STAT5003

Week 3 : Density Estimation

Dr. Justin Wishart

THE UNIVERSITY OF SYDNEY

Readings and functions covered

- For the bias variance tradeoff see Section 2.2 James, Witten, Hastie, and Tibshirani (2013)
- 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

THE UNIVERSITY OF SYDNEY

Discrete distributions

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

The point probabilities 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:

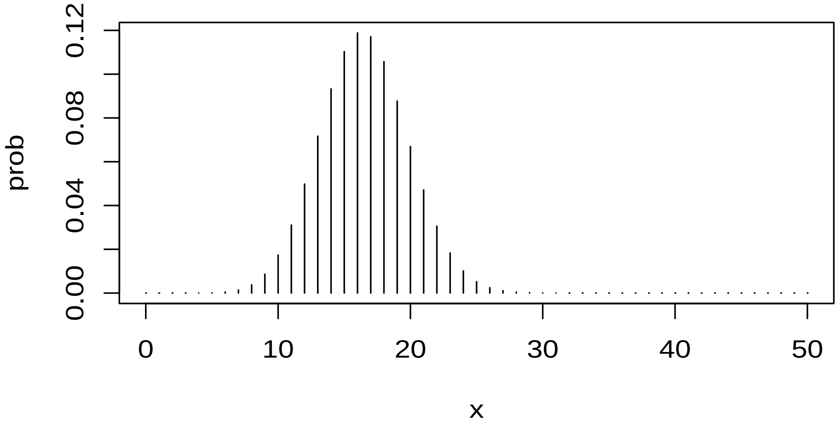
- There is a *countable* number of possible values;
- $\sum_{i=1}^{\infty} p_i = 1$
- $p_i \geq 0$

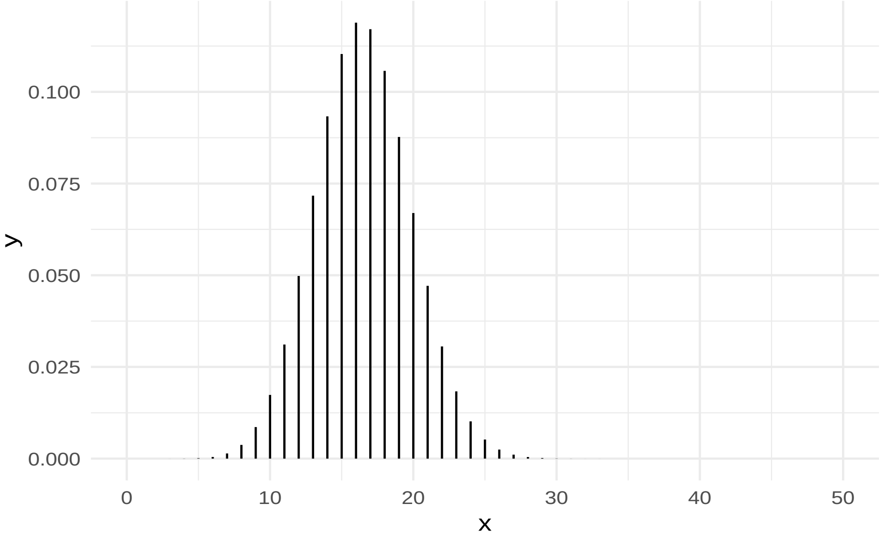
Binomial distribution

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

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

```r
x <- 0:50
prob <- dbinom(x, size = 50, prob = 0.33)
# Base R graphics
plot(x, prob, type = "h")
dat <- data.frame(x = x, y = prob)
# ggplot2 version
ggplot(dat, aes(x = x, y = y, xend = x, yend = 0)) +
  geom_segment() + theme_minimal()
```





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 height of male in cm may be within the range of [50, 300].

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

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

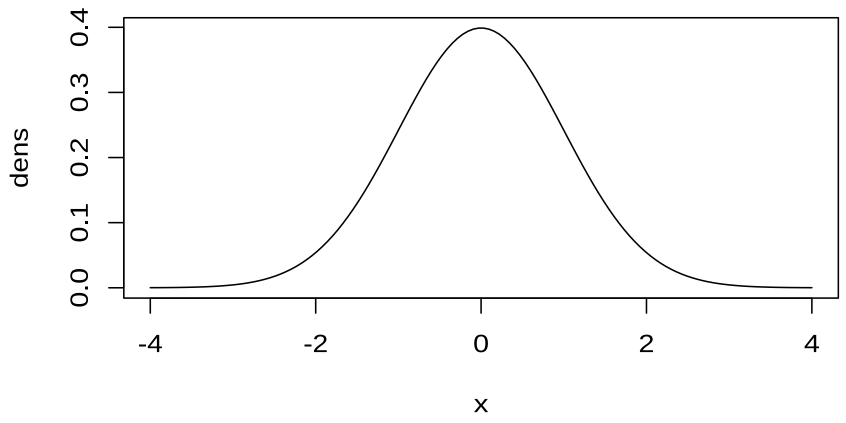
Properties:

