Estimación de densidades por núcleos

Hemos visto que la altura de un histograma en un punto cualquiera x puede escribirse como

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

donde

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

indica cada intervalo de clase.

La sumatoria en j se reduce al único sumando en para el cual $x \in B_j$, esto implica que j = j (x):

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

Para un conjunto de datos $x_1, \ldots, x_n$ tendremos

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

$$= (nh)^{-1} \sum_{i=1}^{n} I(h/2 \leq | x - \tilde{x}_i | < h/2)$$

siendo $\tilde{x}_i$ el centro del intervalo al cual pertenece x_i

Luego

$$\hat{f}_h(x) = (nh)^{-1} \sum_{i=1}^{n} I(x - \tilde{x}_i, h)$$

donde $I(z,h)$ es la función indicadora del intervalo [-h/2,h/2)

Como estimador de una densidad el histograma ha sido criticado por:

- Se desperdicia información al reemplazar cada dato por el centro del intervalo de clase en el cual cae.
- En la mayoría de las situaciones la función de densidad es suave mientras que el histograma no lo es
- El comportamiento del estimador es dependiente de la longitud utilizada para intervalos (o equivalentemente cajas)

Rosemblat (1956), Whittle (1958) y Parzen (1962) desarrollaron un enfoque que eliminan las primeras dos dificultades. Utilizan funciones de núcleos suaves en vez de una caja y estas funciones están centradas directamente sobre cada observación. La estimación por núcleos (**kernel estimation**) tiene la forma:

$$\hat{f}_h(x) = (nh)^{-1} \sum_{i=1}^{n} w(\frac{x - x_i}{h})$$

siendo w(x), el núcleo, una función no negativa, simétrica y centrada en cero, cuya integral es 1.

Por ejemplo w(x) puede ser la densidad normal estándar.

El parámetro h (ancho de la ventana) controla su suavidad y corresponde a la longitud del intervalo de clase en un histograma. Si h es demasiado pequeña la estimación resulta demasiado rugosa. Si h es demasiado grande la estimación resulta demasiado desparramada.

