Quick Recap

- **Chapters 1–3: The basics**
  - Introduction to R
  - Basic R commands
  - Numerical summaries

Quick Recap

- **Chapters 1–3: The basics**
  - Introduction to R
  - Basic R commands
  - Numerical summaries
- **Chapters 4–5: More advanced R stuff**
  - Graphics
  - Functions
  - Loops
  - Control statements

Quick Recap

- **Chapters 1–3: The basics**
  - Introduction to R
  - Basic R commands
  - Numerical summaries
- **Chapters 4–5: More advanced R stuff**
  - Graphics
  - Functions
  - Loops
  - Control statements
- **Chapters 6–7: Randomness**
  - Random number generation
  - Simulating random observations from discrete probability distributions

Quick Recap – Chapter 8: Monte Carlo methods

Basic idea: Repeatedly generate random numbers/samples to approximate a quantity of interest

Quick Recap – Chapter 8: Monte Carlo methods

Basic idea: Repeatedly generate random numbers/samples to approximate a quantity of interest

- Monte Carlo integration (“numerical integration”)

Quick Recap – Chapter 8: Monte Carlo methods

Basic idea: Repeatedly generate random numbers/samples to approximate a quantity of interest

- Monte Carlo integration ("numerical integration")
- Simulation studies:

Quick Recap – Chapter 8: Monte Carlo methods

Basic idea: Repeatedly generate random numbers/samples to approximate a quantity of interest

- Monte Carlo integration ("numerical integration")
- Simulation studies:
  - Monoploy

Quick Recap – Chapter 8: Monte Carlo methods

Basic idea: Repeatedly generate random numbers/samples to approximate a quantity of interest

- Monte Carlo integration ("numerical integration")
- Simulation studies:
  - Monoploy
  - Replicating rolls on a die: the binomial distribution

Quick Recap – Chapter 8: Monte Carlo methods

Basic idea: Repeatedly generate random numbers/samples to approximate a quantity of interest

- Monte Carlo integration ("numerical integration")
- Simulation studies:
  - Monoploy
  - Replicating rolls on a die: the binomial distribution
  - The distribution of the sample mean: CLT

Quick Recap – Chapter 8: Monte Carlo integration

We generate a random data point from a *simulation grid*. Let

$$A = \{\text{The data point lies below the curve}\}.$$

Quick Recap – Chapter 8: Monte Carlo integration

We generate a random data point from a *simulation grid*. Let

$$A = \{\text{The data point lies below the curve}\}.$$

Then

$$\Pr[A] = \frac{\text{Area under curve}}{\text{Area of simulation grid}}$$

Quick Recap – Chapter 8: Monte Carlo integration

We generate a random data point from a *simulation grid*. Let

$$A = \{\text{The data point lies below the curve}\}.$$

Then

$$\Pr[A] = \frac{\text{Area under curve}}{\text{Area of simulation grid}}$$

Area under curve $=$

Quick Recap – Chapter 8: Monte Carlo integration

We generate a random data point from a *simulation grid*. Let

$$A = \{\text{The data point lies below the curve}\}.$$

Then

$$\Pr[A] = \frac{\text{Area under curve}}{\text{Area of simulation grid}}$$

$$\text{Area under curve} = \int_a^b f(x)\,dx =$$

Quick Recap – Chapter 8: Monte Carlo integration

We generate a random data point from a *simulation grid*. Let

$$A = \{\text{The data point lies below the curve}\}.$$

Then

$$\Pr[A] = \frac{\text{Area under curve}}{\text{Area of simulation grid}}$$

$$\text{Area under curve} = \int_a^b f(x)\,dx = \Pr[A] \times \text{Area of simulation grid}$$

Quick Recap – Chapter 8: Monte Carlo integration

We generate a random data point from a *simulation grid*. Let

$$A = \{\text{The data point lies below the curve}\}.$$

Then

$$\Pr[A] = \frac{\text{Area under curve}}{\text{Area of simulation grid}}$$

$$\text{Area under curve} = \int_a^b f(x)\,dx = \Pr[A] \times \text{Area of simulation grid}$$

$$\approx \left[\frac{\text{No. of ``hits''}}{\text{No. of points simulated}}\right]$$

