STAT5003

Week 3: Density Estimation

Jaslene Lin

The University of Sydney

THE UNIVERSITY OF SYDNEY

Readings and R functions covered

Important

- **Introduction to Statistical Learning**
  - Chapter 2.2 the bias-variance tradeoff
- **R** functions
- `dbinom` `dnorm` (density functions)
- `rnorm` (generate random values)
- `hist` (Histogram)
- `density` (nonparametric density estimation)
- `stats4::mle` (Maximum Likelihood estimation)

Review on probability distribution functions

Discrete distributions

For any random variable X with a discrete distribution, there is a sample space Ω with finite or countably infinite number of possible values (outcomes) $x = \{x_1, x_2, \ldots\}$ and associated probabilities $\{p_1, p_2, \ldots\}$

The point probabilities (aka **probability mass function**) for each value of x are denoted $f(x)$ and the cumulative distribution function denoted $F(x)$ where

$$f(x) = P(X = x), \qquad F(x) = P(X \leq x)$$

Properties:

- There is a *countable* number of possible values
- $\sum_{i=1}^{\infty} p_i = 1$, where $p_i = f(x_i) = P(X = x_i)$
- $p_i \geq 0$
- Is it possible that $p_i > 1$?

Discrete distributions

An example: Throwing One Fair Die

Let X be the random variable representing the outcome of throwing one fair six-sided die. The possible values of X are:

$$X \in \{1, 2, 3, 4, 5, 6\}$$

For a fair six-sided die, each outcome is equally likely. Therefore, the probability mass distribution for (X) is:

$$P(X = x) = \begin{cases} \frac{1}{6} & \text{if } x \in \{1, 2, 3, 4, 5, 6\} \\ 0 & \text{otherwise} \end{cases}$$

We can also represent this distribution in a table:

Probability Mass Distribution for a Fair Die

x	P(X = x)
1	0.1666667
2	0.1666667
3	0.1666667
4	0.1666667
5	0.1666667
6	0.1666667

Throwing One Fair Die

Suppose we want to **model** the number of times, S that we roll a 5 in 60 throws of a die.

What is the variable type of S and what its sample space?

- ○ S is a discrete variable and the sample space is $(0, 60]$.
- ○ S is not a random variable.
- ○ S is a continuous variable and the sample space is all the positive values.
- ○ S is a discrete variable and the sample space is $[0, 60]$.

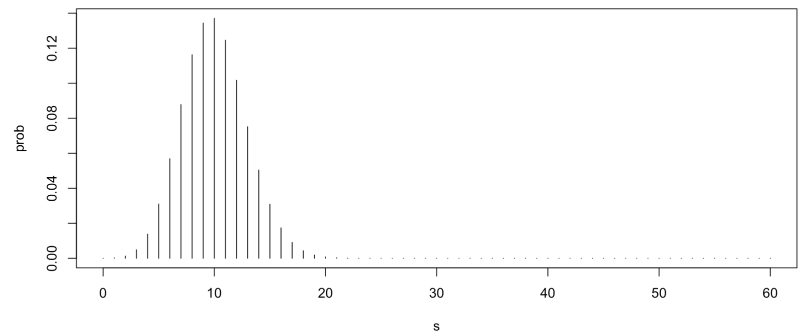
Binomial distribution

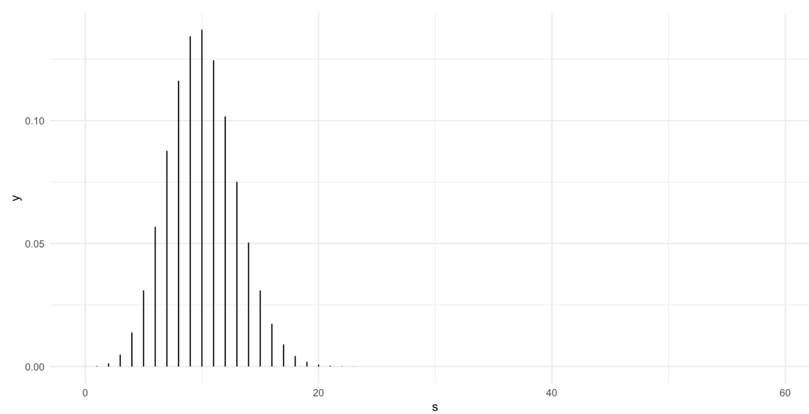
$S \sim Binomial(\frac{1}{6}, 60)$

```r
library(ggplot2)
s <- 0:60
prob <- dbinom(s, size = 60, prob = 1/6)
# Base R graphics
 plot(s, prob, type = "h")
dat <- data.frame(x = s, y = prob)
# ggplot2 version
 ggplot(dat,
       aes(x = s, y = y, xend = s, yend = 0)) +
  geom_segment() + theme_minimal()
```

$$f(s) = \begin{cases} \binom{n}{s} p^s (1-p)^{n-s}, & s = 0, 1, 2, \ldots, n \\ 0, & \text{otherwise} \end{cases}$$

The $\binom{n}{s}$ are known as the binomial coefficients. **The parameter** p is the probability of success.





Quiz

Suppose an online store has a 20% chance of a visitor making a purchase (success) each time they visit the site. If you want to model the number of purchases made by 50 visitors, you can use a binomial distribution where:

the number of trials (n) and the probability of success (p) are

- ○ $n = 50$; p is unknown and to be estimated
- ○ n is the number of purchases made by 50 visitors; $p = 0.2$
- ○ $n = 50; p = 0.8$
- ○ $n = 50; p = 0.2$

Continuous distributions

- A continuous random variable X is where the outcome can take an infinite (uncountable) number of possible values.
  - ➟ These values may be within a fixed or unbounded interval.
- For example, the average temperature range in Sydney is within the range of [8.8, 25.8] celsius.

The point probabilities for each value of x is $P(X = x) = 0$ and the cumulative distribution function

$$F(x) = \int_{-\infty}^{x} f(t)\,dt = P(X \leq x)$$

- There are an uncountable number of possible values
- $f(x)$ is called the probability density function; $f(x) \geq 0$ (non-negative)
- $\int_{-\infty}^{\infty} f(x)\,dx = 1$ (unit measure)
- Is it possible that $f(x) > 1$?

