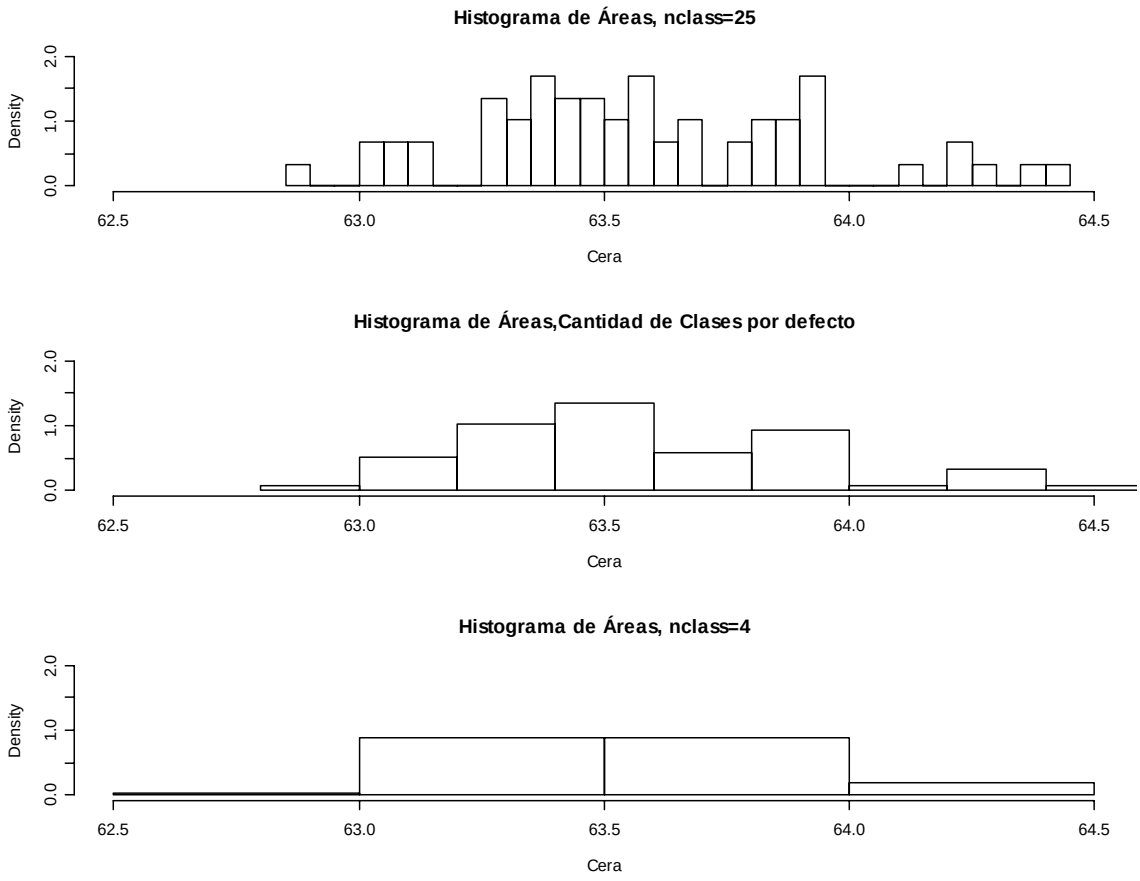
Los histogramas son histogramas de frecuencias absolutas y tienen distintas escalas en los ejes.