$$\times \text{ Area of simulation grid}$$

Part IX

Kernel Density Estimation

9.1 Introduction

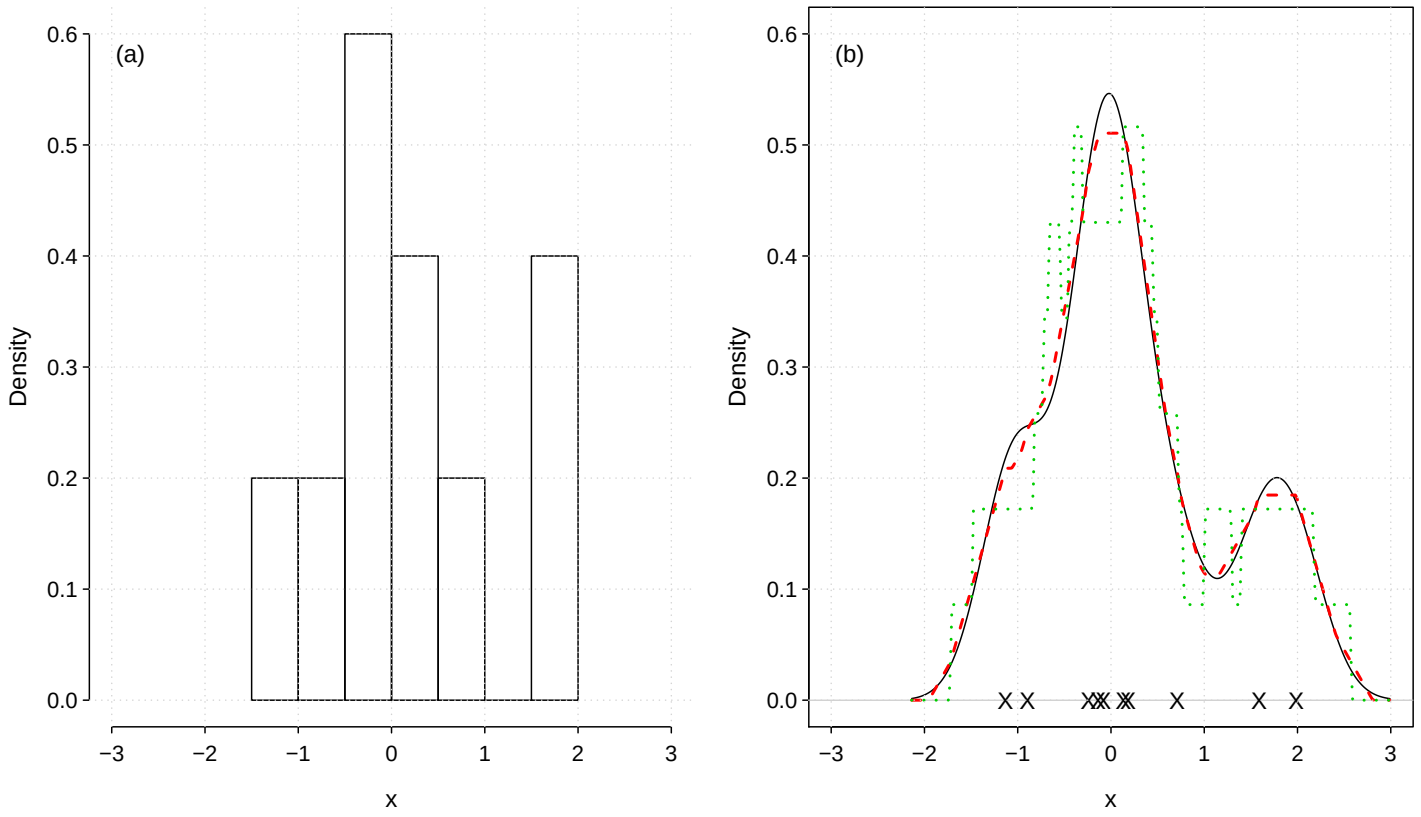


Figure: (a) Histogram of ten values sampled from a $N(0,1)$ distribution. (b) Three different Kernel density estimators. The data are the X's.

9.2 Definition

A kernel is a non-negative real-valued integrable function K which satisfies the following two requirements:

$$\int_{-\infty}^{\infty} K(t)\, dt = 1 \tag{9.1}$$

and

$$K(-t) = K(t) \quad \text{for all values of } t\ . \tag{9.2}$$

Expression 9.1 ensures that the kernel is a pdf, whilst Expression 9.2 makes the distribution symmetric about 0.

