Notas del curso CA-403

I Ciclo 2020

Instrucciones de uso

Este es un archivo de latex normal, salvo que se le pueden agregar códigos de R.

Para hacerlo solo debe encerrar su código con lo siguiente comandos.

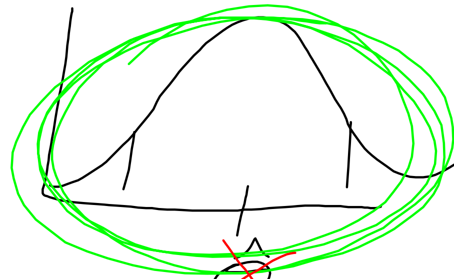
```
<<>>=
   Su código acá.
@
```

Algunas recomendaciones inciales:

- Tratemos de ser ordenados con el texto y el código. Recuerden que esto será usado por ustedes en el exámen.
- No usen comandos proprios (`newcommand`) ya que eso solo haría más díficil que los compañeros puedan editar su trabajo.
- El documento es colaborativo, por lo que está bien editar o escribir "encima" de otro compañero, siempre y cuando esto sea para mejorar el texto.

Creo en el buen juicio de cada uno para hacer estas notas lo mejor posible.

$\cancel{N(\theta, \sigma^2)}$ $\cancel{\hat{\theta}}, \cancel{\hat{\sigma}^2}$



$\cancel{\hat{\theta}}$ $\cancel{\hat{\sigma}^2}$

1 Histograma

El histograma es una de las estructuras básicas en estadística. Básicamente con este objeto se puede visualizar la distribución de los datos sin tener conocimiento previo de los mismos.

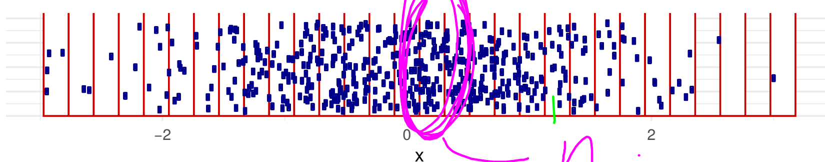
1.1 Construcción Estadística

Suponga que $X_1, X_2, \ldots, X_n$ proviene de una distribución desconocida.

- Seleccione un origen x_0 y divida la línea real en *segmentos*.

$$B_j = [x_0 + (j-1)h, x_0 + jh) \quad j \in \mathbb{Z}$$

- Cuente cuántas observaciones caen en cada segmento. n_j.



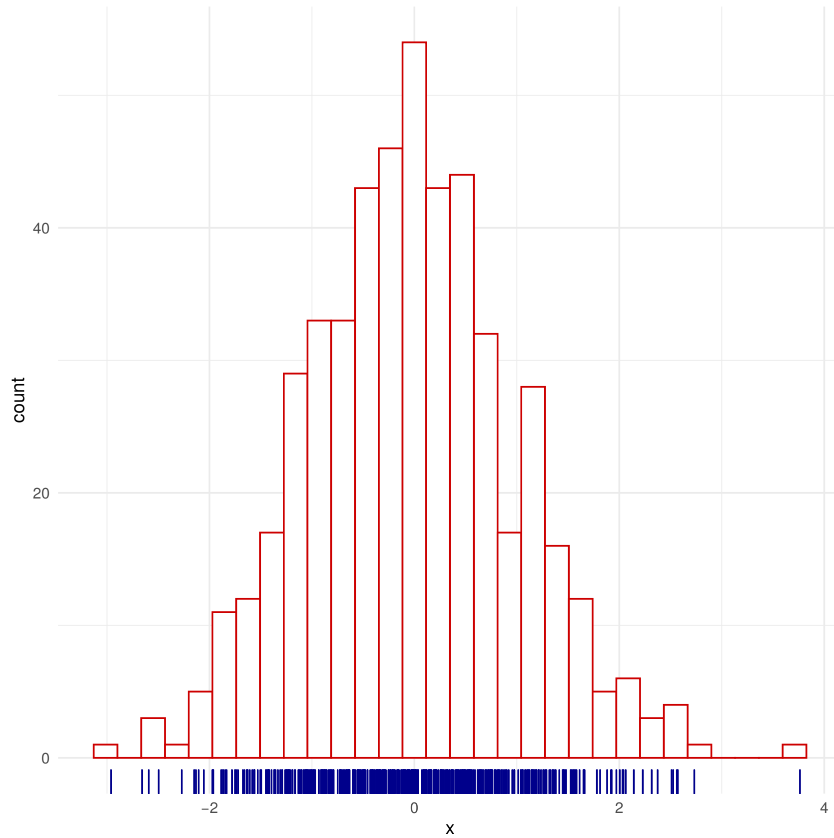
n_j

- Cuente la frecuencia por el tamaño de muestra n y el ancho de banda h.