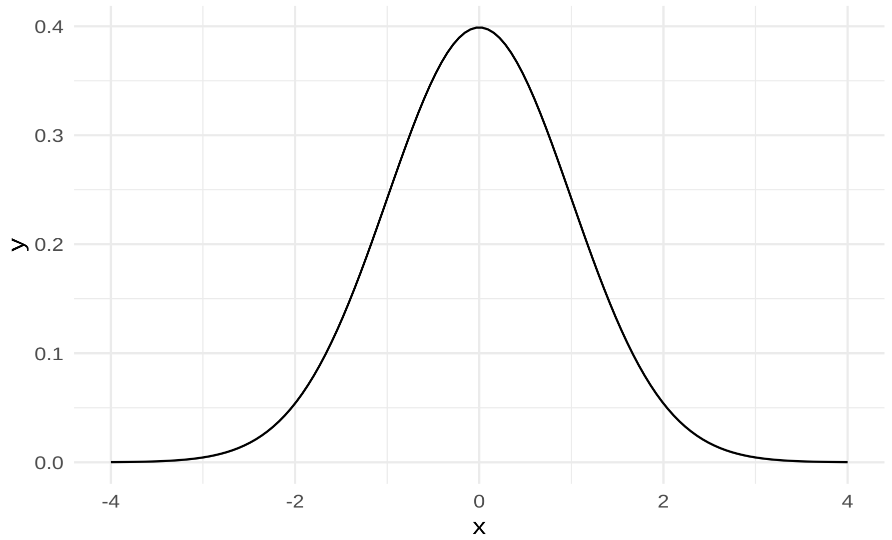
- There are an infinite (uncountable) number of possible values;
- $f(x)$ is called the density function
- $f(x) \geq 0$ (non-negative)
- $\int_{-\infty}^{\infty} f(x)\,dx = 1$ (unit measure)

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 $X \sim N(\mu, \sigma)$,

```
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 - Likelihoood approach

Density estimation

In exploratory data analysis, an estimate of the density function can be used

- 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)

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.

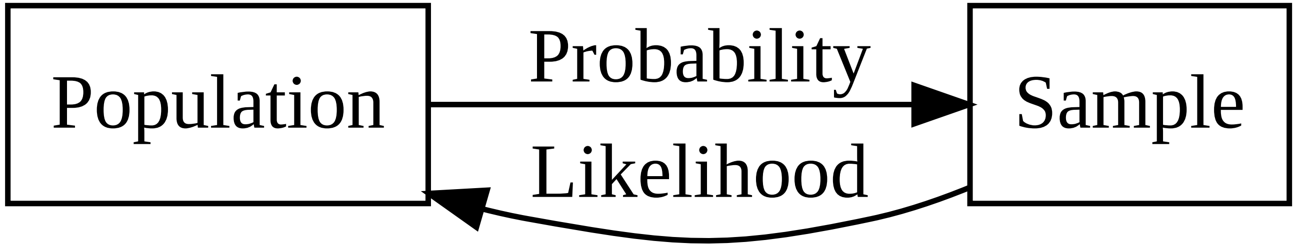
Goal: The estimation of the density function f.

Parametric density estimation

- The parametric approach to density estimation assumed a parametric model.
- That is, $X_1, X_2, \dots, X_n \overset{\text{i.i.d.}}{\sim} f_\theta$ where θ is a parameter vector.
    - For example, $\theta = (\mu, \sigma)$ when $X \sim N(\mu, \sigma)$
- Typically the parameter θ is estimated using the method of maximum likelihood.
- Density function is then estimated as $f(x|\hat{\theta})$

Maximum likelihood the best value for the parameters is the one for which the probability of obtaining the observed samples is the largest.

What is a likelihood?



Simple example:

- Population has girl:boy ratio of 2:1 (100 girls for 50 boys)
- If I draw a sample of 50 people, what is the probability of picking 10 boys
- If I draw a sample of 50 people, and picked 10 boys, what is the likelihood that the girl:boy ratio is 2:1

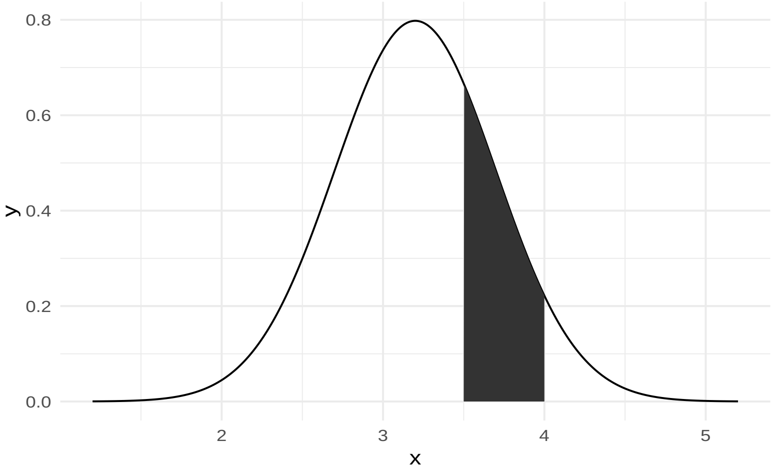
Normal distribution example

- Consider a random variable $X \sim N(3.5, 0.2)$
- 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 $X \sim N(\mu, 0.2)$, say $x = 3.7$
- Determine the likelihood of drawing this value. Flip the perspective $f(x|\theta) \rightsquigarrow L(\theta|x)$

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