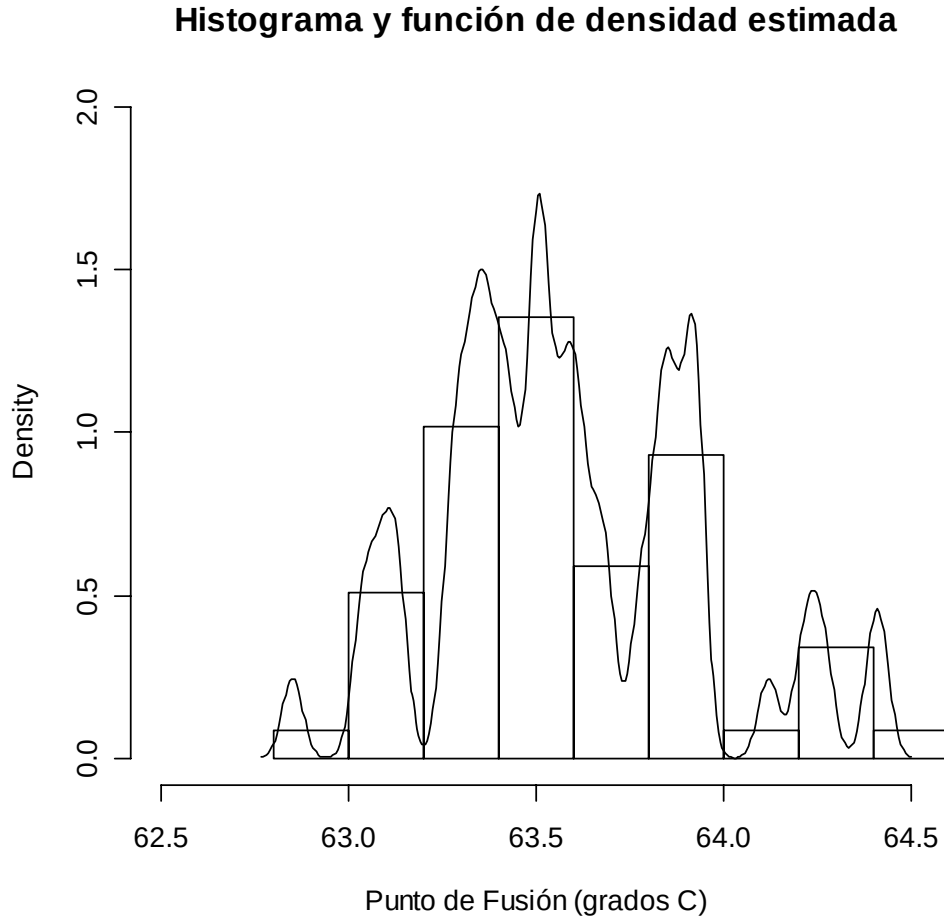


Figura 3. Histograma junto con una función de densidad estimada

La figura 3 muestra el histograma junto con una función de densidad estimada que, aunque es más suave que el histograma, es demasiado “rugosa”. Fue obtenida mediante las siguientes instrucciones:

```
hist(Cera $CERA,probability=T, ylim=c(0,2),
  xlim=c(62.5,64.5),main="",
  xlab="Punto de Fusión (grados C)")
  lines(density(Cera $CERA,n=300,width=0.11))
   title ( "Histograma y función de densidad
estimada")
```

Los gráficos de la figura 4 fueron obtenidos mediante las siguientes instrucciones:

```
 par(mfrow=c(3,1))
 plot(density(Cera
$CERA,n=500,width=0.11,from=62,to=66))

plot(density(Cera$CERA,n=500,width=0.30,from=
62, to=66))

plot(density(Cera$CERA,n=500,width=0.60,from=
62, to=66))
```