9.2 Definition

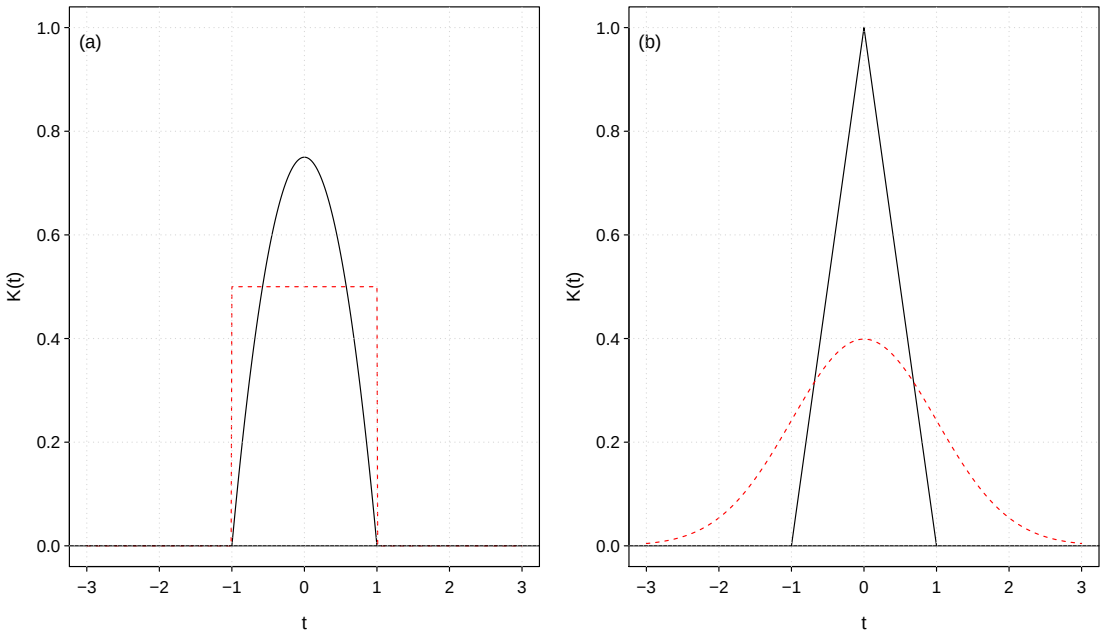


Figure: (a) The Epanechnikov and Uniform kernels. (b) The triangular and Gaussian kernel.

9.2 Definition (Standard kernels):

Epanechnikov:

$$K(t) = \begin{cases} \frac{3}{4}(1 - t^2) & -1 < t < 1 \\ 0 & \text{otherwise.} \end{cases}$$

9.2 Definition (Standard kernels):

Epanechnikov:

$$K(t) = \begin{cases} \frac{3}{4}(1-t^2) & -1 < t < 1 \\ 0 & \text{otherwise.} \end{cases}$$

Uniform:

$$K(t) = \begin{cases} \frac{1}{2} & -1 < t < 1 \\ 0 & \text{otherwise.} \end{cases}$$

9.2 Definition (Standard kernels):

Epanechnikov:

$$K(t) = \begin{cases} \frac{3}{4}(1-t^2) & -1 < t < 1 \\ 0 & \text{otherwise.} \end{cases}$$

Uniform:

$$K(t) = \begin{cases} \frac{1}{2} & -1 < t < 1 \\ 0 & \text{otherwise.} \end{cases}$$

Triangular:

$$K(t) = \begin{cases} 1-|t| & -1 < t < 1 \\ 0 & \text{otherwise.} \end{cases}$$

9.2 Definition (Standard kernels):

Epanechnikov:

$$K(t) = \begin{cases} \frac{3}{4}(1-t^2) & -1 < t < 1 \\ 0 & \text{otherwise.} \end{cases}$$

Uniform:

$$K(t) = \begin{cases} \frac{1}{2} & -1 < t < 1 \\ 0 & \text{otherwise.} \end{cases}$$

Triangular:

$$K(t) = \begin{cases} 1-|t| & -1 < t < 1 \\ 0 & \text{otherwise.} \end{cases}$$

Gaussian:

$$K(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2} .$$

9.2 Definition (Standard kernels):

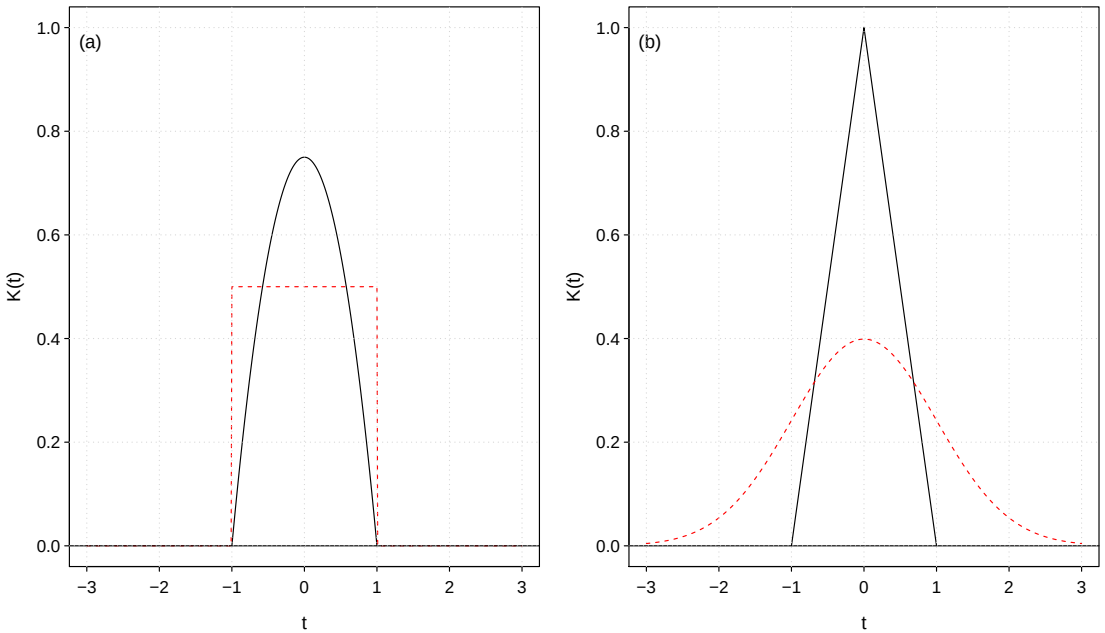


Figure: (a) The Epanechnikov and Uniform kernels. (b) The triangular and Gaussian kernel.

9.3 The general idea

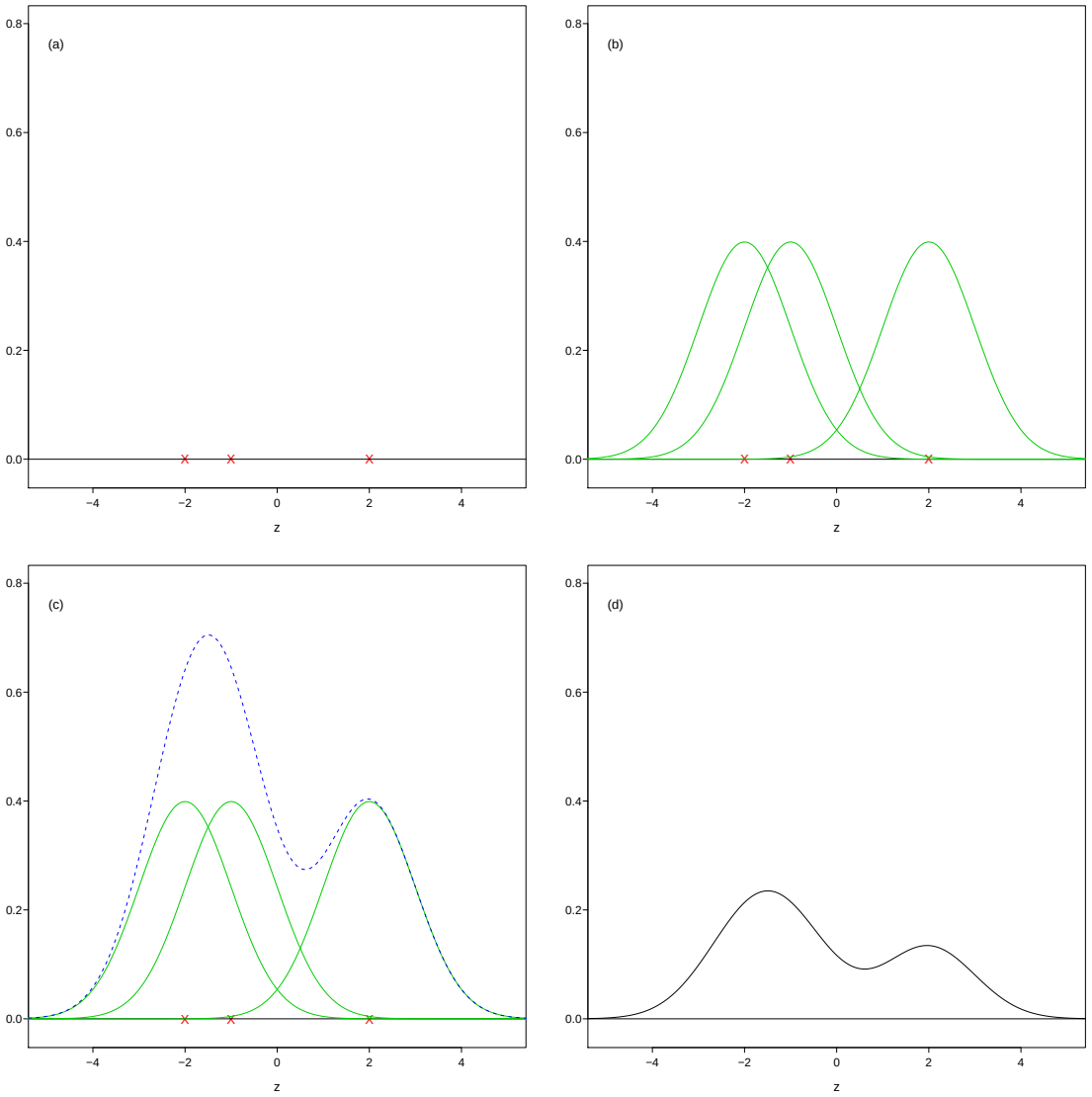


Figure: (a) Plot showing the data. (b) A Gaussian kernel is drawn round each point. (c) The kernels are combined. (d) The function is scaled to give an area of one.