$$f_j = \frac{n_j}{nh}$$

$\to \sum f_j = 1$

- Dibuje el histograma.



Formalmente el histograma es el

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

donde I es la indicadora.

B_j

1.2 Construcción probabilistica

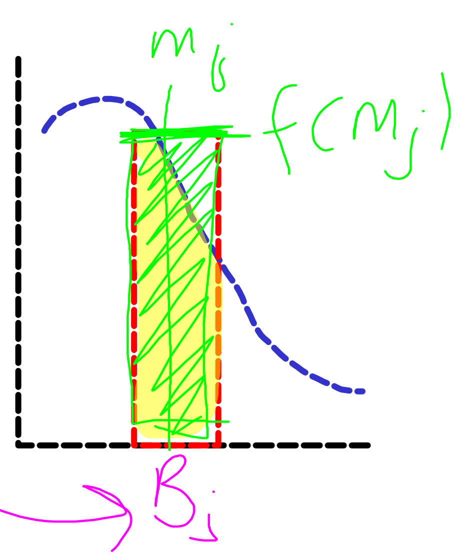
Denote $m_j = jh - h/2$ el centro del segmento,

Desconocida

$[\ m_j\ )$

h

$$\mathbb{P}\left(X \in \left[m_j - \frac{h}{2}, m_j + \frac{h}{2}\right)\right) = \int_{m_j - \frac{h}{2}}^{m_j + \frac{h}{2}} f(u)du$$
$$\approx f(m_j)h$$



Esto se puede aproximar como

n_j

$$\mathbb{P}\left(X \in \left[m_j - \frac{h}{2}, m_j + \frac{h}{2}\right)\right) \approx \frac{1}{n}\#\left\{X \in \left[m_j - \frac{h}{2}, m_j + \frac{h}{2}\right)\right\}$$

Acomodando un poco la expresión

$$\hat{f}_h(m_j) = \frac{1}{nh}\#\left\{X \in \left[m_j - \frac{h}{2}, m_j + \frac{h}{2}\right)\right\}$$

1.3 Propiedades estadísticas

Suponga que $x_0 = 0$ y que $x \in B_j$ fijo, entonces

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

$Bias(\hat{f}_h) = \mathbb{E}[\hat{f}_h] - f(x)$

1.3.1 Sesgo

El cálculo del sesgo es el

$$\begin{aligned}\mathbb{E}\left[\hat{f}_h(m_j)\right] &= \frac{1}{nh}\sum_{i=1}^{n}\mathbb{E}\left[I(X_i \in B_j)\right] \\ &= \frac{1}{\cancel{n}h}\cancel{n}\mathbb{E}\left[I(X_i \in B_j)\right]\end{aligned}$$

1 ó 0

$I(X_i \in B_j)$ es una indicadora con probabilidad de 1 de $\boxed{\int_{(j-1)h}^{jh} f(u)du}$ y 0 sino.

Entonces

$$\mathbb{E}\left[I(X_i \in B_j)\right] = \mathbb{P}\left(I(X_i \in B_j) = 1\right) = \int_{(j-1)h}^{jh} f(u)du.$$

Entonces,

$$\mathbb{E}\left[\hat{f}_h(m_j)\right] = \frac{1}{h}\int_{(j-1)h}^{jh} f(u)du$$

$$Bias(\hat{f}_h(m_j)) = \frac{1}{h}\int_{(j-1)h}^{jh} f(u)du - f(x)$$

Esto se puede aproximar usando Taylor alrededor del centro $m_j = jh - h/2$ de B_j de modo que $f(u) - f(x) \approx f'(m_j)(u - x)$.