```
## [1] 1.209854
```

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

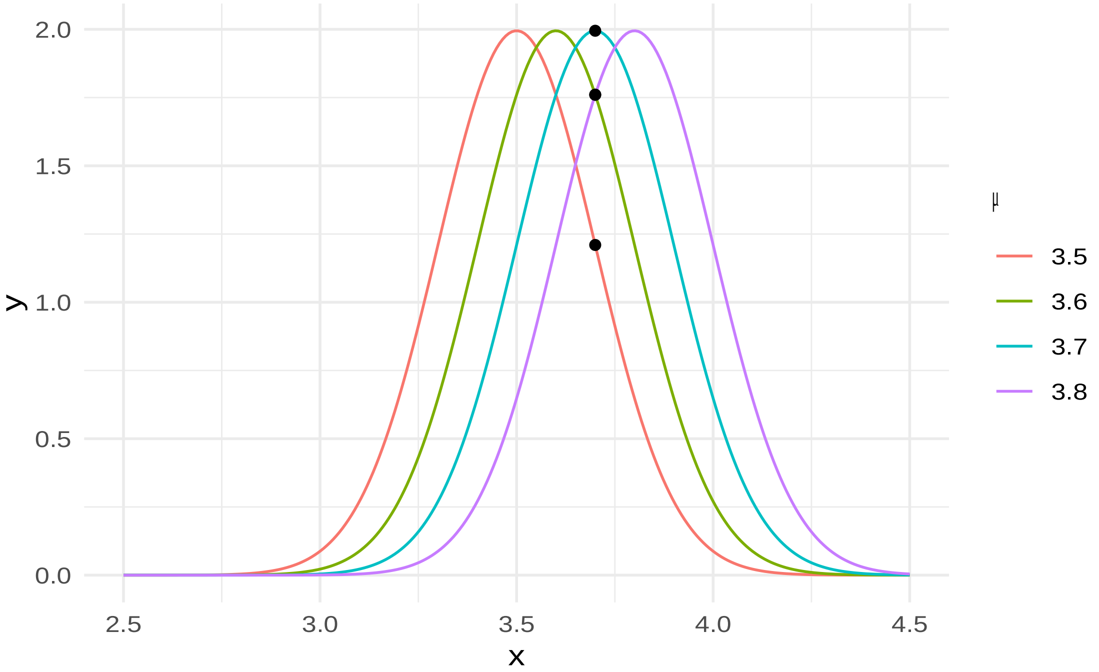
```
## [1] 1.760327
```

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

```
## [1] 1.994711
```

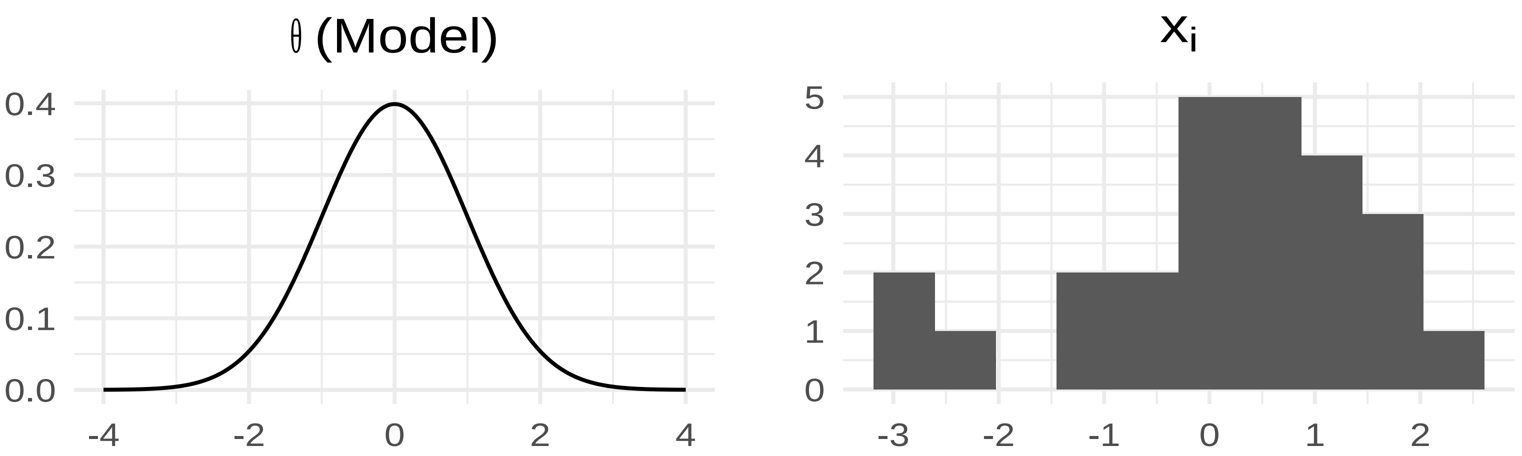
```
dnorm(3.7, mean = 3.8, sd = 0.2)
```