9.3 The general idea

Kernel density estimation can be summarised in four steps:

9.3 The general idea

Kernel density estimation can be summarised in four steps:

1. We have some sample data. In Figure 9.3a we have three points, so $n = 3$.

9.3 The general idea

Kernel density estimation can be summarised in four steps:

1. We have some sample data. In Figure 9.3a we have three points, so $n = 3$.
2. Around each of the data points, we draw a kernel.

9.3 The general idea

Kernel density estimation can be summarised in four steps:

1. We have some sample data. In Figure 9.3a we have three points, so $n = 3$.
2. Around each of the data points, we draw a kernel. In Figure 9.3b we have used a Gaussian kernel. However, we could have used a Uniform, triangular, or Epanechnikov kernel.

9.3 The general idea

Kernel density estimation can be summarised in four steps:

1. We have some sample data. In Figure 9.3a we have three points, so $n = 3$.
2. Around each of the data points, we draw a kernel. In Figure 9.3b we have used a Gaussian kernel. However, we could have used a Uniform, triangular, or Epanechnikov kernel.
3. Next we combine the kernels - the blue dashed line in Figure 9.3c.

9.3 The general idea

Kernel density estimation can be summarised in four steps:

1. We have some sample data. In Figure 9.3a we have three points, so $n = 3$.
2. Around each of the data points, we draw a kernel. In Figure 9.3b we have used a Gaussian kernel. However, we could have used a Uniform, triangular, or Epanechnikov kernel.
3. Next we combine the kernels - the blue dashed line in Figure 9.3c.
4. The final step is to normalise the distribution.

9.3 The general idea

Kernel density estimation can be summarised in four steps:

1. We have some sample data. In Figure 9.3a we have three points, so $n = 3$.
2. Around each of the data points, we draw a kernel. In Figure 9.3b we have used a Gaussian kernel. However, we could have used a Uniform, triangular, or Epanechnikov kernel.
3. Next we combine the kernels - the blue dashed line in Figure 9.3c.
4. The final step is to normalise the distribution. In our example, since we have three points, the total area under the blue dashed curve is 3.

9.3 The general idea

Kernel density estimation can be summarised in four steps:

1. We have some sample data. In Figure 9.3a we have three points, so $n = 3$.
2. Around each of the data points, we draw a kernel. In Figure 9.3b we have used a Gaussian kernel. However, we could have used a Uniform, triangular, or Epanechnikov kernel.
3. Next we combine the kernels - the blue dashed line in Figure 9.3c.
4. The final step is to normalise the distribution. In our example, since we have three points, the total area under the blue dashed curve is 3. Hence, to recover a density we divide by 3 to get the black curve in Figure 9.3d.

9.4 Estimation

Let K be a kernel

9.4 Estimation

Let K be a kernel and suppose our sample contains n values:

9.4 Estimation

Let K be a kernel and suppose our sample contains n values: $x_1, \ldots, x_n$.

9.4 Estimation

Let K be a kernel and suppose our sample contains n values: $x_1, \ldots, x_n$. Then our estimate of the true pdf $f(x)$ is

9.4 Estimation

Let K be a kernel and suppose our sample contains n values: $x_1, \ldots, x_n$. Then our estimate of the true pdf $f(x)$ is

$$\hat{f}(x) =$$

9.4 Estimation

Let K be a kernel and suppose our sample contains n values: $x_1, \ldots, x_n$. Then our estimate of the true pdf $f(x)$ is

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} K(x - x_i) \quad . \tag{9.3}$$

9.4 Estimation

Let K be a kernel and suppose our sample contains n values: $x_1, \ldots, x_n$. Then our estimate of the true pdf $f(x)$ is

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} K\left(x - x_i\right) \quad . \tag{9.3}$$

Notice we use $K\left(x - x_i\right)$, since we draw a kernel around *each* x_i.

9.4 Estimation

It's fairly straightforward to see that $\hat{f}(x)$ is also a pdf, namely

9.4 Estimation

It's fairly straightforward to see that $\hat{f}(x)$ is also a pdf, namely

$$\int_{-\infty}^{\infty} \hat{f}(x)\, dx \quad =$$

9.4 Estimation

It's fairly straightforward to see that $\hat{f}(x)$ is also a pdf, namely

$$\int_{-\infty}^{\infty} \hat{f}(x)\, dx \quad = \quad \frac{1}{n} \sum_{i=1}^{n} \int_{-\infty}^{\infty} K\left(x - x_i\right)\, dx$$