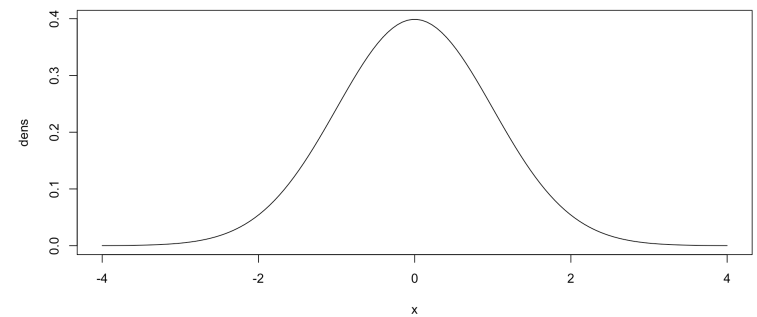
```r
dnorm(10, 10, 0.1)
```

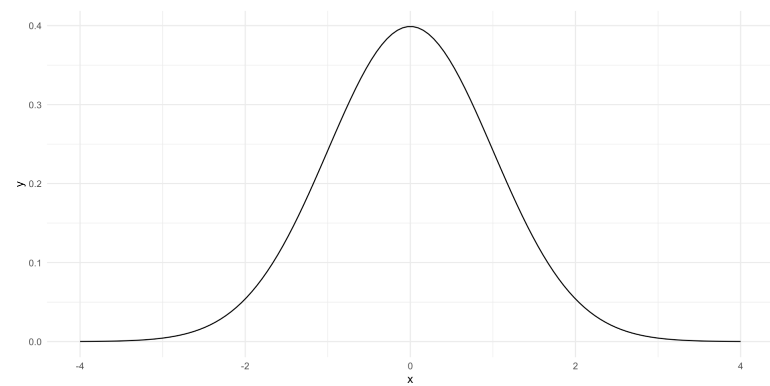
```
[1] 3.989423
```

Normal (Gaussian) distribution: $f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

- The most famous continuous distribution
- Fully specified by two parameters
  - μ the location parameter (mean)
  - σ the scale parameter (sd)
- Notation $X \sim \mathcal{N}(\mu, \sigma^2)$

```r
mu <- 0; sig <- 1
x <- seq(from = mu - 4 * sig, to = mu + 4 * sig,
         length.out = 128)
dens <- dnorm(x, mean = mu, sd = sig)
# Base R graphics
plot(x, dens, type = "l")
dat <- data.frame(x = x, y = dens)
# ggplot2 version
ggplot(dat, aes(x = x, y = y)) +
  geom_line() + theme_minimal()
```





Density estimation - Likelihood approach

Density estimation

Suppose random variables $X_1, X_2, \ldots, X_n$ have been observed and assumed to be sampled independently from the distribution with density f

Goal: The estimation of the density function f

Applications of density estimation in exploratory data analysis (EDA):

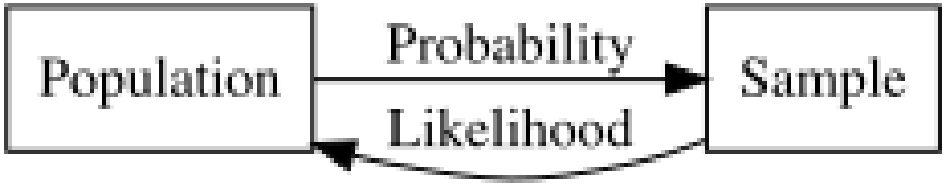
- to assess multimodality, skew, tail behaviour, etc.
- in decision making, classification, and summarizing Bayesian posteriors
- as a useful visualisation tool (a simple summary of a distribution)

Parametric density estimation

- **The parametric approach** to density estimation assumed parametric model.
- That is, $X_1, X_2, \ldots, X_n \overset{\text{i.i.d.}}{\sim} f\boldsymbol{\theta}$ where $\boldsymbol{\theta}$ is a parameter vector.
  - For example, $\boldsymbol{\theta} = (\mu, \sigma)$ when $X \sim \mathcal{N}(\mu, \sigma^2)$.
  - For example, $\boldsymbol{\theta} = p$ when $X \sim \mathcal{B}inomial(n, p)$.

Density Function vs Likelihood Function

- **Density Function**: $f(X|\boldsymbol{\theta})$ represents the probability density of observing data X given parameters $\boldsymbol{\theta}$.
- **Likelihood Function**: $L(\boldsymbol{\theta}|x)$ is the function used to estimate $\boldsymbol{\theta}$ based on observed data x.
- Maximum likelihood estimator (θ_{mle}) is the value of $\boldsymbol{\theta}$ that maximise the likelihood function.



Simple example:

Assuming the population has girl:boy ratio of 2:1 $(\theta_{boy} = \frac{1}{3}; \theta_{girl} = \frac{2}{3})$

- If I draw a sample of 50 people, what is the probability of picking 10 boys

$$P(Y = 10|\theta = \tfrac{1}{3}; n = 50)$$

- If I draw a sample of 50 people, and picked 10 boys, what is the likelihood that the girl:boy ratio is 2:1

$$P(\theta = \tfrac{1}{3}|y = 10; n = 50)$$

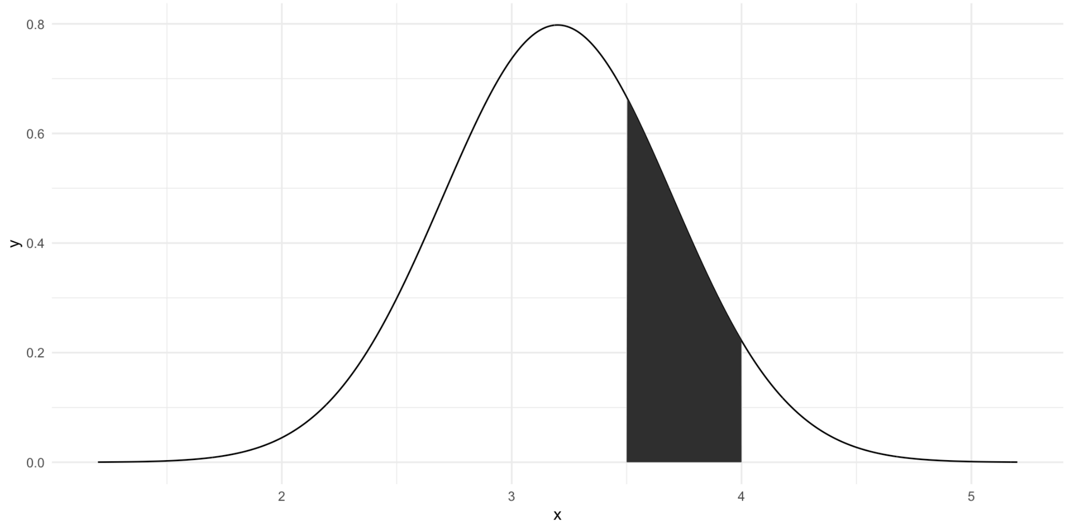
Normal distribution example

- Consider a random variable $X \sim \mathcal{N}(3.2, 0.5^2)$
- What is the probability that X is between 3.5 and 4?
    - ➟ Compute the area under the density: $P(3.5 \leq X \leq 4) = \int_{3.5}^{4} f(t)\, dt$

```r
mu = 3.2; sig = 0.5
pnorm(4, mean = mu, sd = sig) -
  pnorm(3.5, mean = mu, sd = sig)
```

```
[1] 0.2194538
```

```r
# Or in one line
## diff(pnorm(c(3.5, 4), mean = mu, sd = sig))
```