```
## [1] 1.760327
```



Maximum likelihood approach

- $f(x_1, x_2 \dots, x_n|\theta)$ is the probability density of observing $x_1, x_2, \dots, x_n$ given the parameter θ.



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

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

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

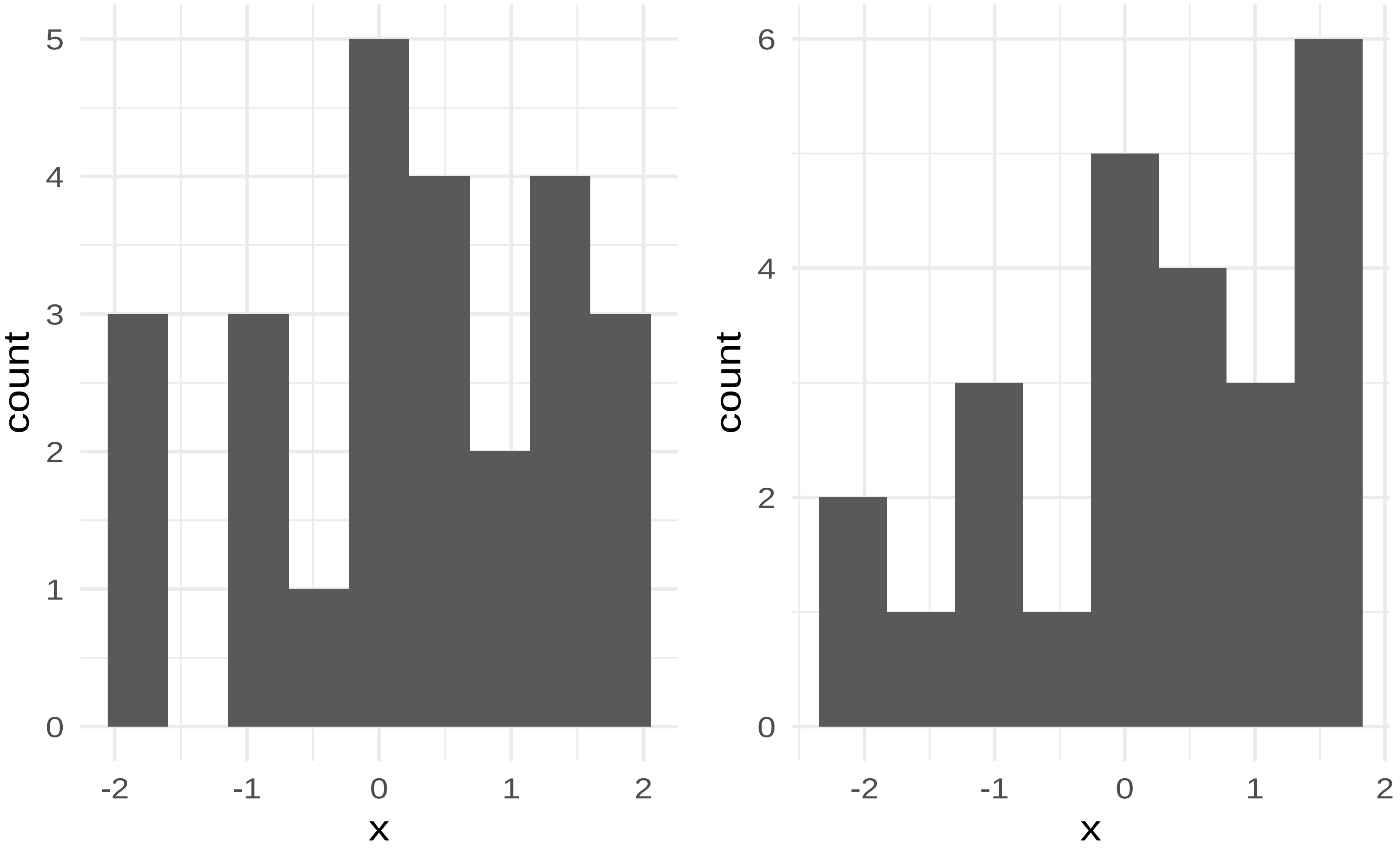
Density estimation - Non-parametric approach

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 are
  - one type of nonparametric density estimators
  - piecewise constant density estimators
  - produced automatically by most software packages

Histograms

