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) y j ∈ 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:

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

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.

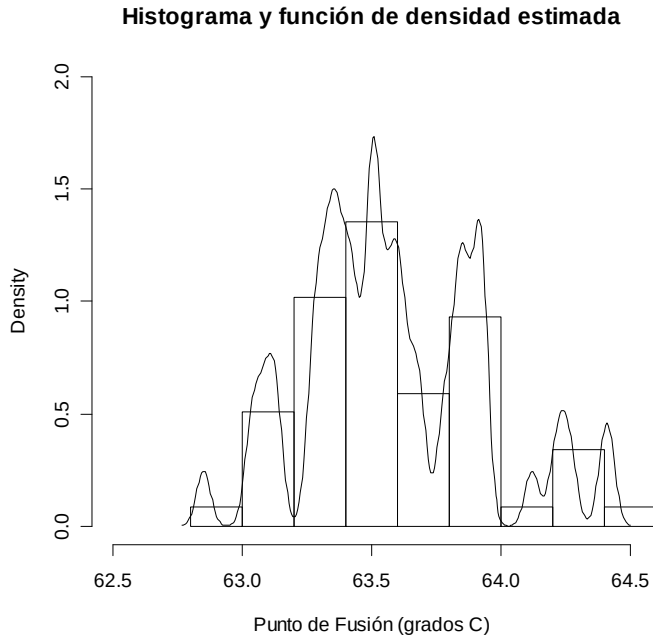


Figura 3. Histograma junto con una función de densidad estimada

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))
```

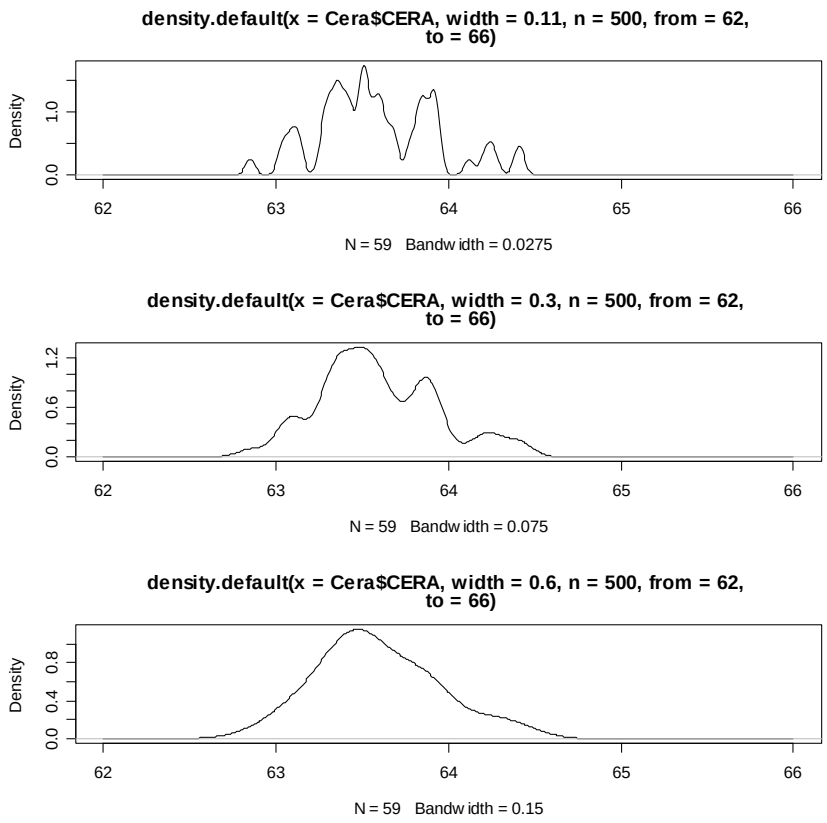


Figura 4. Estimaciones de densidad de probabilidad para los datos del punto de fusión de ceras naturales, para distintos tamaños del ancho de la ventana.

Sesgo y Varianza de los estimadores de densidad por núcleos

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

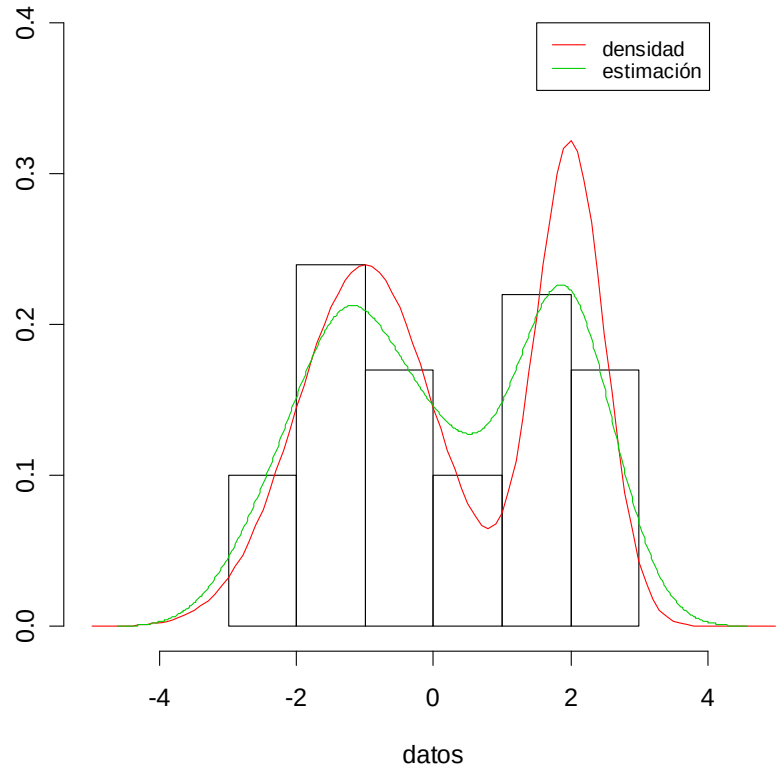
donde $\sigma_w^2 = \int x^2 w(x)dx$.

- 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.

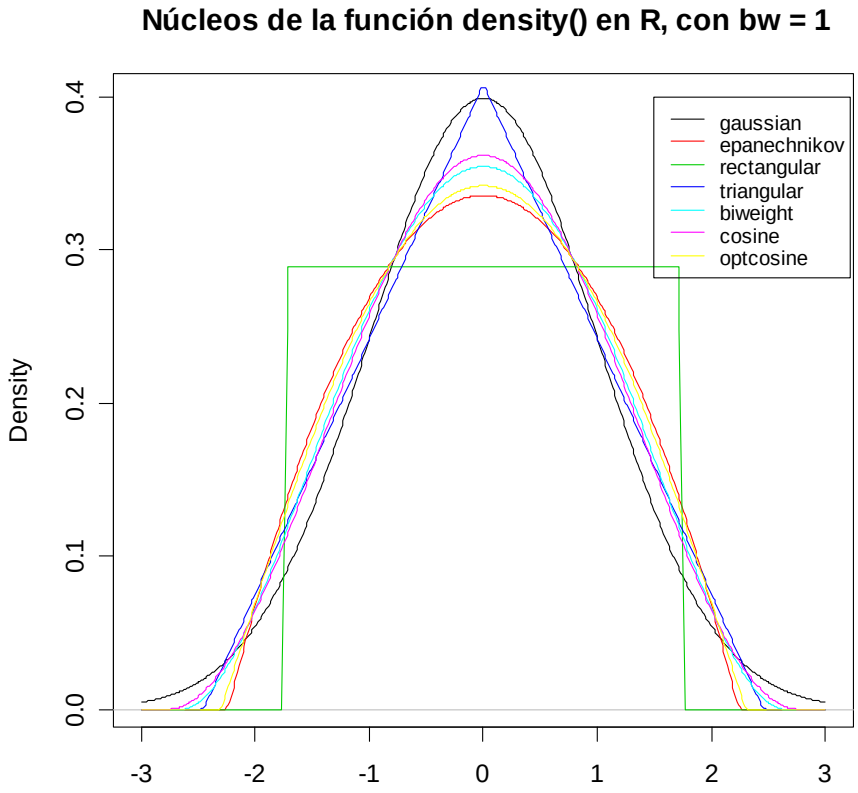
Igual que con los histogramas, tenemos que $h \to 0$ para reducir el sesgo, $\frac{1}{nh} \to 0$ para reducir la varianza y ambos dependen de la función de densidad que queremos estimar.

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)
        })
```