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**

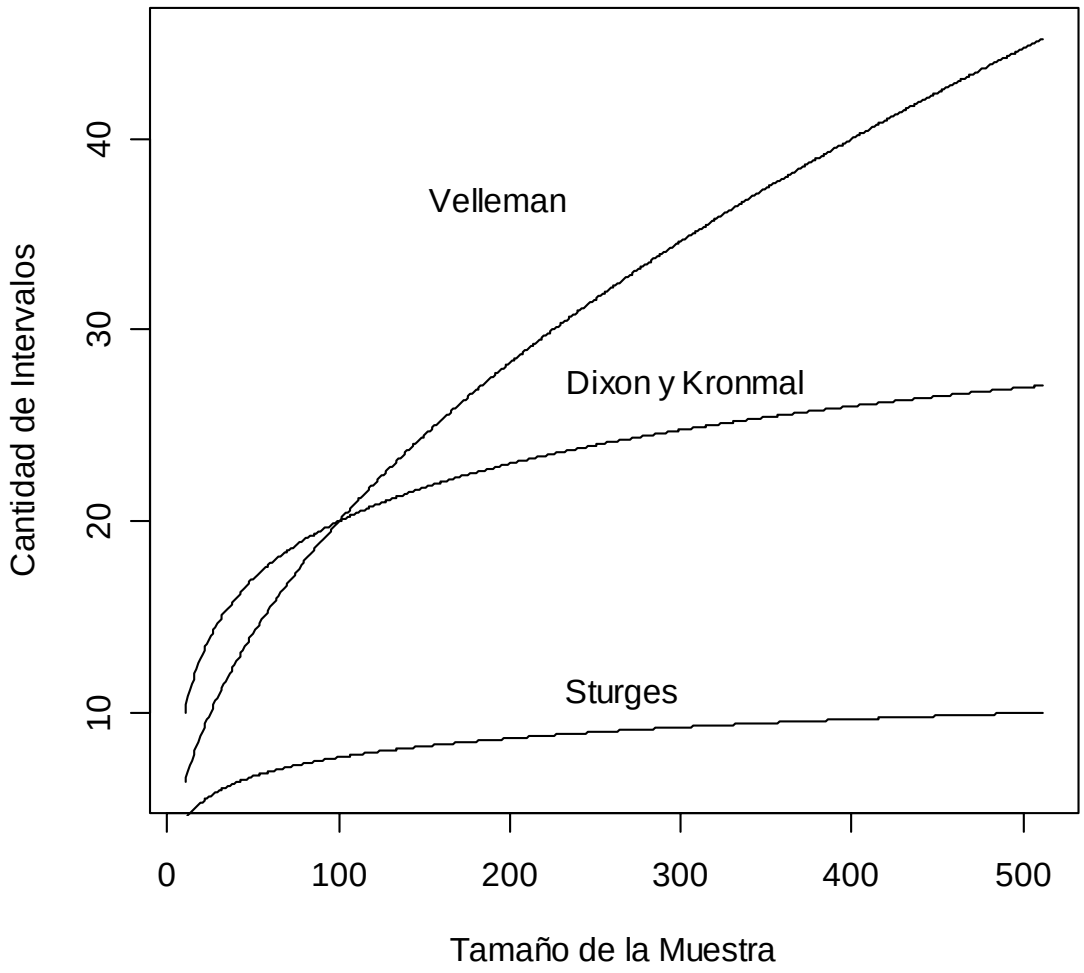
La primera, L = [10 $\log_{10} n$], da una cota superior (si n<100) para la cantidad de intervalos que es generalmente bastante efectiva en la práctica (en lo que sigue [ ] indica parte entera).

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 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`).

en R por ejemplo hist(Cera$CERA,nclass=nclass.FD) Las opciones son

nclass.Sturges(x)
nclass.scott(x)
nclass.FD(x)

El aspecto más interesante de estas 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),$1 + \log_2 n$.

Diagramas de Tallo-Hoja (Stem-and-Leaf)

Un diagrama de tallo-hoja (Tukey, 1977) es un histograma que conserva información numérica.

De manera similar al histograma permite ver el lote como un todo y advertir aspectos como:

- Cuán aproximadamente simétricos son los datos.
- Cuán dispersos están los valores.
- La aparición de valores inesperadamente más frecuentes.

- Si algunos valores están alejados del resto.
- Si hay concentraciones de valores.
- Si hay grupos separados.

Al utilizar los dígitos de los valores de los mismos datos, en vez de simplemente encerrando áreas, ofrece **ventajas**:

- Es más fácil de construir a mano.
- Facilita el ordenamiento de los datos.
- Permite, por lo tanto, hallar la mediana y otras medidas resumen basadas en el lote ordenado.
- Permite ver la distribución de los datos dentro de cada intervalo como patrones dentro de los datos.

Por ejemplo podríamos descubrir que todos los valores son múltiplos de 3.

- Facilita la identificación de una observación y la información que la acompaña.