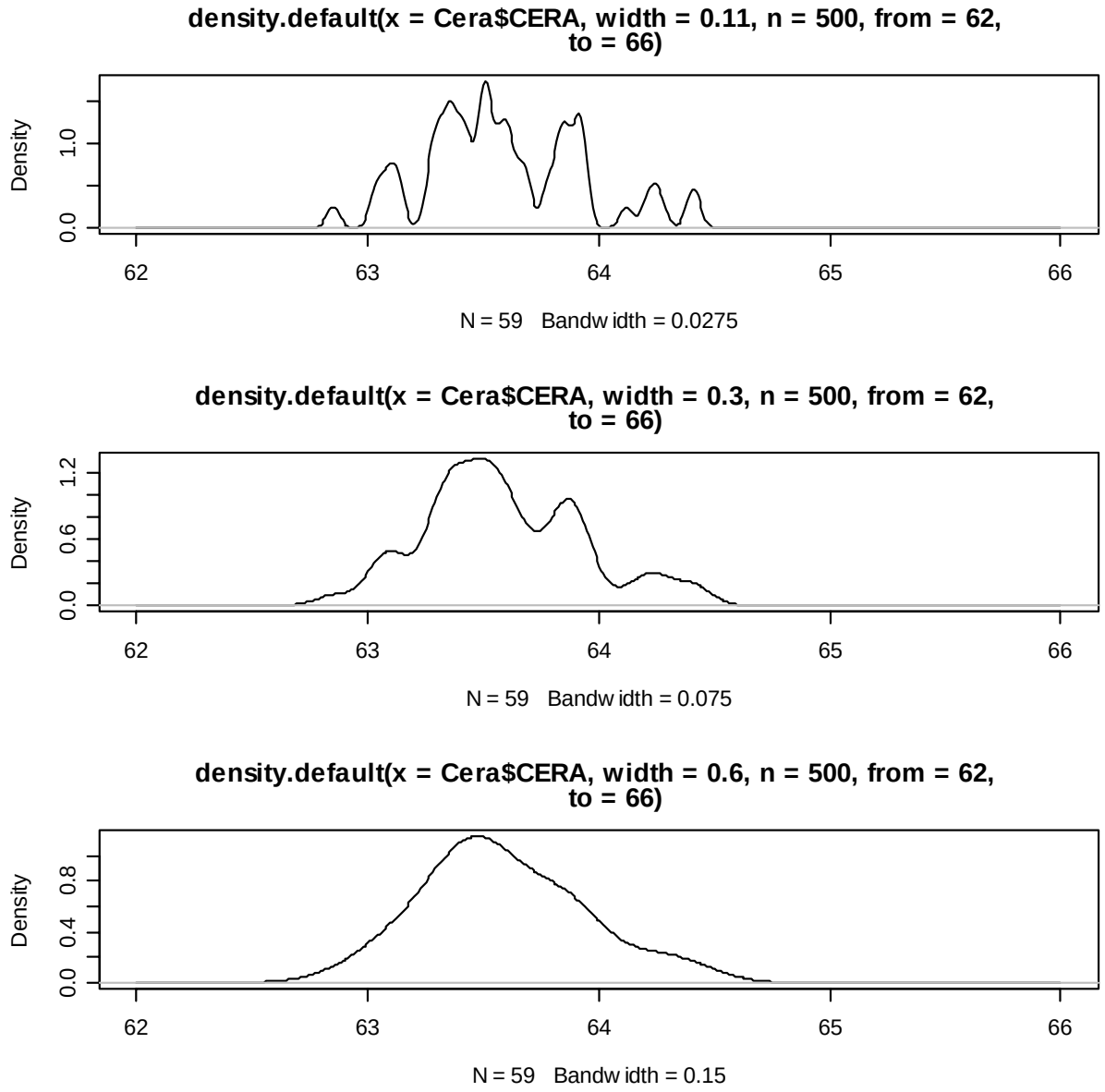


Figura 4. Estimaciones de densidad de probabilidad para los datos del punto de fusión de ceras naturales, para distintos tamaños del ancho de la ventana.

Sesgo y Varianza de los estimadores de densidad por núcleos

$$Sesgo(\hat{f}_h(x)) = \frac{h^2}{2}\sigma_w^2 \, f''(x) + o(h^2) \qquad h \to 0$$

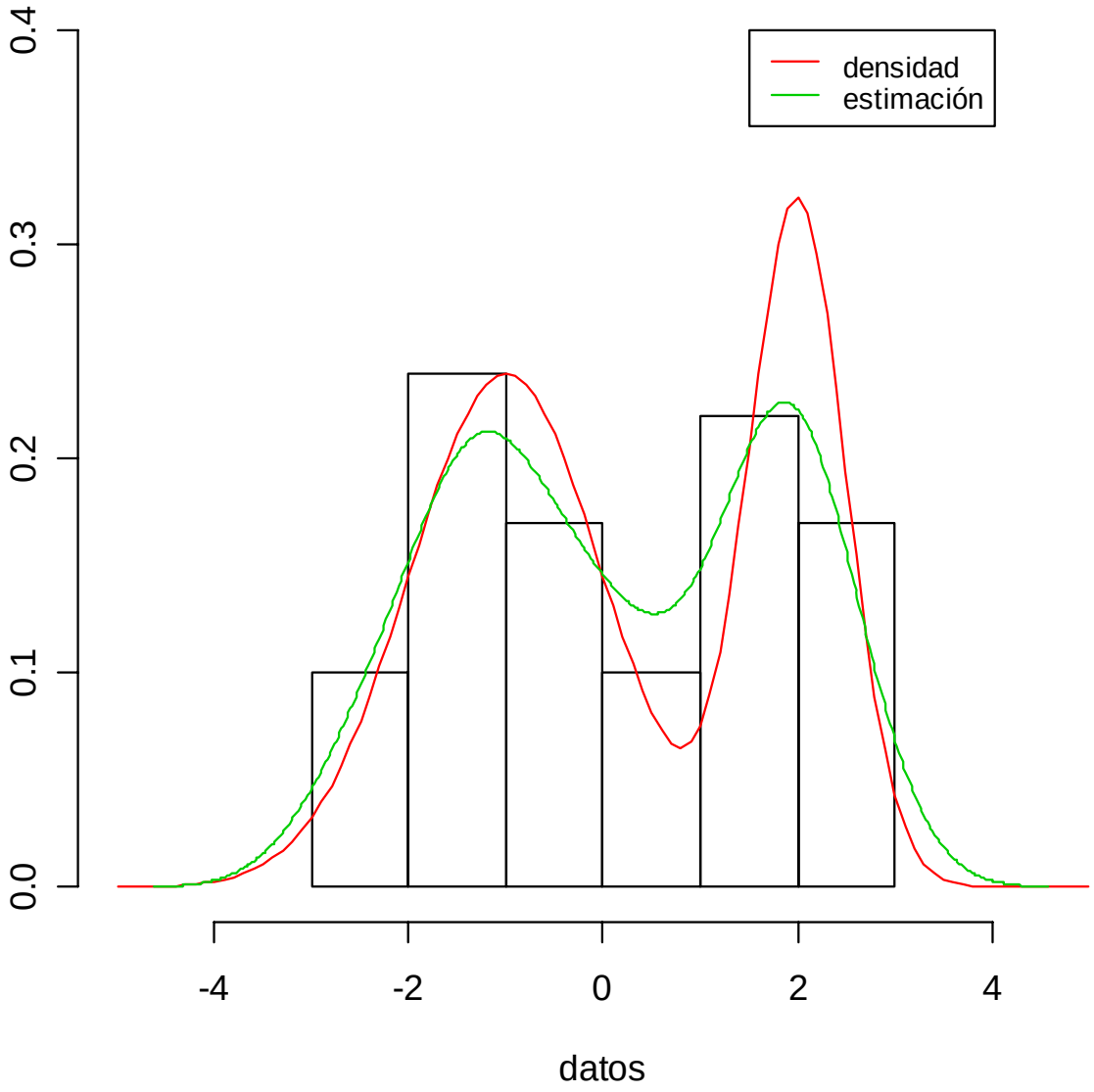
donde $\sigma_w^2 = \int x^2 w(x)dx$.