Likelihood

- Consider a single value is observed from $X \sim \mathcal{N}(\mu, 0.2^2)$, say $x = 3.7$
- Determine the likelihood of drawing this value. Flip the perspective $f(x|\theta)$ to $L(\theta|x)$

```r
dnorm(3.7, mean = 3.5, sd = 0.2)
```

```
[1] 1.209854
```

```r
dnorm(3.7, mean = 3.6, sd = 0.2)
```

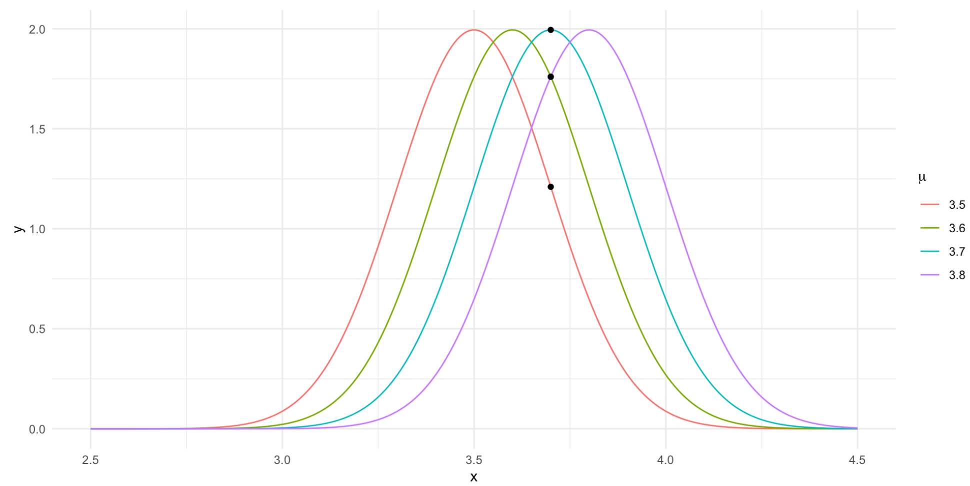
```
[1] 1.760327
```

```r
dnorm(3.7, mean = 3.7, sd = 0.2)
```

```
[1] 1.994711
```

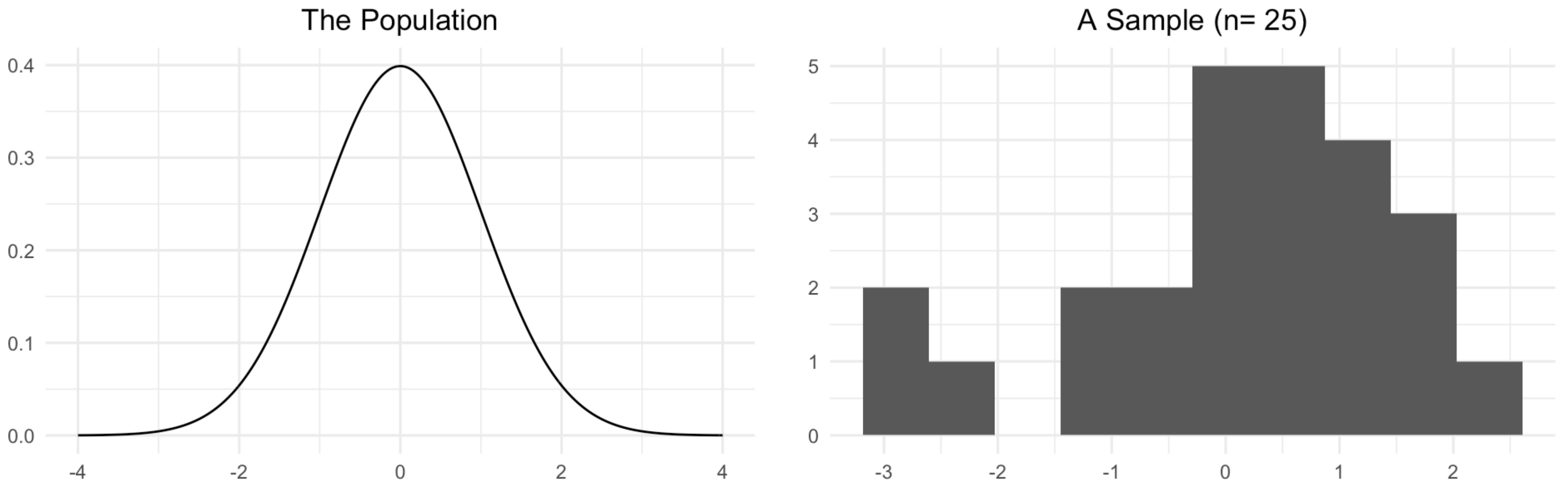
```r
dnorm(3.7, mean = 3.8, sd = 0.2)
```

```
[1] 1.760327
```



Maximum likelihood approach

- $f(x_1, x_2 \ldots, x_n|\boldsymbol{\theta})$ is the probability density of observing $x_1, x_2, \ldots, x_n$ given the parameter $\boldsymbol{\theta}$



- Assuming independent and identically distributed variables $f(x_1, x_2, \ldots, x_n|\boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i|\boldsymbol{\theta})$

Maximising the log-likelihood is often easier so it is common to maximise

$$L(\boldsymbol{\theta}|\boldsymbol{x}) = \prod_{i=1}^{n} f(x_i|\boldsymbol{\theta}) \rightarrow \mathcal{L}(\boldsymbol{\theta}|\boldsymbol{x}) = \ln L(\boldsymbol{\theta}|\boldsymbol{x}) = \sum_{i=1}^{n} \ln f(x_i|\boldsymbol{\theta})$$

Maximum likelihood approach

Denote Y as the number of boys picked from a sample 50 people

```r
# Given data
n <- 50  # total number of people
y <- 10  # number of boys

# Define the negative log-likelihood function
neg_log_likelihood <- function(p) {
  -dbinom(y, n, p, log = TRUE)
}

# Initial guess for the proportion of boys
initial_guess <- 0.5

# Optimize the negative log-likelihood function
result <- optim(initial_guess, neg_log_likelihood, method = "Brent", lower = 0, upper = 1)

# Extract the MLE for the proportion of boys
mle_proportion_boys <- result$par

# Print the MLE for the proportion of boys
mle_proportion_boys
```

```
[1] 0.2
```