Figura 2: Puntos de fusión de ceras de abeja **(tabla 1)**

N = 59 Median = 63.53 Quartiles = 63.36, 63.84 Decimal point is 1 place to the left of the colon	El primer dato de la tabla 1 (63.78) aparece en la décima fila de la figura 2 como 637:8. El punto decimal está un

628 : 5 629 : 630 : 358 631 : 033 632 : 77 633 : 001446669 634 : 01335 635 : 0000113668 636 : 0013689 637 : 88 638 : 334668 639 : 22223 640 : 641 : 2 642 : 147 643 : 644 : 02	lugar a la izquierda de los dos puntos (:), esto se indica con "unidad = 0.01 ^{o}C". Los 3 primeros dígitos de los puntos de fusión forman *el tallo*, el cuarto forma *la hoja*. Los tallos están ordenados, en columna, y en líneas separadas, aparecen todos los valores posibles de tallos dentro del rango observado. Las hojas, en cada tallo, son el cuarto dígito de todos los números con ese tallo.

```
> stem(Cera)
```

En su apariencia global el diagrama se asemeja a un histograma con ancho de intervalo igual a 0.1 ^{0}C.

El 95% (56/59) de los puntos de fusión de la cera natural de abeja se encuentra entre 62.9 y 64.3.

Profundidades y cantidad de hojas por tallo

A cada dato se le puede asignar un *rango*, contando desde cada extremo en el lote ordenado.

Por ejemplo, en la figura 3, el 63.03 tiene rango 2 contando desde 62.85 hacia valores crecientes y rango 58 contando desde 64.42 hacia valores decrecientes. La *profundidad* es el menor de los dos valores.

Figura 3

Decimal point is 1 place to the left of the colon

PROF.	# hojas	TALLO	HOJAS
1	1	628	: 5
1	0	629	:
4	3	630	: 358
7	3	631	: 033
9	2	632	: 77
18	9	633	: 001446669
23	5	634	: 01335
	10	635	: 0000113668
26	7	636	: 0013689
19	2	637	: 88
17	6	638	: 334668
11	5	639	: 22223
6	0	640	:
6	1	641	: 2
5	3	642	: 147
2	0	643	:
2	2	644	: 02

La primera columna (PROF.) de profundidad, muestra en cada fila, excepto en la línea central que contiene la mediana, la máxima profundidad correspondiente a los datos de esa fila. Facilita hallar estadísticos de orden.

La segunda columna (# hojas) da la cantidad de hojas en cada tallo.

```
> stem(Cera,depth=T)
```

Observación: Los diagramas de tallo-hoja no son adecuados para datos cuyo rango tiene varios órdenes de magnitud, en esos casos es conveniente construir un diagrama tallo-hoja para el logaritmo de los datos.

Organización del esquema

$$L = [10 \times \log_{10} n\ ]$$

Esta regla parece dar esquemas efectivos sobre el rango $20 \leq n \leq 300$. Los valores n menores que 20 pueden necesitar un

tratamiento especial. Para lotes de 300 o más el uso de diagramas tallo-hoja es generalmente incómodo.

Para el ejemplo, que tiene $n = 59$, resulta cantidad de lineas

$$L = [10 \times \log_{10} 59\ ] = [10 \times 1.77] = 17$$

Este valor coincide con la cantidad de líneas del esquema considerado, podría no coincidir exactamente.

Para determinar el intervalo de valores para cada línea dividimos R el rango del lote por L y redondeamos hacia arriba a la potencia de 10 más próxima.

En el ejemplo el rango $R = 64.42 - 62.85 = 1.57$ y $L=17$, de manera que $R / L = 0.09$. Redondeando a la potencia de 10 más próxima da 0.1 como ancho de los intervalos. Este es el valor utilizado.

Algunas variaciones

Ejemplo: Consideremos los datos (UREDA pág 13) de la dureza de 30 incrustaciones de aluminio presentadas en un estudio de control de calidad (Shewhart, 1931),

$L = [10 \times \log_{10} 30] = [14.77] = 14$, $R = 95.4 - 51.1 = 44.3$, y
$R / L = 44.3 / 14 = 3.16$.

Redondeando hacia arriba a la potencia de 10 más próxima, obtendríamos 10 como la longitud indicada para los intervalos.

Esta longitud es utilizada en el esquema tallo-hoja básico dado por la figura 4-a.

Figura 4 a

El punto decimal está 1 lugar a la derecha de los dos puntos (:)

```
11    11    5 : 11233345669
       5    6 : 34799
 14    8    7 : 00123488
  6    5    8 : 23467
  1    1    9 : 5
```

Obtenido con

```
> stem(dur,depth=T)
```

Figura 4 b

El punto decimal está 1 lugar a la derecha de los dos puntos (:)

```
 7    7    5 : 1123334
11    4    5 : 5669
13    2    6 : 34
      3    6 : 799
14    6    7 : 001234
 8    2    7 : 88
 6    3    8 : 234
 3    2    8 : 67
 1    0    9 :
 1    1    9 : 5
```

```
> stem(dur,5,depth=T)
```

Como el esquema de la figura 4-a tiene relativamente pocas líneas, utilizamos 2 líneas por tallo, o equivalentemente **5 dígitos** en cada línea, obteniendo el esquema de la figura 4-b.