$$Bias(\hat{f}_h(m_j)) = \frac{1}{h}\int_{(j-1)h}^{jh} f(u) - f(x)du \approx f'(m_j)(m_j - x)$$

1.3.2 Varianza

Dado que todos los X_i son i.i.d., entonces

$$\begin{aligned}\mathrm{Var}\left(\hat{f}_h(m_j)\right) &= \mathrm{Var}\left(\frac{1}{nh}\sum_{i=1}^{n} I(X_i \in B_j)\right) \\ &= \frac{1}{n^2h^2} n\mathrm{Var}\left(I(X_i \in B_j)\right)\end{aligned}$$

Var (Ber)
= p (1 - p)

La variable I es una bernoulli con parametro $\int_{(j-1)h}^{jh} f(u)du$ por lo tanto su varianza es el

$$\mathrm{Var}\left(\hat{f}_h(x)\right) = \frac{1}{nh^2}\left(\int_{(j-1)h}^{jh} f(u)du\right)\left(1 - \int_{(j-1)h}^{jh} f(u)du\right)$$

Tarea 1. *Usando un desarrollo de Taylor como en la parte anterior, pruebe que*

$$\mathrm{Var}\left(\hat{f}_h(x)\right) \approx \frac{1}{nh} f(x)$$

2 Error cuadrático medio

El error cuadrático medio del histograma es el

$$\mathrm{MSE}\left(\hat{f}_h(x)\right) = \mathrm{E}\left[\left(\hat{f}_h(x) - f(x)\right)^2\right] = \mathrm{Bias}^2\left(\hat{f}_h(x)\right) + \mathrm{Var}\left(\hat{f}_h(x)\right).$$

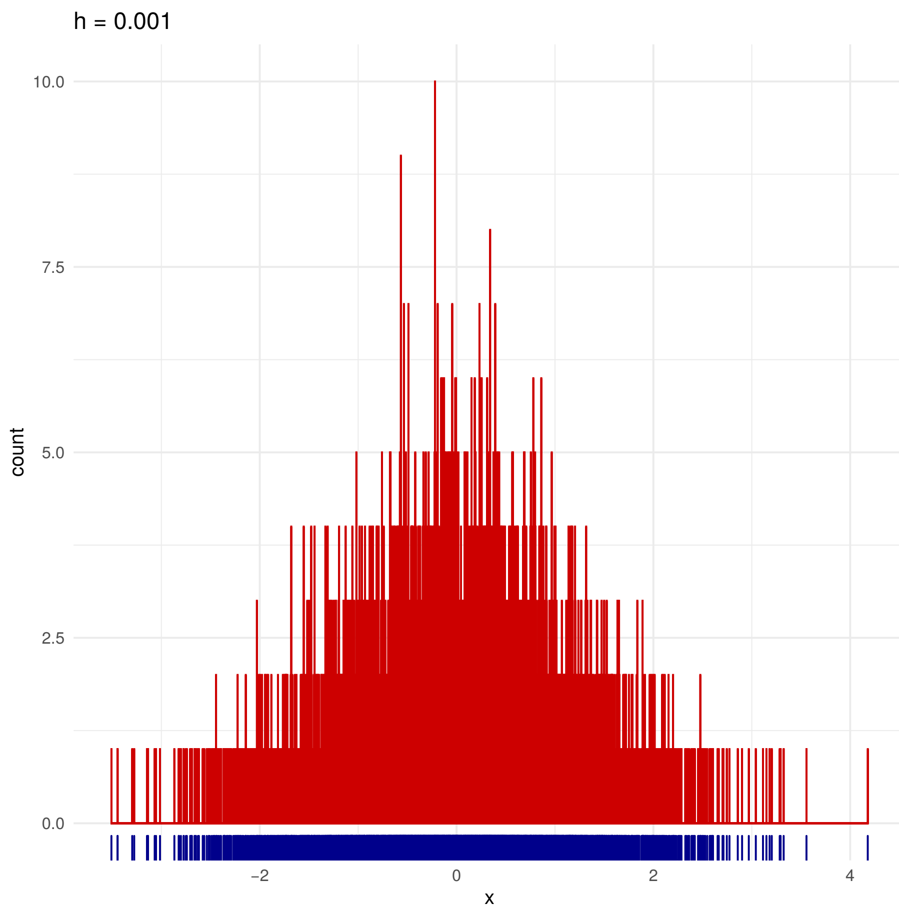
Tarea 2. *Pueden probar la segunda igualdad de la expresión anterior?*