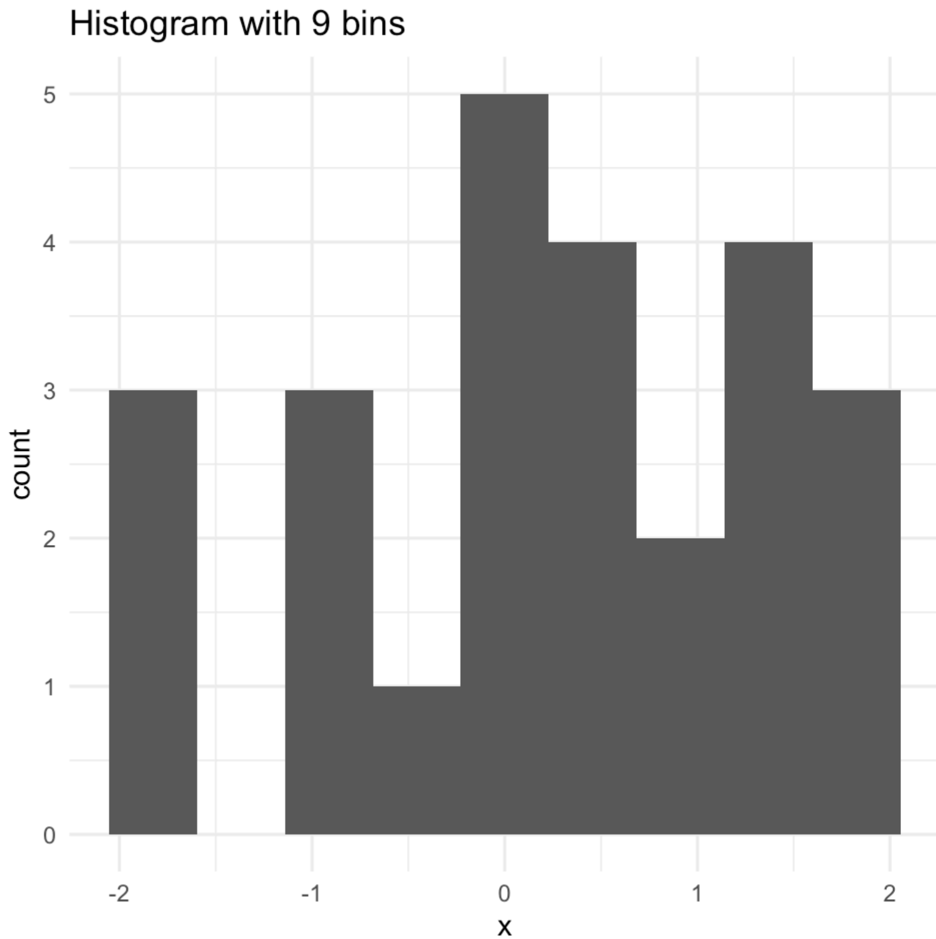
Density estimation - Non-parametric approach

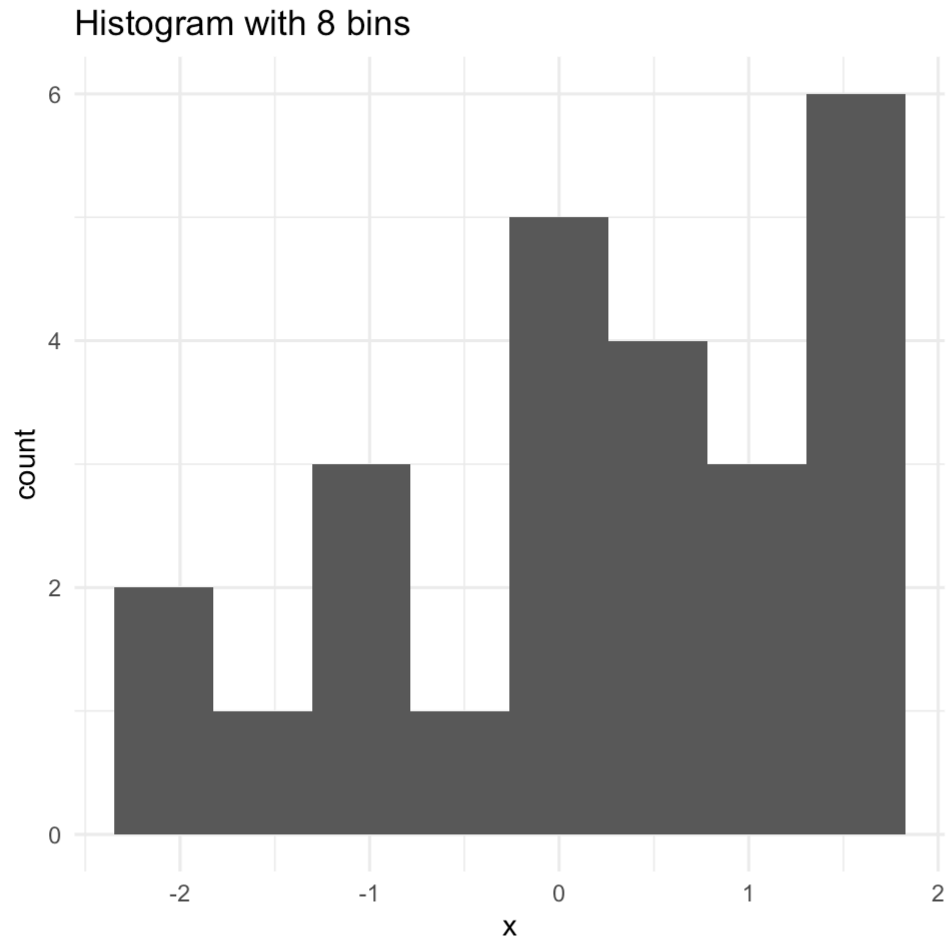
Non-parametric density estimation

- Danger of misspecification with parametric approach
  - ➟ If the assumed f_θ is incorrect
  - ➟ Serious danger of inferential errors
- Non-parametric approaches to density estimations
  - ➟ Assume little about the structure of f
  - ➟ Use *local information* to estimate f at a point x

Histograms

- one type of nonparametric density estimators
- piecewise constant density estimators
- Very simple visualization and easy to produce
- Sensitive to **the number of bins chosen** and **bin width**