9.4 Estimation

It's fairly straightforward to see that $\hat{f}(x)$ is also a pdf, namely

$$\begin{aligned}\int_{-\infty}^{\infty} \hat{f}(x)\,dx &= \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{\infty} K\left(x - x_i\right)\,dx \\ &= \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{\infty} K(y)\,dy\end{aligned}$$

9.4 Estimation

It's fairly straightforward to see that $\hat{f}(x)$ is also a pdf, namely

$$\begin{aligned}
\int_{-\infty}^{\infty} \hat{f}(x)\, dx &= \frac{1}{n}\sum_{i=1}^{n} \int_{-\infty}^{\infty} K\left(x - x_i\right)\, dx \\
&= \frac{1}{n}\sum_{i=1}^{n} \int_{-\infty}^{\infty} K(y)\, dy \\
&= \frac{1}{n}\sum_{i=1}^{n} 1
\end{aligned}$$

9.4 Estimation

It's fairly straightforward to see that $\hat{f}(x)$ is also a pdf, namely

$$\begin{aligned}
\int_{-\infty}^{\infty} \hat{f}(x)\, dx &= \frac{1}{n}\sum_{i=1}^{n} \int_{-\infty}^{\infty} K\left(x - x_i\right)\, dx \\
&= \frac{1}{n}\sum_{i=1}^{n} \int_{-\infty}^{\infty} K(y)\, dy \\
&= \frac{1}{n}\sum_{i=1}^{n} 1 \\
&= \frac{1}{n}\left(1 + 1 + \ldots + 1\right)
\end{aligned}$$

9.4 Estimation

It's fairly straightforward to see that $\hat{f}(x)$ is also a pdf, namely

$$\begin{aligned}
\int_{-\infty}^{\infty} \hat{f}(x)\, dx &= \frac{1}{n} \sum_{i=1}^{n} \int_{-\infty}^{\infty} K\left(x - x_i\right)\, dx \\
&= \frac{1}{n} \sum_{i=1}^{n} \int_{-\infty}^{\infty} K(y)\, dy \\
&= \frac{1}{n} \sum_{i=1}^{n} 1 \\
&= \frac{1}{n} \left(1 + 1 + \ldots + 1\right) \\
&= \frac{1}{n} \times n = 1 \,. \qquad (9.4)
\end{aligned}$$

Example: River Monsters

The data shown below are the lengths (to the nearest cm) of 10 Giant Groupers caught by expert angler Jeremy Wade for the TV series *River Monsters*.

This sample was taken in 2013 in a lake near the Chernobyl nuclear disaster of 1986. These fish usually grow to around 75cm in length, but genetic mutations caused by the nuclear explosions are thought to have increased the size of this species.

101	97	99	104	103	94	102	94	102	106

Example: River Monsters

Example: River Monsters

Produce a density plot for these data using the Gaussian kernel.

Example: River Monsters

Produce a density plot for these data using the Gaussian kernel.

Let X represent the length of a fish. We have

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} K\left(x - x_i\right),$$

where

$$K(x - x_i) = \frac{1}{\sqrt{2\pi}} \exp\left\{ -\frac{(x - x_i)^2}{2} \right\}.$$

Example: River Monsters

Produce a density plot for these data using the Gaussian kernel.

Let X represent the length of a fish. We have

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} K\left(x - x_i\right),$$

where

$$K(x - x_i) = \frac{1}{\sqrt{2\pi}} \exp\left\{ -\frac{(x - x_i)^2}{2} \right\}.$$

The range of our data is 94$\longrightarrow$106, so let's plot over the range 90$\longrightarrow$110.

Example: River Monsters

Produce a density plot for these data using the Gaussian kernel.

Let X represent the length of a fish. We have

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} K\left(x - x_i\right),$$

where

$$K(x - x_i) = \frac{1}{\sqrt{2\pi}} \exp\left\{ -\frac{(x - x_i)^2}{2} \right\}.$$

The range of our data is $94 \longrightarrow 106$, so let's plot over the range $90 \longrightarrow 110$. For example,

Example: River Monsters

Produce a density plot for these data using the Gaussian kernel.

Let X represent the length of a fish. We have

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} K\left(x - x_i\right),$$

where

$$K(x - x_i) = \frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{(x - x_i)^2}{2}\right\}.$$

The range of our data is $94 \longrightarrow 106$, so let's plot over the range $90 \longrightarrow 110$. For example,

$$\begin{aligned}\hat{f}(\mathbf{90}) &= \frac{1}{10}\left[\frac{1}{\sqrt{2\pi}}e^{-(\mathbf{90}-101)^2/2} + \ldots + \frac{1}{\sqrt{2\pi}}e^{-(\mathbf{90}-106)^2/2}\right] \\ &= 0.0000267\end{aligned}$$

Example: River Monsters

As another example:

Example: River Monsters

As another example:

$$\begin{aligned}\hat{f}(\mathbf{94}) &= \frac{1}{10}\left[\frac{1}{\sqrt{2\pi}}e^{-(\mathbf{94}-101)^2/2} + \ldots + \frac{1}{\sqrt{2\pi}}e^{-(\mathbf{94}-106)^2/2}\right] \\ &= 0.08023.\end{aligned}$$

Example: River Monsters

As another example:

$$\hat{f}(\mathbf{94}) = \frac{1}{10}\left[\frac{1}{\sqrt{2\pi}}e^{-(\mathbf{94}-101)^2/2} + \ldots + \frac{1}{\sqrt{2\pi}}e^{-(\mathbf{94}-106)^2/2}\right]$$

$$= 0.08023.$$

Also,

$$\hat{f}(\mathbf{102}) = \frac{1}{10}\left[\frac{1}{\sqrt{2\pi}}e^{-(\mathbf{102}-101)^2/2} + \ldots + \frac{1}{\sqrt{2\pi}}e^{-(\mathbf{102}-106)^2/2}\right]$$

$$= 0.13404.$$

Example: River Monsters

This gives:

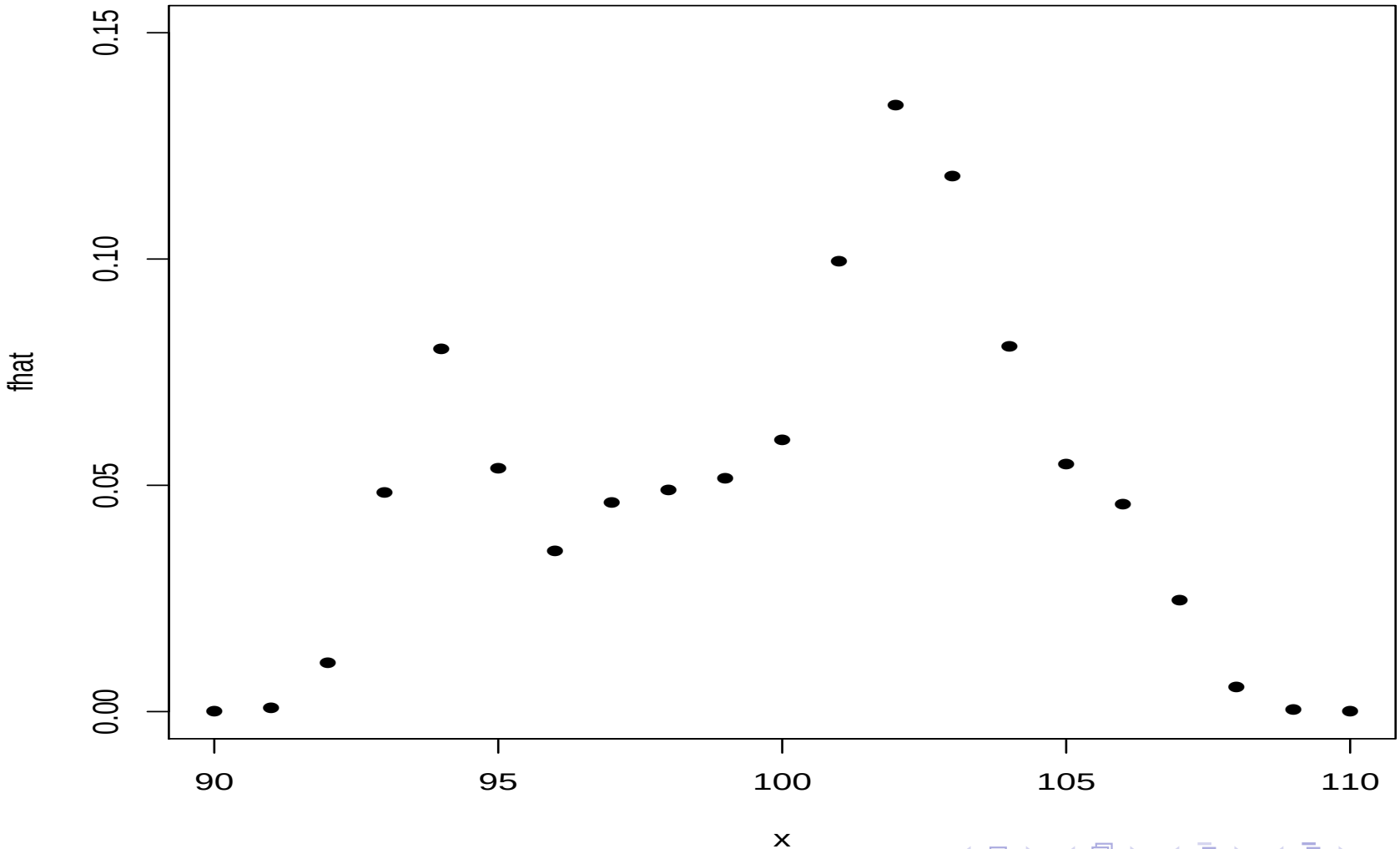
x	$\hat{f}(x)$	x	$\hat{f}(x)$
90	0.0000267	101	0.0995432
91	0.0000886	102	0.1340384
92	0.0107983	103	0.1183411
93	0.0484075	104	0.0807319
94	0.0802317	105	0.0546929
95	0.0538066	106	0.0457634
96	0.0354386	107	0.0246539
97	0.0461933	108	0.0054126
98	0.0488910	109	0.0004433
99	0.0515926	110	0.0000138
100	0.0600920		