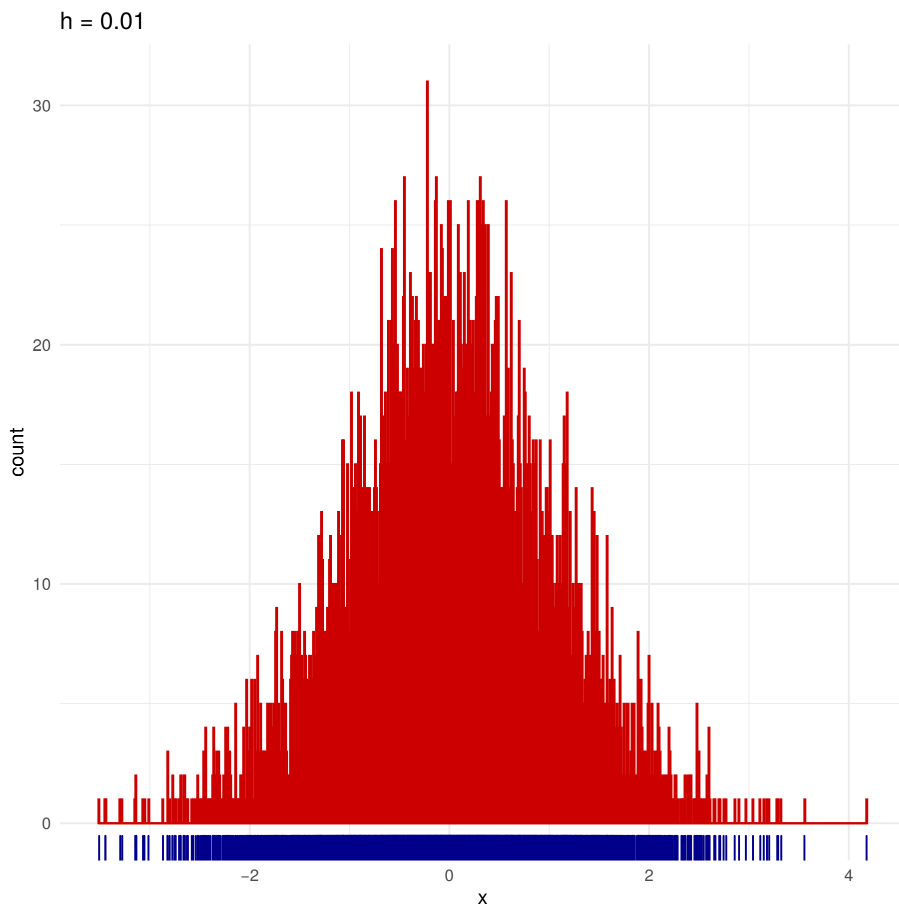
Retomando los términos anteriores se tiene que