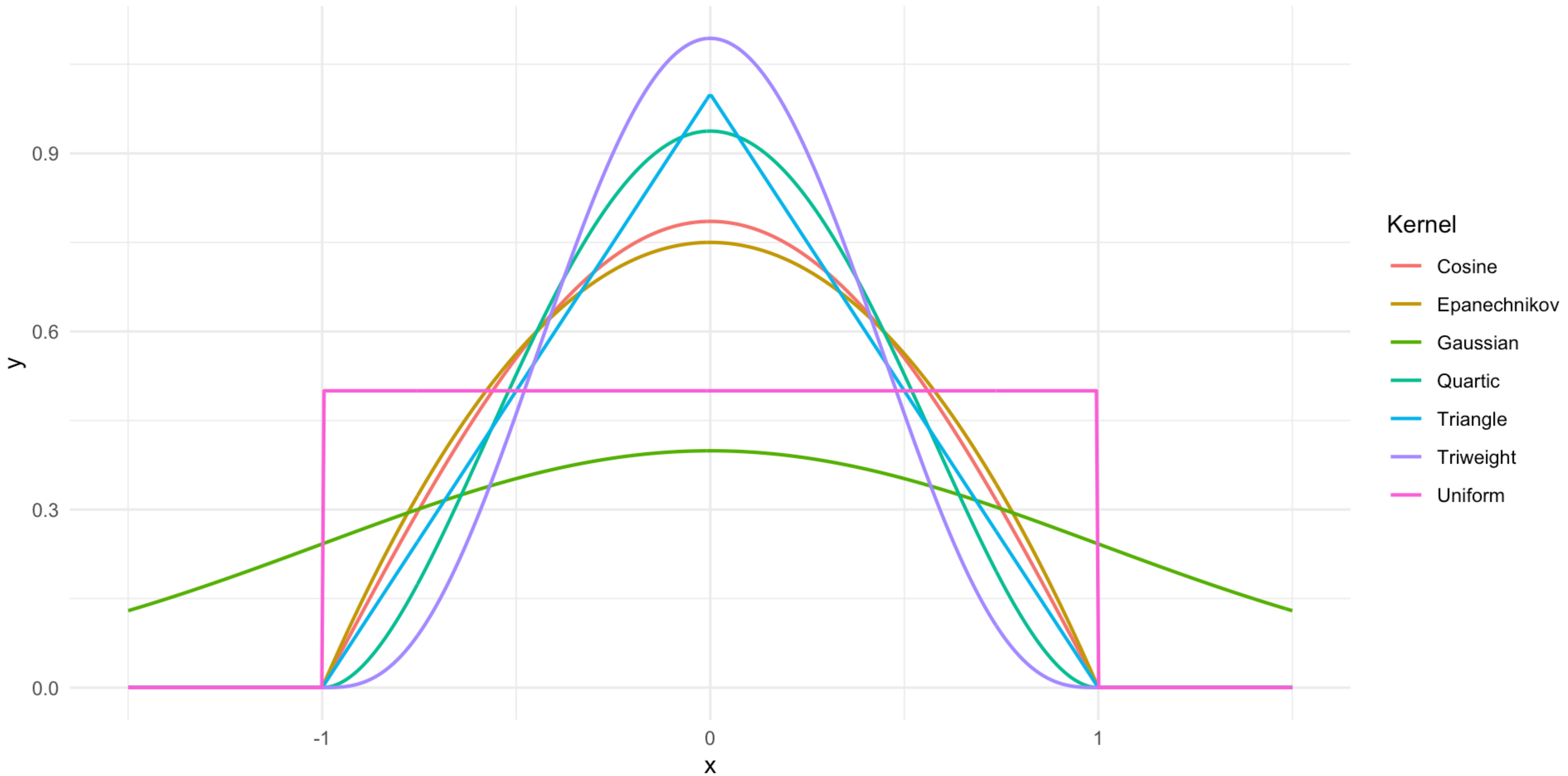




- Preferable to have a smooth estimate and not have columns

Kernel functions

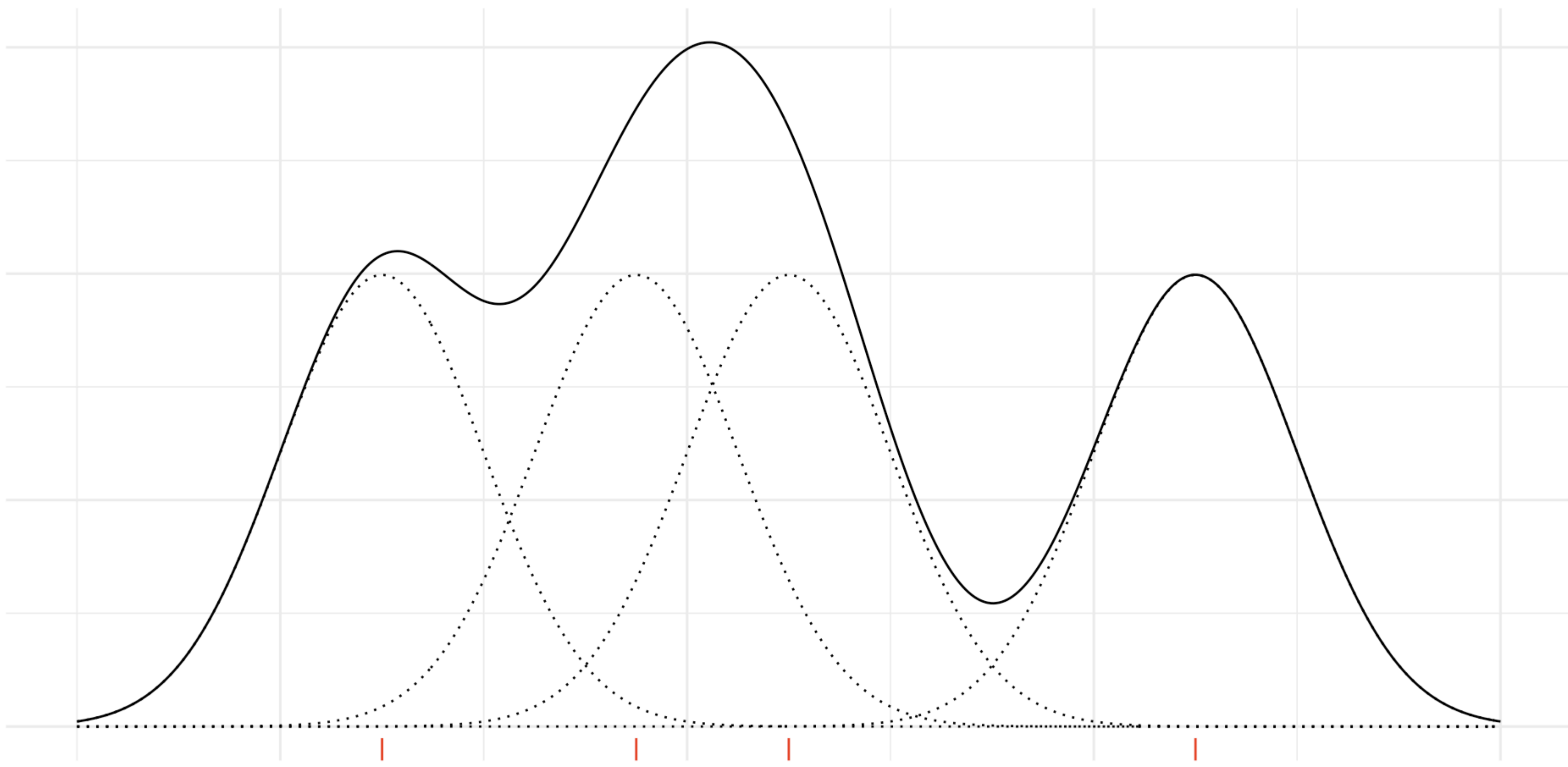
- A kernel is a special type of probability density function (PDF) having the properties.
  - ➟ non-negative $K(x) \geq 0$, symmetric $K(-x) = K(x)$, unit measure $\int K(x)\, dx = 1$