- El sesgo es cuadrático en *h*, debe elegirse pequeño para reducir el sesgo.

- El sesgo es proporcional a *f”*, la estimación por núcleos subestima la densidad (el sesgo es negativo) en los picos de la densidad verdadera *f* y la subestima en los valles, como vemos en el ejemplo siguiente.

Ejemplo:

```
#el estimador por núcleos subestima en los picos y
sobreestima en los valles
fn1 <- rnorm(60,-1,1)
fn2 <- rnorm(40,2,0.5)

datos <- c(fn1,fn2)

densidad <- 0.6*dnorm(seq(-5,5,0.1),-1,1)+
0.4*dnorm(seq(-5,5,0.1),2,0.5)

par(mfrow=c(1,1))

hist(datos, probability=T, xlim=c(-5,5),
ylim=c(0,0.4), main="", ylab="")

lines(seq(-5,5,0.1),densidad,col=2)
lines(density(datos,n=500),col=3)

legend(1.5,.4, legend = c("densidad","estimación"),
col = 2:3,lty = 1, cex = .8, y.intersp = 1)
```



$$Var(\hat{f}_h(x)) = \frac{1}{nh} f(x)\alpha(w) + o((nh)^{-1}) \qquad nh \to \infty$$

donde $\alpha(w) = \int w^2(x)dx$.

- La varianza es inversamente proporcional al tamaño de la muestra como ocurre siempre. El término *nh* puede pensarse como controlando el tamaño muestral local.

- En forma análoga a lo que ocurre con el histograma, la varianza es proporcional a la altura de la densidad.

Igual que con los histogramas, tenemos que $h \to 0$ para reducir el sesgo, $\frac{1}{nh} \to 0$ para reducir la varianza y ambos dependen de la función de densidad que queremos estimar.