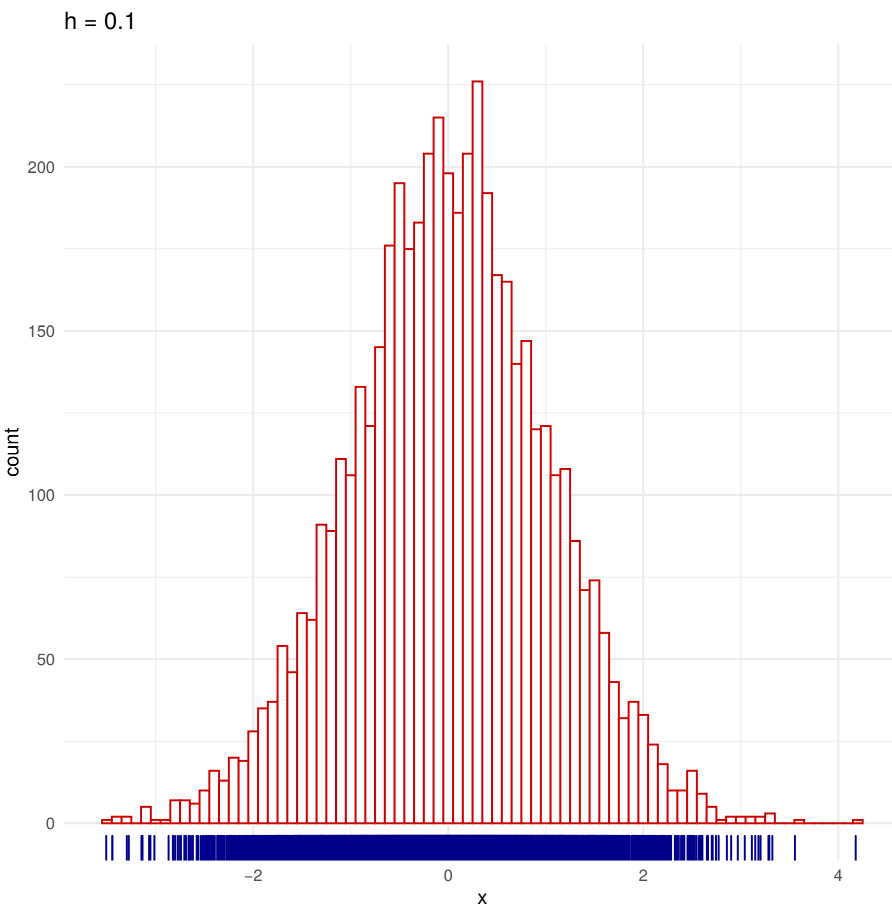
$$\begin{aligned}\mathrm{MSE}\left(\hat{f}_h(x)\right) = \frac{1}{nh} f(x) + f'\left\{(j-1/2)\,h\right\}^2\left\{(j-1/2)\,h - x\right\}^2 \\ + o\,(h) + o\left(\frac{1}{nh}\right)\end{aligned}$$

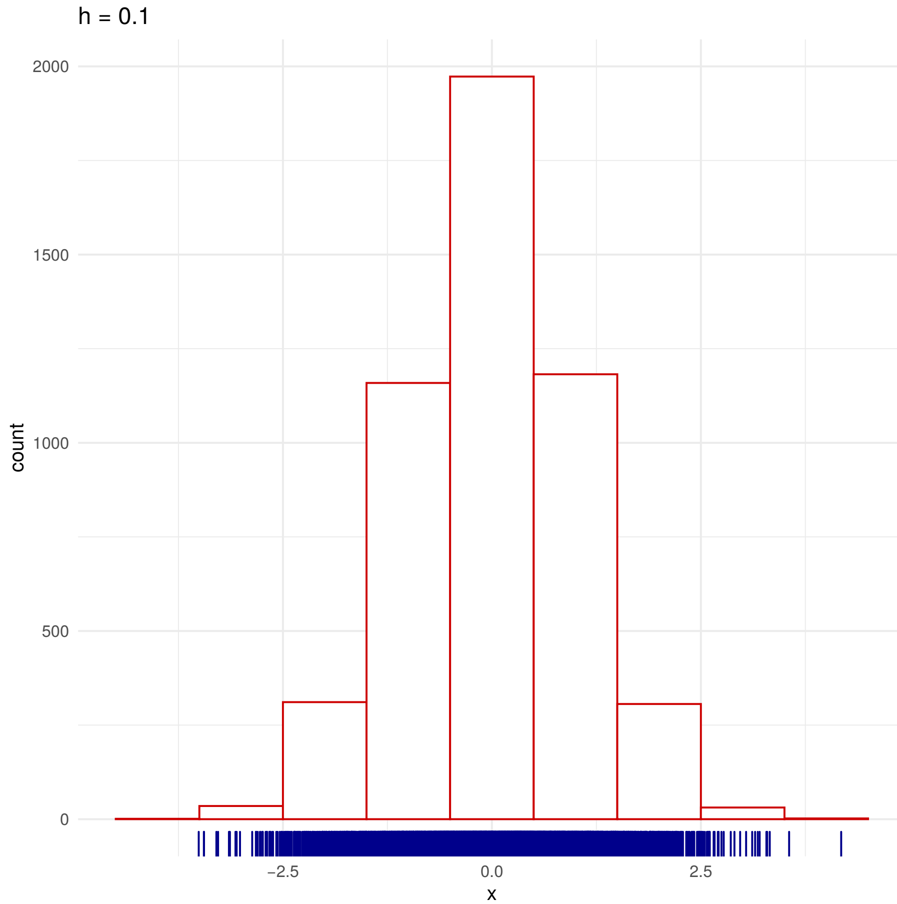
La fórmula anterior tiene la siguiente particularidad