Kernel density esimation

- Kernel density estimation is a non-parametric approach estimating densities
  - ➟ Knowledge of the structure of f is not required
- **Essentially, at every data point, a kernel function is created with the point at its centre**
- The PDF is estimated by **adding all of these kernel functions and dividing by the number of data** to ensure that it satisfies:
  - ➟ every possible value of the PDF is non-negative
  - ➟ the definite integral of the PDF over its support set equals 1

Normal kernel density estimate



- Example: Four sampled variables marked in red with Gaussian weights sum together to give the overall density estimate

Kernel density estimator (KDE)

- A simple one weights all points within a window h of x equally

$$\hat{f}(x) = \frac{1}{2nh} \sum_{i=1}^{n} \mathbf{1}_{\{|X_i - x| < h\}}$$

  - ➟ $\mathbf{1}_A = 1$ if A is true and $\mathbf{1}_A = 0$ otherwise

- More generally a univariate kernel density estimator has a general weight function (Kernel)

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

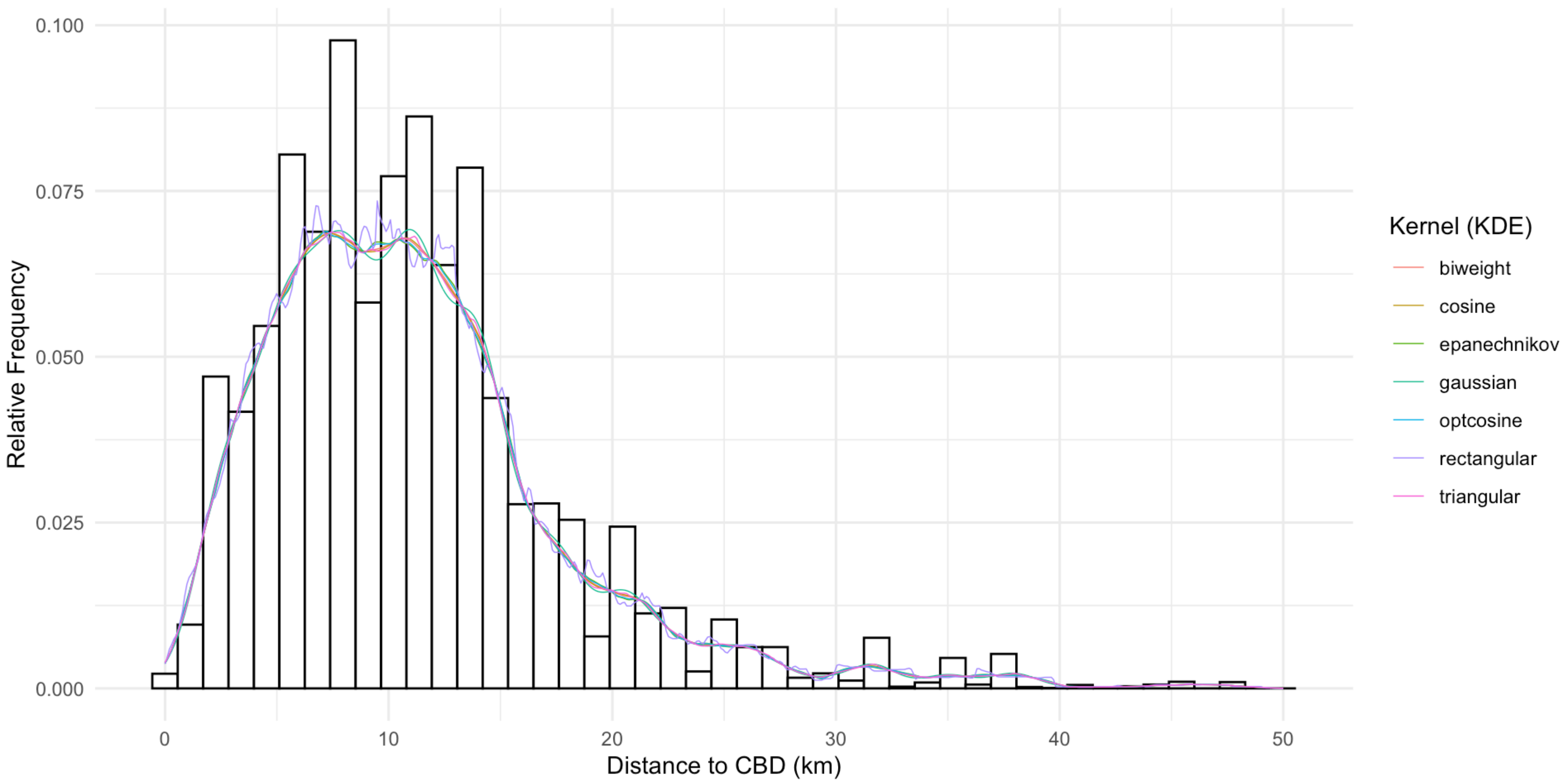
  - ➟ K is a Kernel function
  - ➟ h is a bandwidth parameter (possibly fixed or varying)
  - ➟ Consider only h fixed for this course

Tuning the Kernel density estimator (KDE)

- There are two main components for the KDE $\hat{f}(x) = \dfrac{1}{nh}\sum_{i=1}^{n} K\left(\dfrac{X_i - x}{h}\right)$
  - ➡ The choice of K
  - ➡ The choice of h
- The choice of Kernel is less important and generally gives similar results
- The choice of bandwidth is important and can vary the result greatly
- Some standard kernels

$$
\begin{array}{rl}
\text{Uniform} & K(x) = \frac{1}{2}\mathbf{1}_{\{|x|\leq 1\}} \\
\text{Gaussian} & K(x) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\} \\
\text{Epanechnikov} & K(x) = \frac{3}{4}(1 - x^2)\mathbf{1}_{\{|x|\leq 1\}}
\end{array}
$$