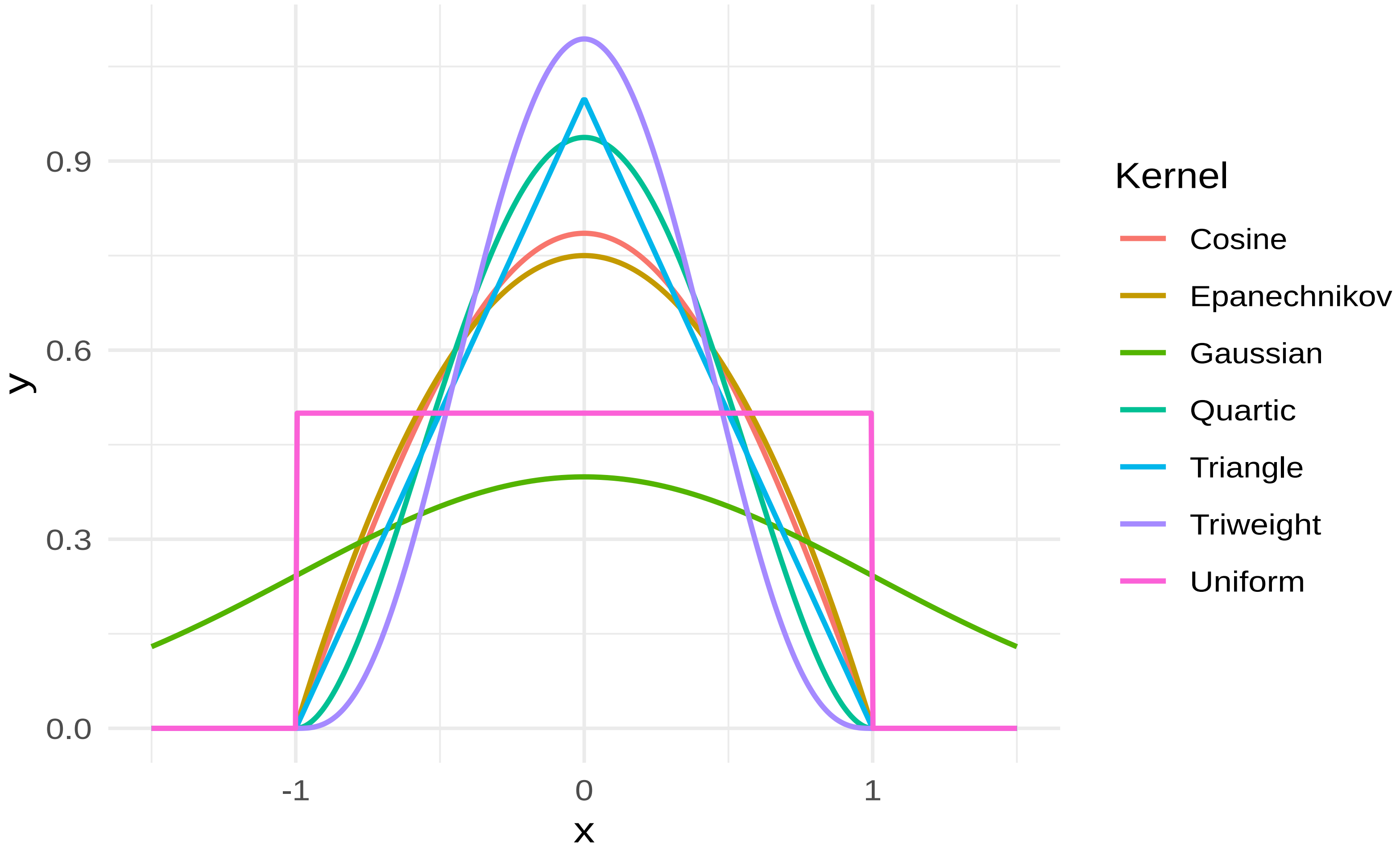
- Very simple visualization
- Sensitive to the number of bins chosen and bin width



- Preferable to have a smooth estimate and not have columns

Kernel functions

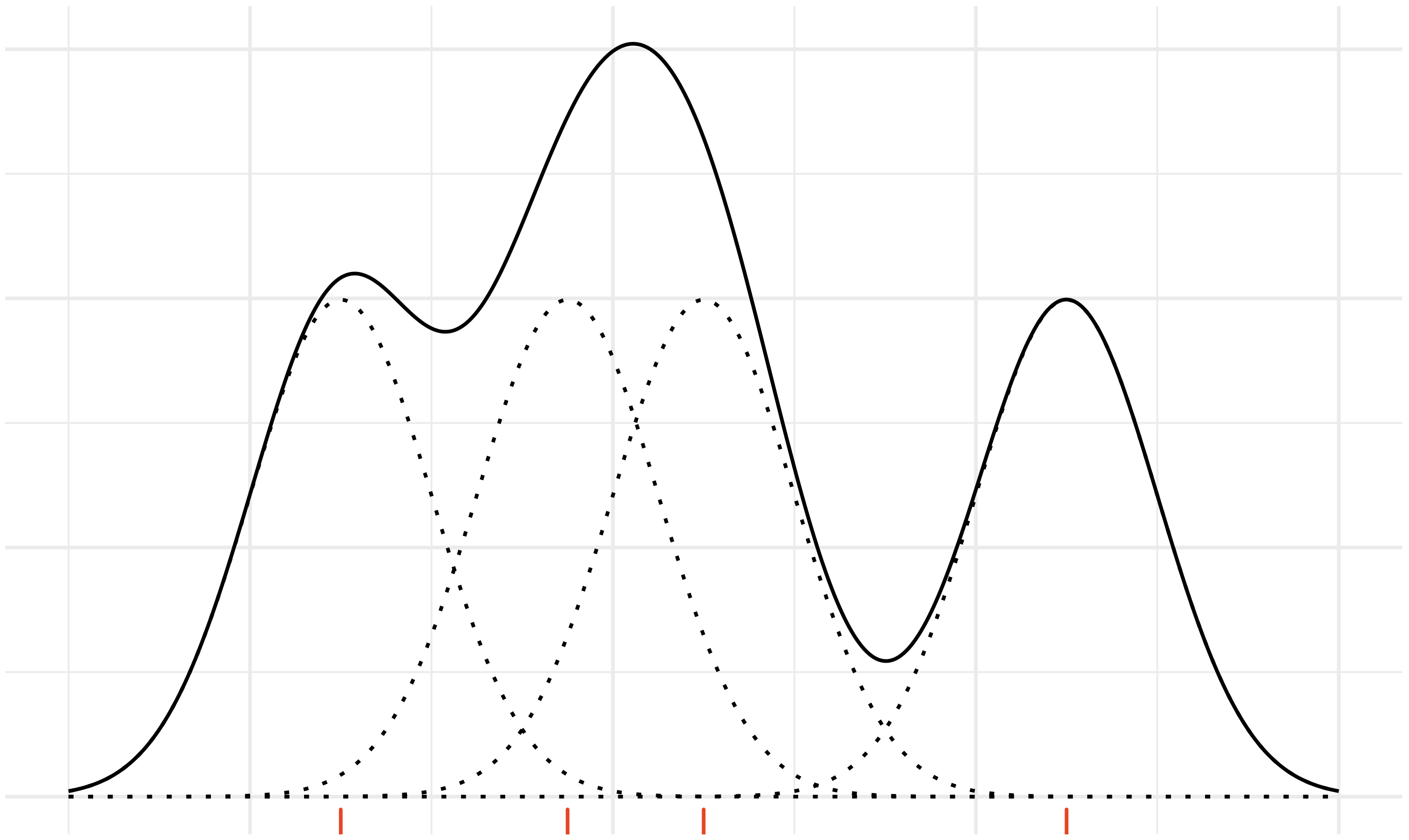
- A kernel is a special type of probability density function (PDF) having the properties.
  - non-negative $K(x) \geq 0$, symmetric $K(-x) = K(x)$, unit measure $\int K(x)\,dx = 1$



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



- E.g. 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

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

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