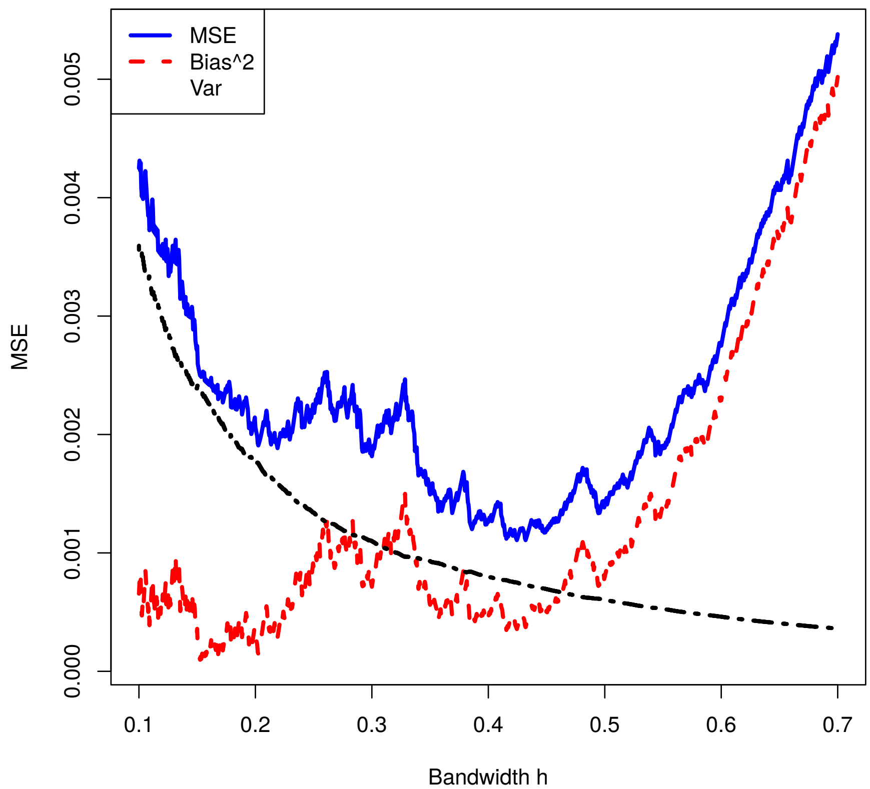
- Si $h \to 0$, la varianza crece (converge a ∞) y el sesgo decrece (converge a $f'(0)x^2$).
- Si $h \to \infty$, la varianza decrece (hacia 0) y el sesgo crece (hacia ∞)