Different choices of Kernel function with same bandwidth



Choosing the bandwidth

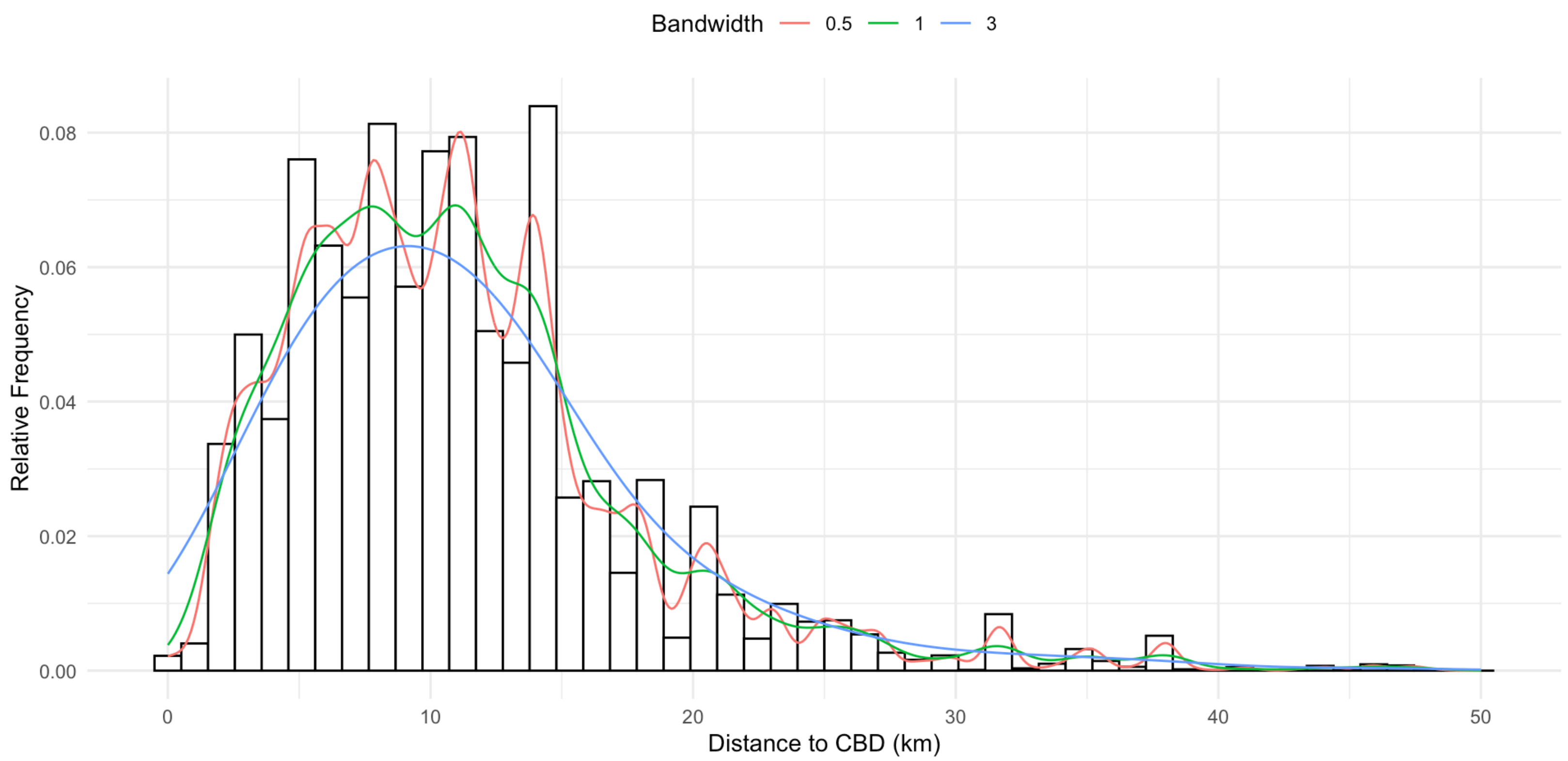
- The density estimator

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

  - ➟ is a fixed-bandwidth kernel density estimator since h is constant
- If h is too small, the density estimator will tend to assign probability density too locally near observed data
  - ➟ a wiggly estimated density function with many false modes
- If h is too large, the density estimator will spread probability density contributions too diffusely
  - ➟ smooths away important features of f

Choice of bandwidth

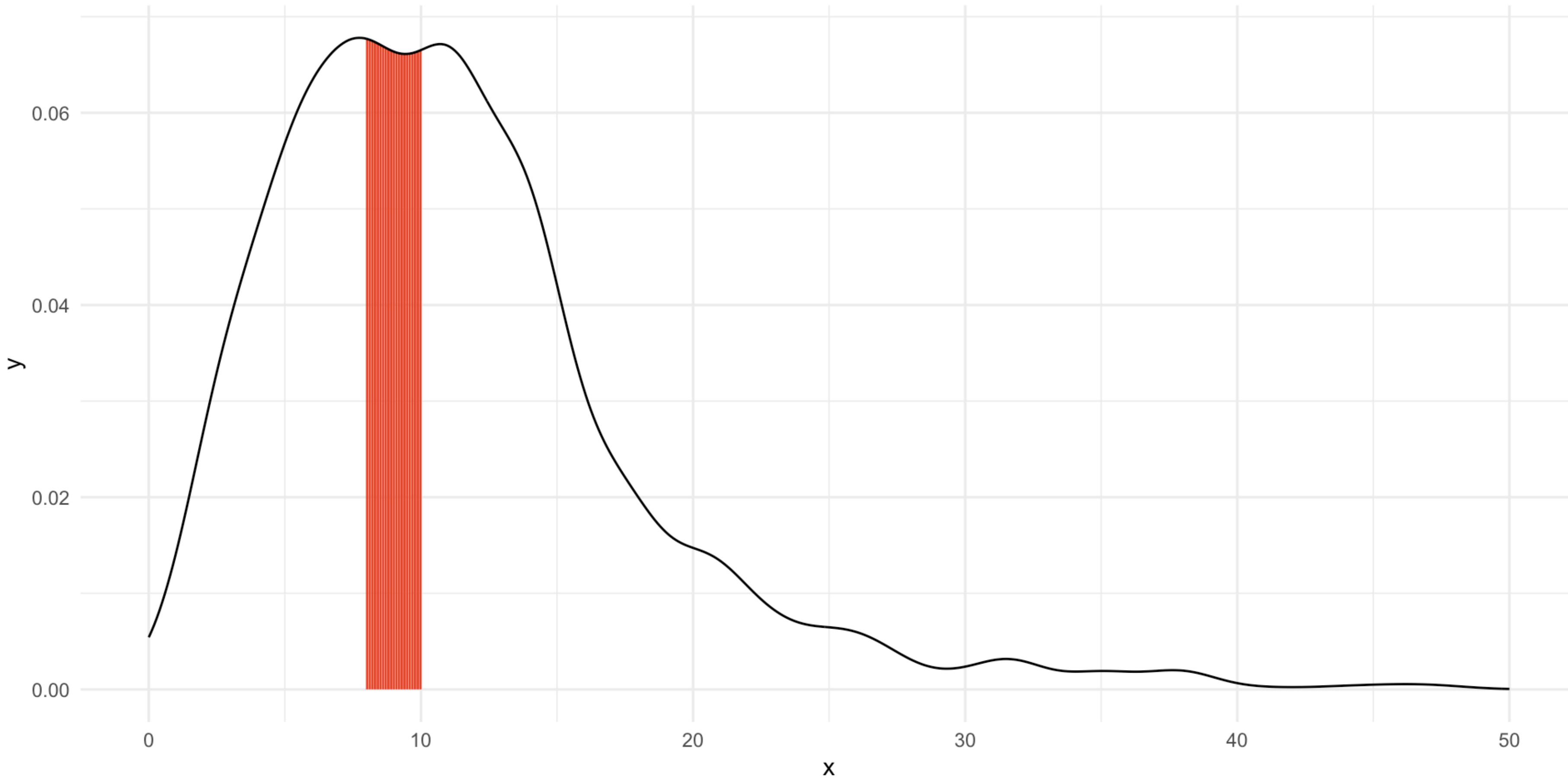
- Consider the distance from CBD variable again with three bandwidths