Obtenemos la figura anterior con las siguientes instrucciones:

```
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)
```

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.

> stem(Cera$CERA) The decimal point is 1 digit(s) to the left of the | 628 | 5	> stem(Cera$CERA,scale=2) The decimal point is 1 digit(s) to the left of the | 628 | 5

630 | 358033	629 |
632 | 77001446669	630 | 358
634 | 013350000113668	631 | 033
636 | 00136898	632 | 77
638 | 3346682223	633 | 001446669
640 | 2	634 | 01335
642 | 147	635 | 0000113668
644 | 02	636 | 0013689
	637 | 88
	638 | 334668
	639 | 22223
	640 |
	641 | 2
	642 | 147
	643 |
	644 | 02

En su apariencia global el diagrama se asemeja a un histograma (con ancho de intervalo igual a 0.2 ^{0}C o 0.1 ^{0}C respectivamente)

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

Profundidad

A cada dato se le puede asignar un *rango*, contando desde cada extremo en el lote ordenado. La *profundidad* es el menor de los dos valores.

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.

Figura 3

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

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

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

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 líneas

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

Ya vimos que para obtener la misma cantidad de líneas en R es necesario fijar el argumento scale = 2 en la función stem:

```
> 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.

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

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

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

Redondeando hacia arriba a 5 tendremos el ancho de intervalo que se obtiene por defecto con R

```
> 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
```

Controlamos la cantidad de tallos (o filas) del diagrama con “scale =0.5” y obtenemos intervalos de clase de longitud =10:

```
> 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
```

Controlamos la cantidad de tallos (o filas) del diagrama ahora con “scale =2” y obtenemos intervalos de clase de longitud =2:

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

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
```