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

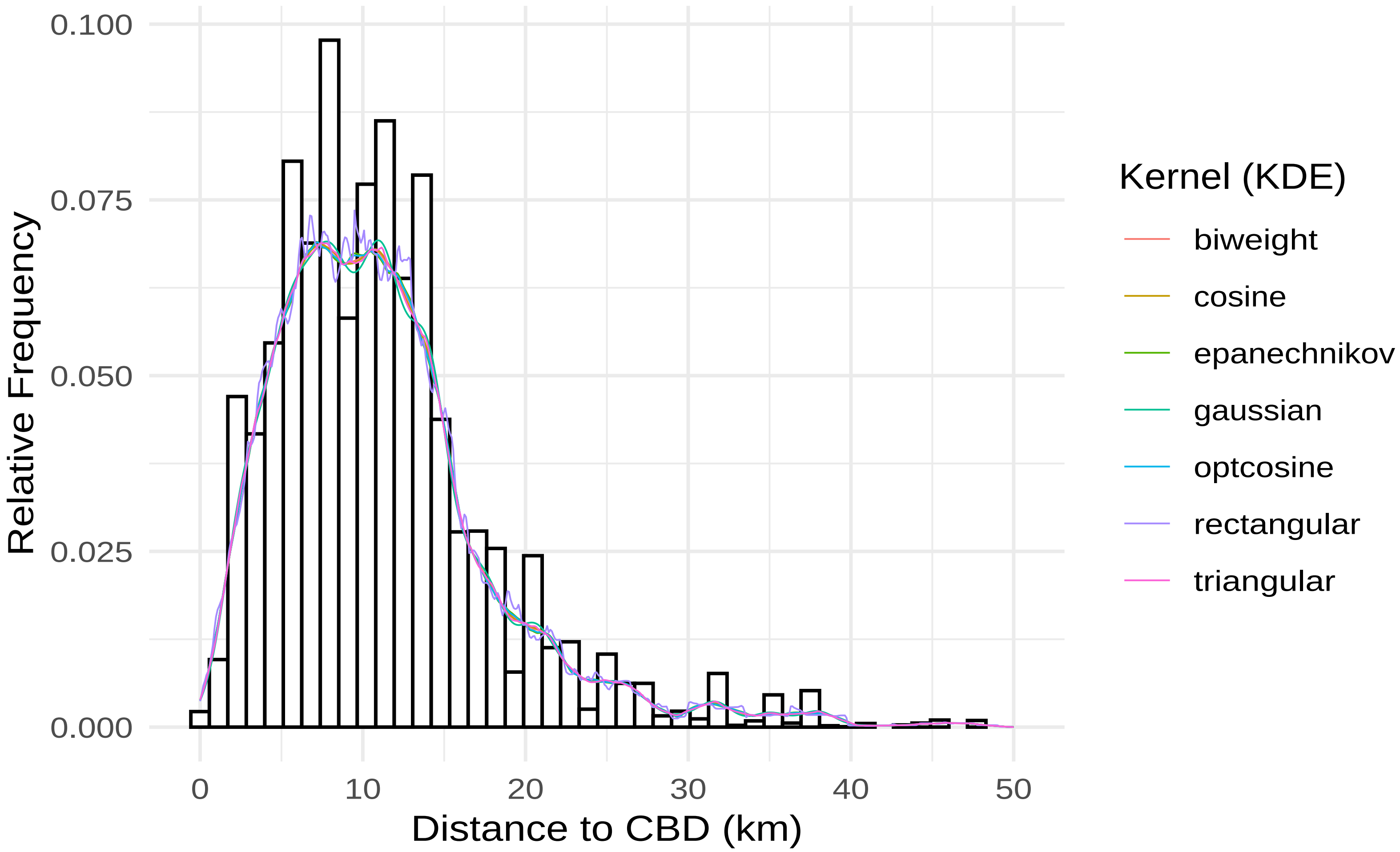
  - K is a Kernel function
  - h is a bandwidth parameter (possibly fixed or varying)
  - Consider only h fixed for this course.

Tuning the Kernel density estimator (KDE)

- There are two main components for the KDE $\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)$
    - 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

$$
\begin{aligned}
&\text{Uniform} && K(x) = \tfrac{1}{2}1_{\{|x|\le 1\}} \\
&\text{Gaussian} && K(x) = \tfrac{1}{\sqrt{2\pi}}\exp\{-\tfrac{x^2}{2}\} \\
&\text{Epanechnikov} && K(x) = \tfrac{3}{4}(1 - x^2)1_{\{|x|\le 1\}}