- A bias and variance trade-off
  - A small bandwidth gives high variance
  - A large bandwidth gives high bias

Uses of the density estimate

- Compute probabilities: Consider the probability a property is between 8 - 10 km of CBD
- Integrate the density function between 8 and 10 yields $p = 0.13$, meaning 13% chance of finding a property between 8 - 10 km of CBD



Computing density in `r-project`

- Base `r-project` there is `density`
  - `density` computes the KDE
  - Can specify the bandwidth with `bw` argument
  - Can inspect details in `summary`

```r
density.cbd <- density(x = distance.cbd, bw = 1.25, from = 0, to = 50)
```

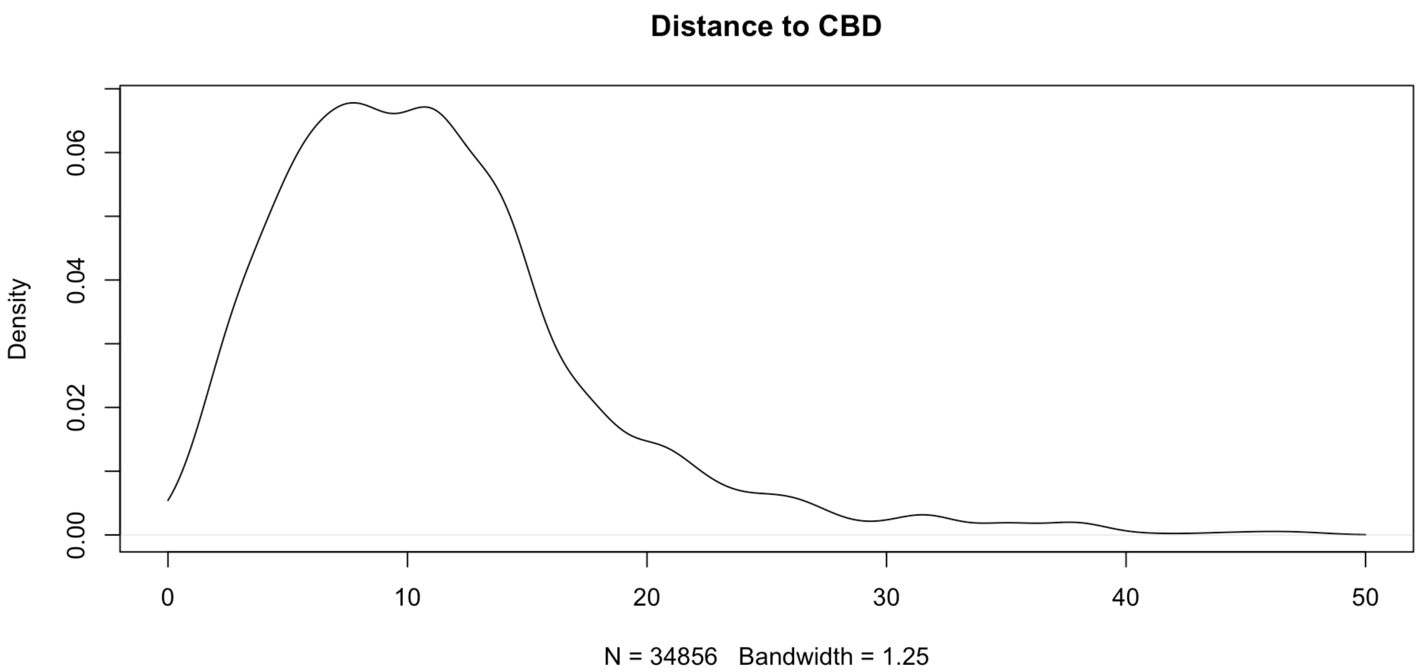
```r
summary(density.cbd)
```

```
           Length Class  Mode
x          512    -none- numeric
y          512    -none- numeric
bw           1    -none- numeric
n            1    -none- numeric
old.coords   1    -none- logical
call         5    -none- call
data.name    1    -none- character
has.na       1    -none- logical
```

Computing density in `r-project`

- Visualization
  - Can wrap in `plot`, i.e. `plot(density(x))`, to visualize
  - For plotting ggplot there is `geom_density`

```
plot(density.cbd, main = "Distance to CBD")
```



- For plotting ggplot there is `geom_density`

Mean squared error, Bias, and Variance

We can decompose the mean squared error (MSE) into the sum of three quantities: The variance, the squared bias, and the variance of the error:

Assume $Y = f(X) + \epsilon$ and we have an estimator $\hat{f}(X)$ of $f(X)$

$$\mathbb{E}\left(Y - \hat{f}(X)\right)^2 = \mathrm{Var}(\hat{f}(X)) + \left[\mathrm{Bias}(\hat{f}(X))\right]^2 + \mathrm{Var}(\epsilon)$$

- Variance here denoting how much would $\hat{f}(x)$ change if we estimate using a different training set.
- Bias: Error introduced by approximating the data using a model.

Model Flexility and Prediction Accuracy

```
Linear Model Training MSE: 0.6597052
Smoothing Spline Model Training MSE: 0.2175801
Random Forest Model Training MSE: 0.09278985
Linear Model Test MSE: 1.306374
Smoothing Spline Model Test MSE: 1.074712
Random Forest Model Test MSE: 1.085234
```