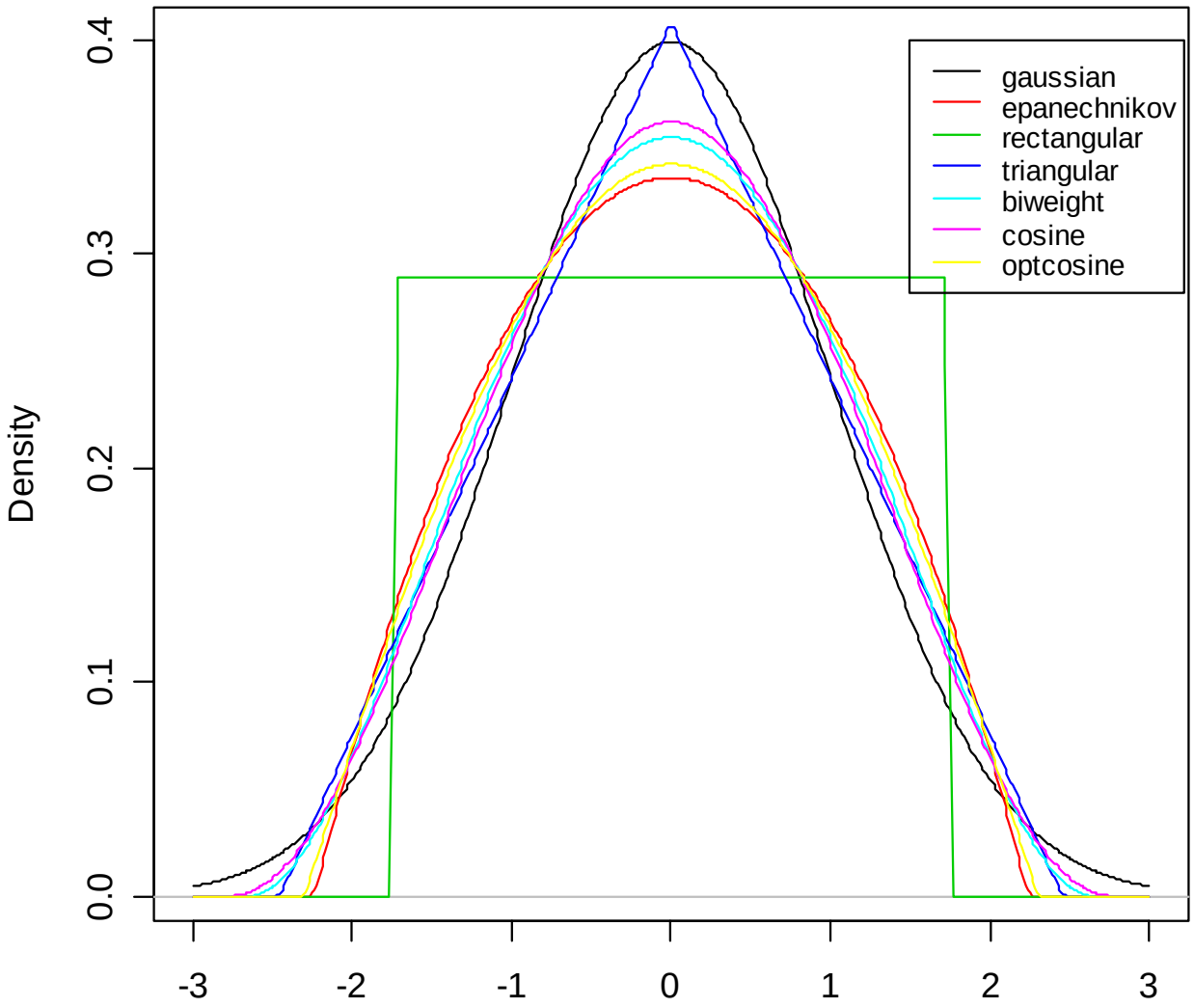
Tipos de núcleos incorporados en R

con `> density.default` se obtienen detalladamente las expresiones de los núcleos incorporados

```
kords <- switch(kernel, gaussian = dnorm(kords, sd = bw),
        rectangular = {
            a <- bw * sqrt(3)
            ifelse(abs(kords) < a, 0.5/a, 0)
        }, triangular = {
            a <- bw * sqrt(6)
            ax <- abs(kords)
            ifelse(ax < a, (1 - ax/a)/a, 0)
        }, epanechnikov = {
            a <- bw * sqrt(5)
            ax <- abs(kords)
            ifelse(ax < a, 3/4 * (1 - (ax/a)^2)/a, 0)
        }, biweight = {
            a <- bw * sqrt(7)
```

```
    ax <- abs(kords)
    ifelse(ax < a, 15/16 * (1 - (ax/a)^2)^2/a, 0)
}, cosine = {
    a <- bw/sqrt(1/3 - 2/pi^2)
    ifelse(abs(kords) < a, (1 + cos(pi * kords/a))/(2 *
        a), 0)
}, optcosine = {
    a <- bw/sqrt(1 - 8/pi^2)
    ifelse(abs(kords) < a, pi/4 * cos(pi * kords/(2 *
        a))/a, 0)
})
```

Núcleos de la función density() en R, con bw = 1



```
 kernels <- eval(formals(density.default)$kernel)
plot (density(0, bw = 1), xlab = "",
  main=" Núcleos de la función density() en R, con bw = 1")
for(i in 2:length(kernels))
   lines(density(0, bw = 1, kernel =  kernels[i]), col = i)
legend(1.5,.4, legend = kernels, col = seq(kernels),
       lty = 1, cex = .8, y.intersp = 1)
```

del help tenemos:

bw the smoothing bandwidth to be used. The kernels are scaled such that this is the standard deviation of the smoothing kernel. (Note this differs from the reference books cited below, and from S-PLUS.)