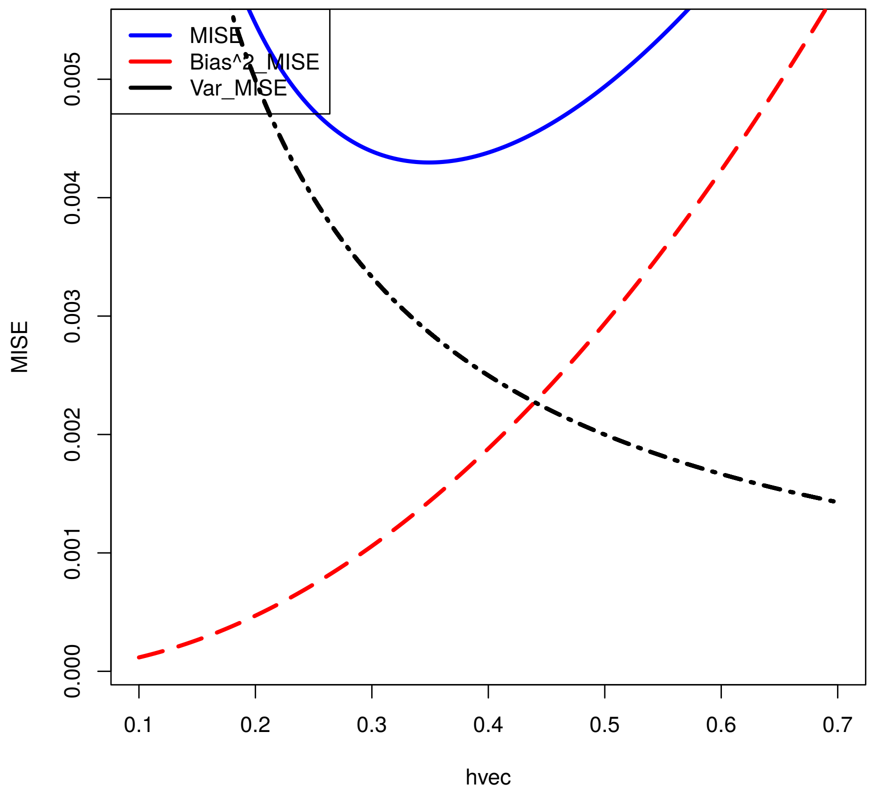

h = 0.001
count
10.0
7.5
5.0
2.5
0.0
−2
0
2
4
x

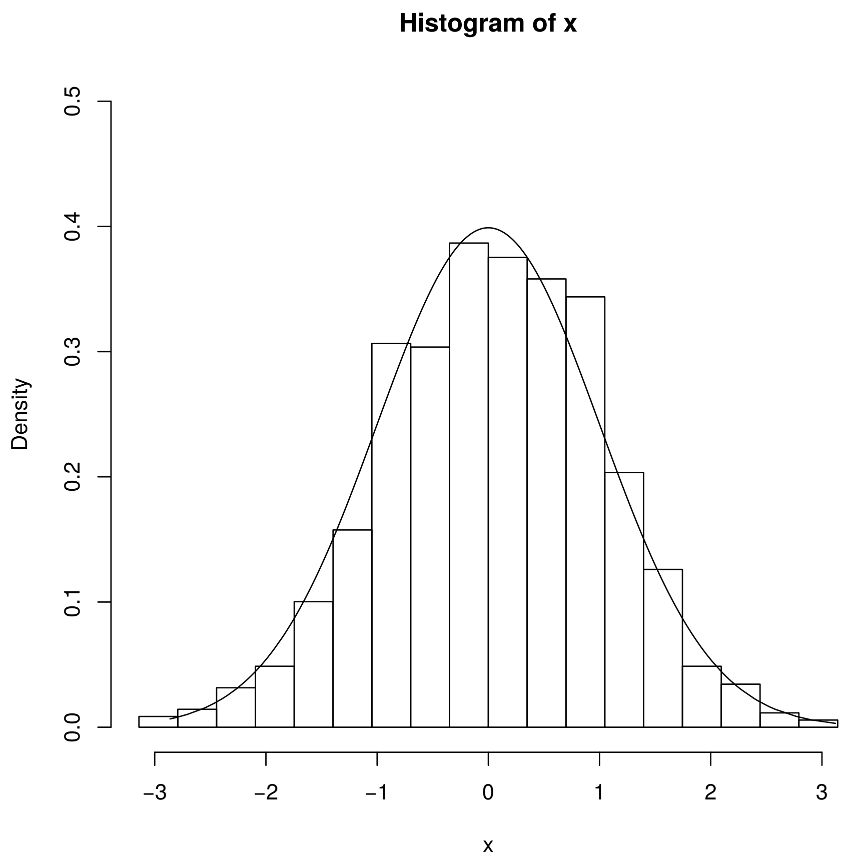

h = 0.01
count
30
20
10
0
−2
0
2
4
x


h = 0.1
count
200
150
100
50
0
−2
0
2
4
x


h = 0.1
2000
1500
1000
500
0
count
−2.5
0.0
2.5
x


MSE
Bias^2
Var
0.000
0.001
0.002
0.003
0.004
0.005
MSE
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Bandwidth h


MISE
Bias^2_MISE
Var_MISE
0.000
0.001
0.002
0.003
0.004
0.005
MISE
0.1
0.2
0.3
0.4
0.5
0.6
0.7
hvec


Histogram of x
Density
0.0
0.1
0.2
0.3
0.4
0.5
−3
−2
−1
0
1
2
3
x