Figura 5

```
5 : 07,11,24,30,34,35,41
5 : 53,57,57,95
6 : 35,43
6 : 73,91,95
7 : 02,05,14,23,30,44
7 : 78,85
8 : 25,27,43
8 : 58,75
9 :
9 : 54
```

Algunas veces interesa conservar un dígito adicional en cada hoja. En este diagrama aparecen dos dígitos en las hojas por cada observación.

```
> stem(dur,5,twodig=T)
```

Figura 6

El punto decimal está un lugar a la derecha de los dos puntos(:)

```
5 : 11
5 : 2333
```

Si el esquema está muy amontonado con dos líneas por tallo y muy raleado con una línea por tallo, en la

5 : 45 5 : 66 5 : 9 6 : 6 : 3 6 : 4 6 : 7 6 : 99 7 : 001 7 : 23 7 : 4 7 : 7 : 88 8 : 8 : 23 8 : 4 8 : 67 8 : 9 : 9 : 9 : 5	siguiente potencia de 10, en ese caso se lo puede construir con **5 líneas por tallo**: poniendo las hojas 0 y 1 en la 1ra línea 2 y 3 en la 2da línea 4 y 5 en la 3ra línea 6 y 7 en la 4ta línea 8 y 9 en la 5ta línea **> stem(dur,2,depth=T)** Equivalentemente **2 dígitos** por línea.

The decimal point is 1 digit(s) to the right of the | **5 | 1123344** **5 | 566** **6 | 044** **6 | 79** **7 | 0011234** **7 | 89** **8 | 334** **8 | 68** **9 |** **9 | 5**	En R > dur <- read.table(file.choose(), header=T) > stem(dur$dureza)

The decimal point is at the |

```
  50 | 71
  52 | 4045
  54 | 1377
  56 |
  58 | 5
  60 |
  62 | 5
  64 | 3
  66 | 3
  68 | 15
  70 | 254
  72 | 30
  74 | 4
  76 | 8
  78 | 5
  80 |
  82 | 57
  84 | 38
  86 | 5
  88 |
  90 |
  92 |
  94 | 4
```

```
> stem(dur$dureza,scale=2)
```

Observemos como han quedado ubicados los datos que a continuación se presentan ordenados

50.7 51.1 52.4 53.0 53.4
53.5 54.1 55.3 55.7 55.7
59.5 63.5 64.3 67.3 69.1
69.5 70.2 70.5 71.4 72.3
73.0 74.4 77.8 78.5 82.5
82.7 84.3 85.8 87.5 95.4

Resistencia del diagrama

Es inadecuado que la escala del diagrama de tallo-hoja se base en los valores mayores y menores de los datos.

Comenzamos excluyendo los datos inusuales y basamos la elección de la escala en el resto de los datos.

Veremos métodos para detectarlos.

Esos valores aparecen en las líneas con las etiquetas “bajo” y “alto” fuera del conjunto de tallos.

Si ahora los datos de la dureza del aluminio tuvieran un valor inusual:

Tabla 5

53.0	70.2	84.3	55.3	78.5	63.5	71.4	53.4
82.5	67.3	69.5	73.0	55.7	85.8	95.4	51.1
74.4	54.1	77.8	52.4	69.1	53.5	64.3	82.7
55.7	70.5	87.5	50.7	72.3	59.5	9.2	

obtendríamos el siguiente diagrama (sin las columnas de profundidad y cantidad de hojas):

Figura 7

Decimal point is 1 place to the right of the colon Low: 9.2 5 : 11233345669 6 : 34799 7 : 00123488 8 : 23467 9 : 5	El punto 9.2 es un valor detectado como “bajo” (low).

El diagrama de tallo-hoja ha demostrado ser una técnica versátil para el primer vistazo del analista de un lote de números.

Las tres maneras de factorizar 10 (1 x 10, 2 x 5 y 5 x 2) proveen un control adecuado sobre la escala (1, 2 ó 5 líneas por tallo) especialmente cuando es combinado con líneas de “bajos” y “altos” para valores inusuales.

Excluyendo posibles outliers es posible obtener un esquema más detallado y efectivo.