\end{aligned}
$$

Different choices of Kernel function with same bandwidth



Computing density in R

- Base R there is `density`
  - `density` computes the KDE
  - Can wrap in `plot`, i.e. `plot(density(x))`, to visualize
  - Can inspect details in `summary`
- For plotting ggplot there is `geom_density`
  - Can specify the bandwidth with `bw` argument

Choosing the bandwidth

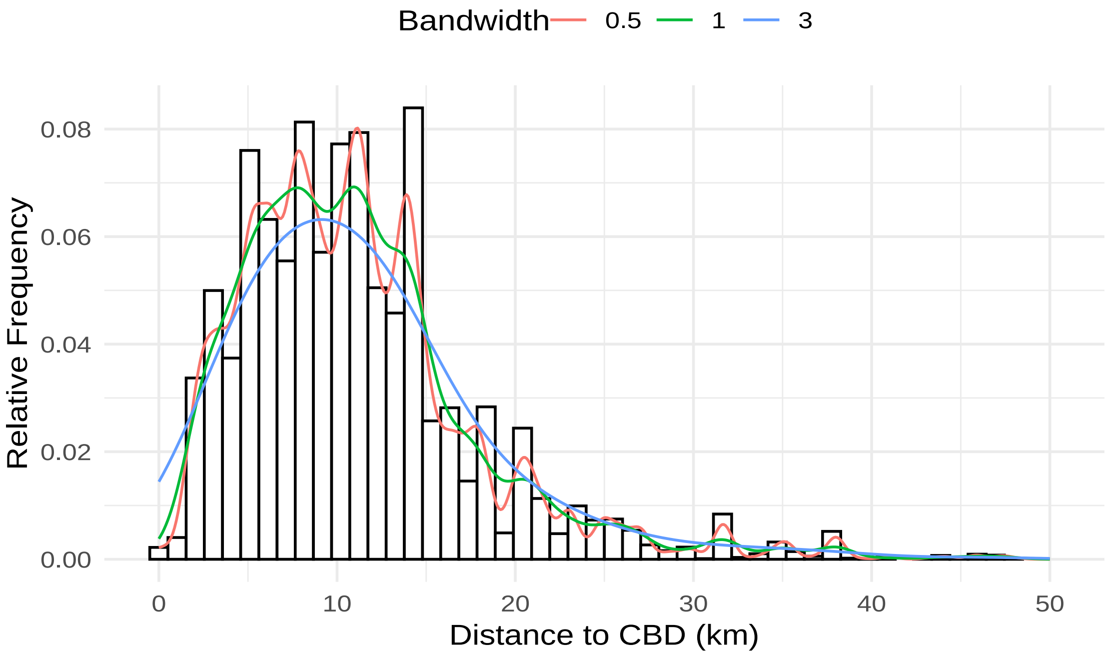
- The density estimator

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

  - 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

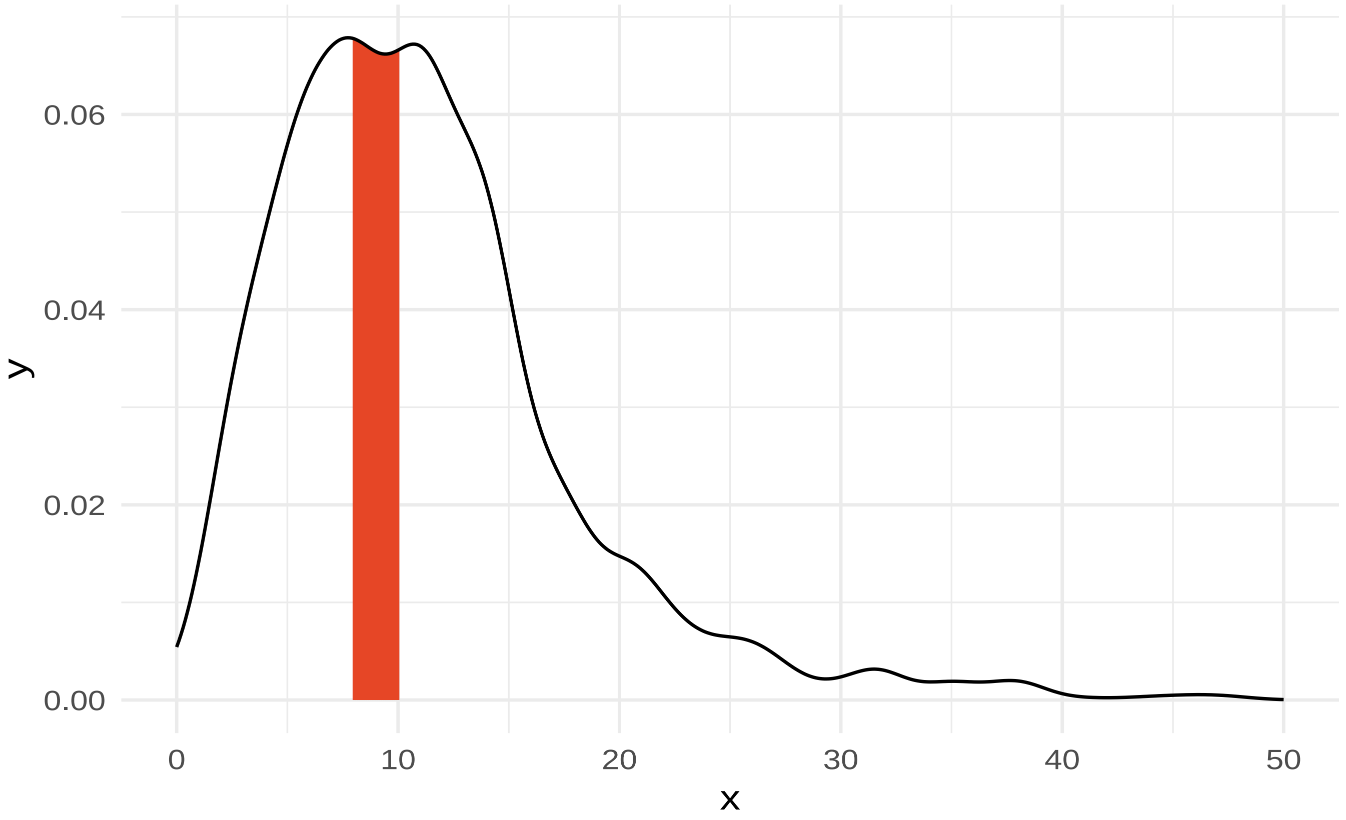
- Consider the distance from CBD variable again with three bandwidths



- 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 8-10km of CBD
- Integrate the density function between 8 and 10 yields $p = 0.13 \rightsquigarrow$ 13% chance of finding a property between 8-10km of CBD



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.

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

- 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.

Kernel density estimation type equivalent

$$Var(\hat{f}(x)) = O\left(\frac{1}{nh}\right)$$
$$Bias(\hat{f}(x)) = O(h)$$

References

James, G., D. Witten, T. Hastie, et al. (2013). *An introduction to statistical learning*. Vol. 112. Springer.