Núcleo cuadrático de Epanechnikov:

$$w(x) = \begin{cases} \frac{3}{4}(1-x^2) & \text{si } -1 \le x \le 1 \\ 0 & \text{fuera} \end{cases},$$

minimiza el error cuadrático medio integrado asintótico entre todos los núcleos no negativos con soporte compacto.

Función bicuadrada, introducida por Tukey en el contexto de estimación robusta:

$$w(x) = \begin{cases} (1-x^2)^2 & \text{si } -1 \le x \le 1 \\ 0 & \text{fuera} \end{cases}$$

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.

En R	En S-plus
> stem(Cera$CERA)	> stem (Cera$CERA)
	N = 59 Median = 63.53 Quartiles = 63.36, 63.84
The decimal point is 1 digit(s) to the left of the |	Decimal point is 1 place to the left of the colon
628 | 5 630 | 358033 632 | 77001446669 634 | 013350000113668 636 | 001368988 638 | 33466822223 640 | 2 642 | 147 644 | 02	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

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 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	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 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.

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

La primera columna (PROF.) de profun-didad, 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 colum-na (# hojas) da la

5	3	642 : 147	cantidad de hojas en cada tallo.
2	0	643 :	
2	2	644 : 02	

En S-plus

```
> 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

La regla de Dixon y Kronmal(1965) para la cantidad de intervalos

$$L = [10 \text{ x } \log_{10} n\ ]$$

se traduce en el caso de los diagramas tallo-hoja en cantidad de líneas o tallos del diagrama. Es razonable 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

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

Este valor coincide con la cantidad de líneas del diagrama presentado por S-plus por defecto. Para obtener la misma cantidad de líneas R:

```
> stem (Cera$CERA,scale=2)
```

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

```
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
```

El parámetro “scale” controla la longitud del diagrama tallo-hoja en R.

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 por el S-plus

Algunas variaciones

Ejemplo: Consideremos los datos (UREDA pág 13)

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

de la dureza de 30 incrustaciones de aluminio presentadas en un estudio de control de calidad (Shewhart, 1931),

L = [10 x $\log_{10}30$] = [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, obtendremos la longitud presentada por S-plus (por defecto) para los intervalos y si redondeamos a 5 tendremos la del R

R	S-plus
> stem(dur$dureza) The decimal point is 1 digit(s) to the right of the | 5 | 1123344 5 | 566 6 | 044 6 | 79 7 | 0011234 7 | 89 8 | 334 8 | 68 9 | 9 | 5	> stem(dur$dureza) N = 30 Median = 69.3 Quartiles = 55.3, 77.8 Decimal point is 1 place to the right of the colon 5 : 11233345669 6 : 34799 7 : 00123488 8 : 23467 9 : 5

En R controlamos la cantidad de tallos (o filas) del diagrama con “scale”

```
> stem(dur$dureza,scale=0.5)

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

  5 | 1123344566
  6 | 04479
  7 | 001123489
  8 | 33468
  9 | 5
```

En S.plus el segundo parámetro “nl” de stem, indica la cantidad de dígitos diferentes en cada que puede tomar cada hoja en un tallo y por lo tanto controla la cantidad de tallos.

Figura 4

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 5

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

Figura 6

El punto decimal está un lugar a la derecha de los dos puntos(:) 5 : 11 5 : 2333 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	Si el esquema está muy amontonado con dos líneas por tallo y muy raleado con una línea por tallo, en la 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.

En R

The decimal point is at the |	> stem(dur$dureza,scale=2)
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**	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

Puede resultar inadecuado que la escala del diagrama de tallo-hoja se base en los valores mayores y menores de los datos. Veremos como detectar datos inusuales de manera de excluirlos y basar la elección de la escala en el resto de los datos.

```
> duroloco<- c(dur$dureza,9.2)
> stem(duroloco)

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

  0 | 9
  1 |
  2 |
  3 |
  4 |
  5 | 1123344566
  6 | 04479
  7 | 001123489
  8 | 33468
  9 | 5
```

```
> duroloco2 <- c(dur$dureza,920)
> stem(duroloco2)

  The decimal point is 2 digit(s) to the right of the
|

  0 | 55555556666667777777778888899
  1 | 0
  2 |
  3 |
  4 |
  5 |
  6 |
  7 |
  8 |
  9 | 2
```