Example: River Monsters

Plotting $\hat{f}(x)$ against x gives:

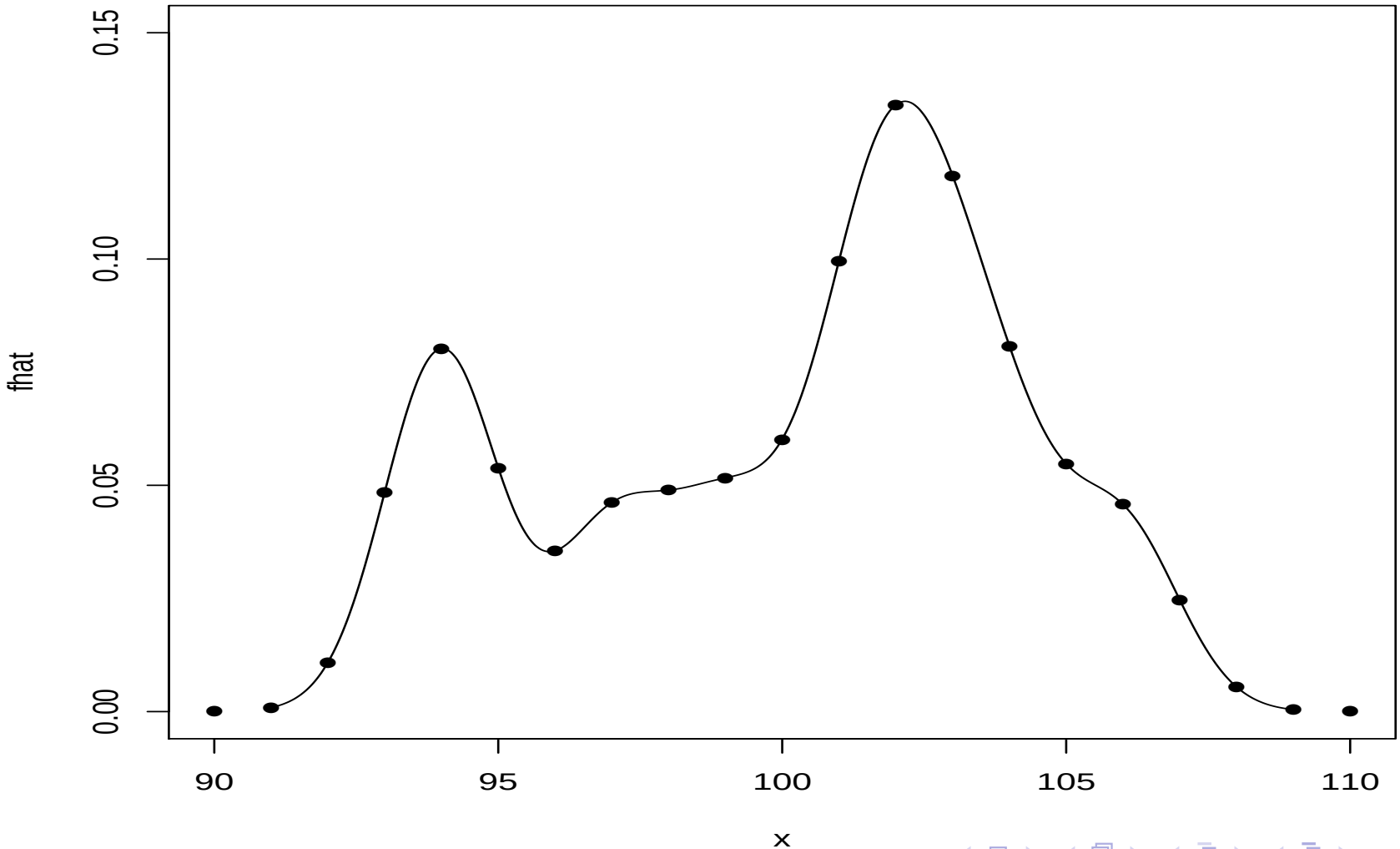
Example: River Monsters

Plotting $\hat{f}(x)$ against x gives:



Example: River Monsters

Plotting $\hat{f}(x)$ against x gives:



Example: River Monsters

Plotting $\hat{f}(x)$ against x gives: