University of BRISTOL

Exploratory data analysis

Fundamental tools for describing and summarizing data.

Henry W J Reeve

henry.reeve@bristol.ac.uk

Statistical Computing & Empirical Methods (EMATM0061)

MSc in Data Science, Teaching block 1, 2021.

What will we cover today?

- We will give a taxonomy of the basic data types.
- We will explore methods for estimating the overall location of a feature in a data set.
- We will explore methods for estimating overall variability of a feature in a data set.
- We will explore methods for estimating how connected two features in a data set are.

A taxonomy of data types

Continuous Data that can take any value on an interval – e.g. height of a person in mm

Discrete Data with a minimum distance between possible values – e.g. number of restaurant meals in a month

Categorical Data that can take on only a specific set of values representing distinct categories e.g. brand.

Binary Categorical data with exactly two categories e.g. pass or fail a driving test.

Ordinal Categorical data with an ordering e.g. “How was your meal?” on a Likert scale.

Exploratory data analysis

Let’s suppose we want to begin exploring a data set …. also known as a data sample.

We begin by understanding the meaning and data type of each of the variables aka. features.

We can use visualisation to efficiently identify the shape of distributions and key relationships.

We can learn more by computing **statistics.**

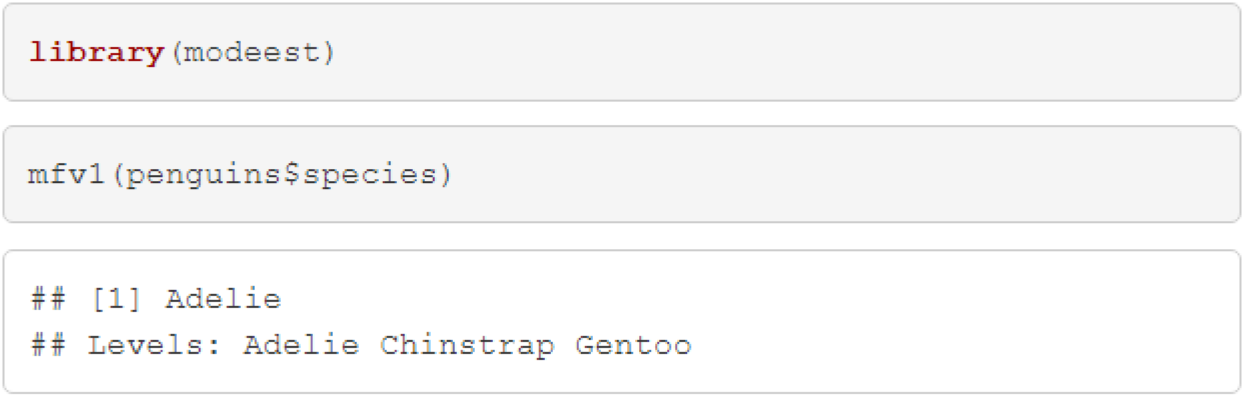
Statistics are functions of the data intended to provide useful information.

Estimates of location for categorical data

Which single value is most representative or typical?

For categorical data the natural answer is the sample mode.

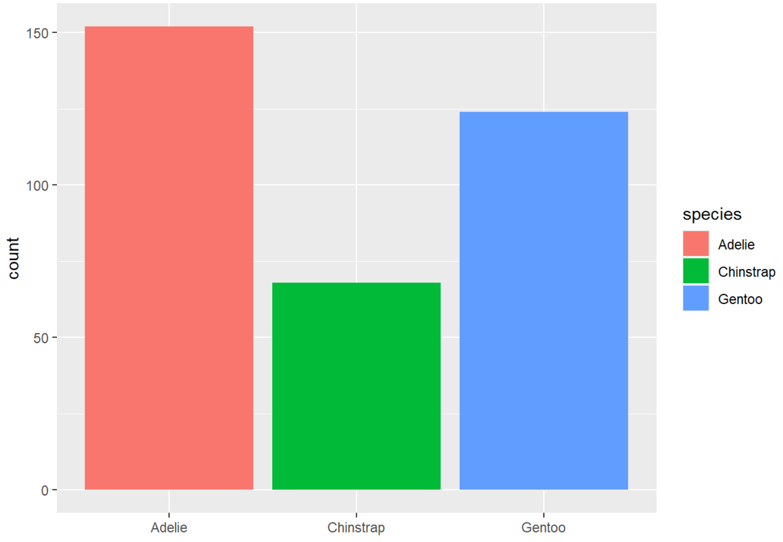
The sample mode is the value which occurs with the highest frequency for a feature within a data set.



```r
library(modeest)
```

```r
mfv1(penguins$species)
```

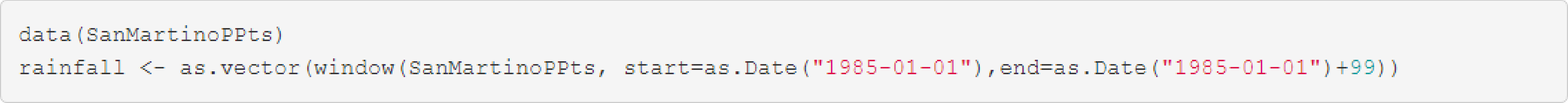
```
## [1] Adelie
## Levels: Adelie Chinstrap Gentoo
```



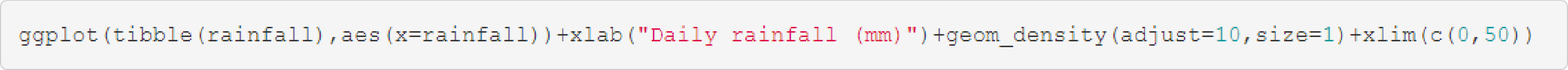
Note: The inbuilt "mode" function determines the type of an object.

Estimates of location for continuous data

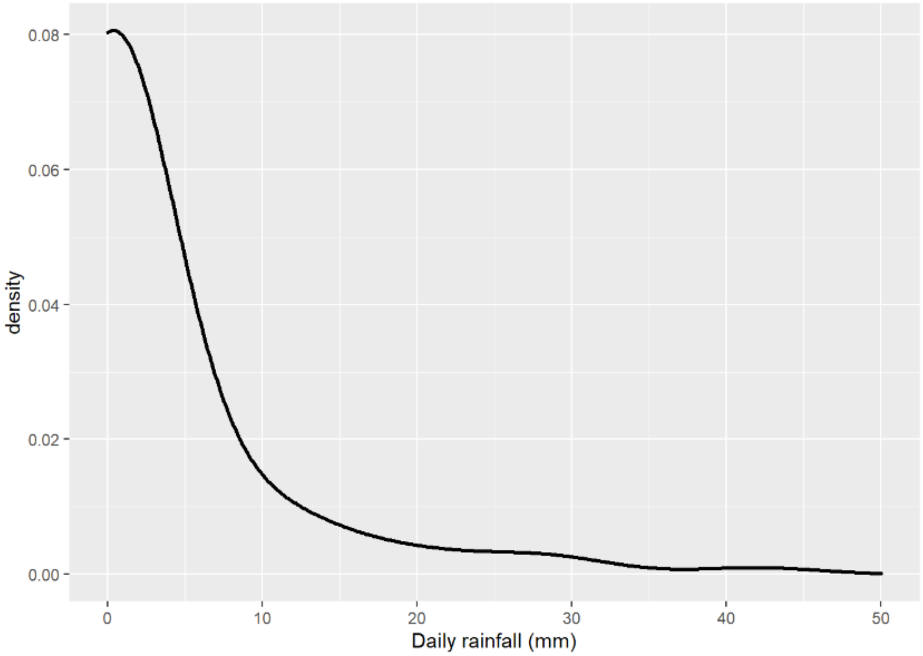
Let’s look at rainfall data for San Martino for the first 100 days of 1985 using the hydroTSM library.



```r
data(SanMartinoPPts)
rainfall <- as.vector(window(SanMartinoPPts, start=as.Date("1985-01-01"),end=as.Date("1985-01-01")+99))
```

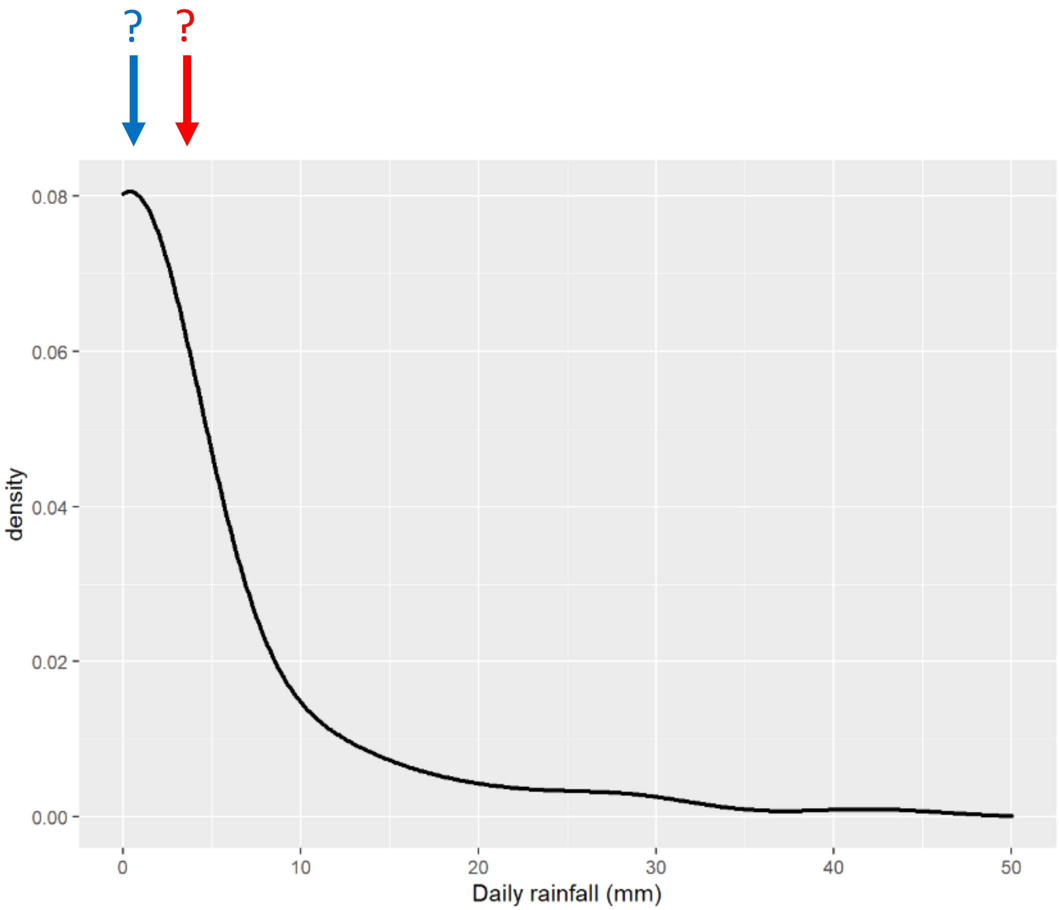


```r
ggplot(tibble(rainfall),aes(x=rainfall))+xlab("Daily rainfall (mm)")+geom_density(adjust=10,size=1)+xlim(c(0,50))
```



Estimates of location for continuous data

What single number best represents “typical rainfall”?

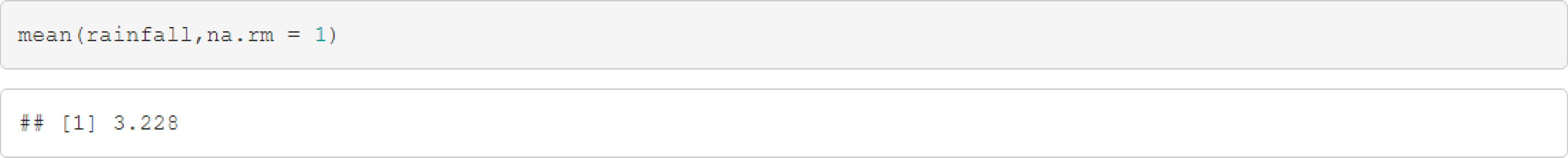


The sample mean

The most well known estimate of location is the sample mean (the arithmetic mean):

Suppose our variable of interest has values $X_1, \cdots, X_n$

$$\text{Sample mean} := \frac{1}{n} \cdot (X_1 + \cdots + X_n)$$

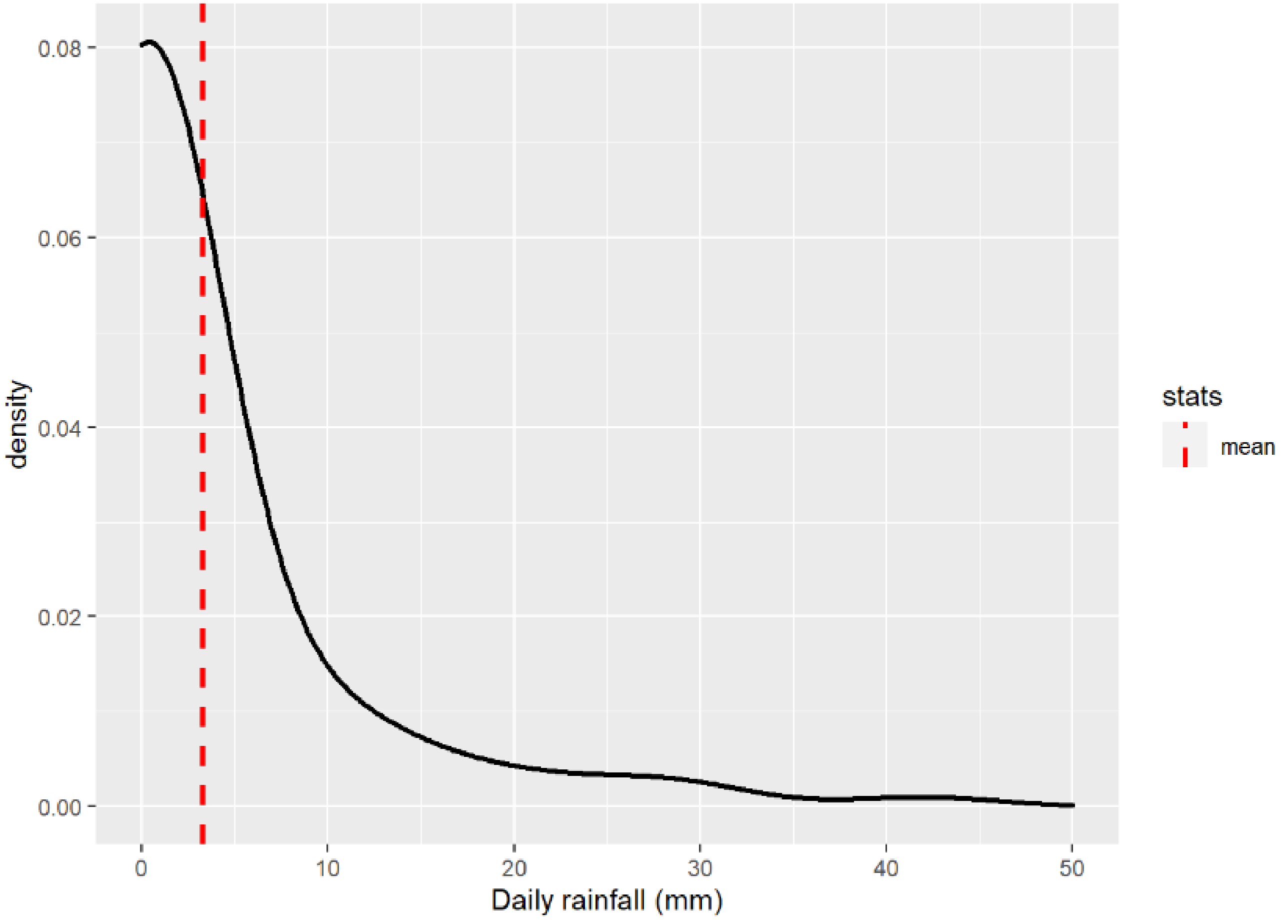


```
mean(rainfall,na.rm = 1)
```

```
## [1] 3.228
```

The sample mean

The most well known estimate of location is the sample mean:



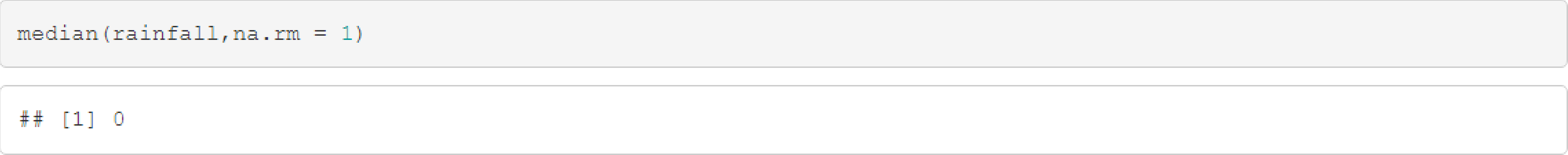
The sample median

The sample median is the middle value after sorting the values by numerical order.

Starting with $X_1, \cdots, X_n$ we apply a sorting algorithm to obtain:

$$X^{(1)} \leq X^{(2)} \leq \cdots \leq X^{(n)}$$

$$\text{Sample median} := \frac{1}{2} \cdot (X^{(\lfloor n/2 \rfloor)} + X^{(\lceil n/2 \rceil)})$$

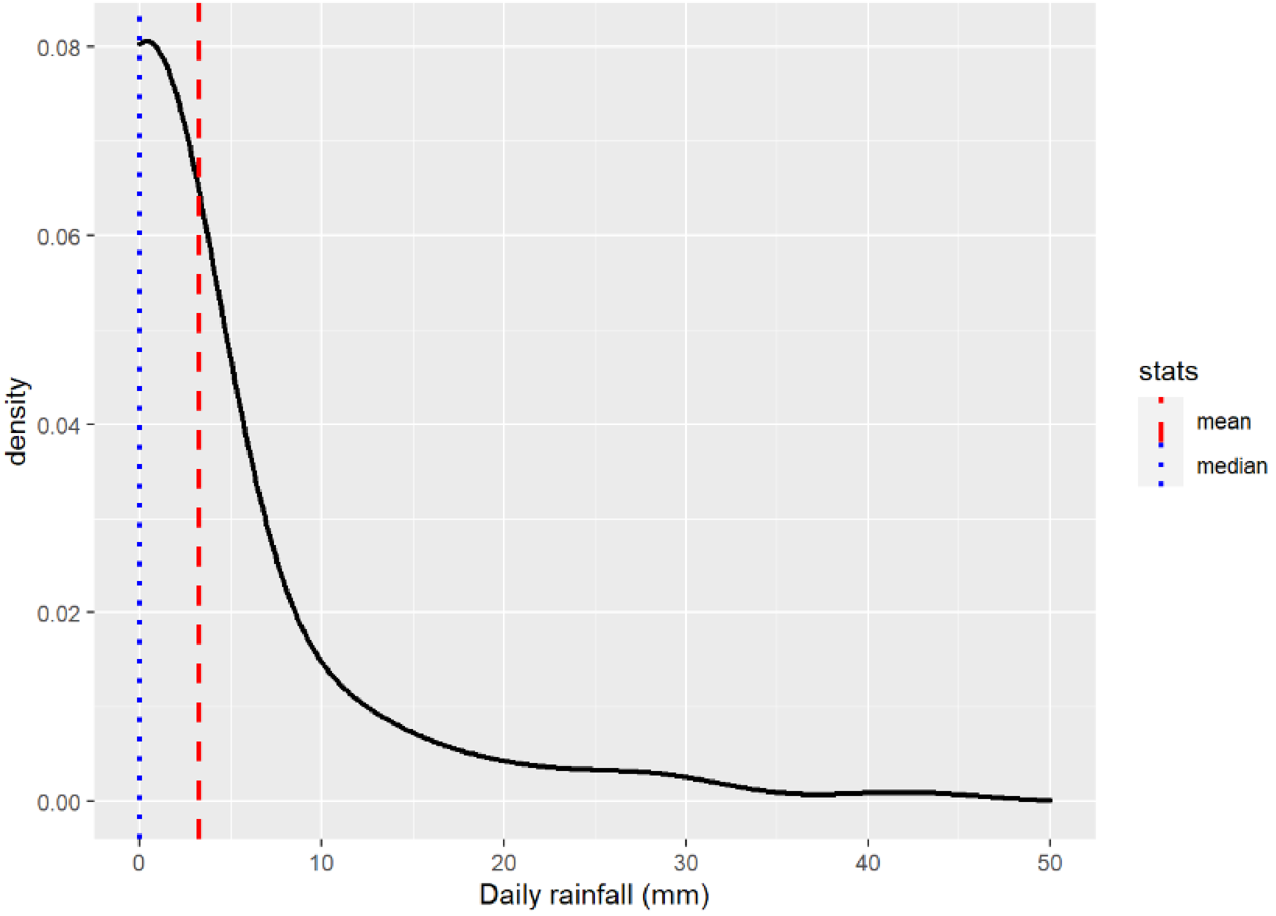


```
median(rainfall,na.rm = 1)
```

```
## [1] 0
```

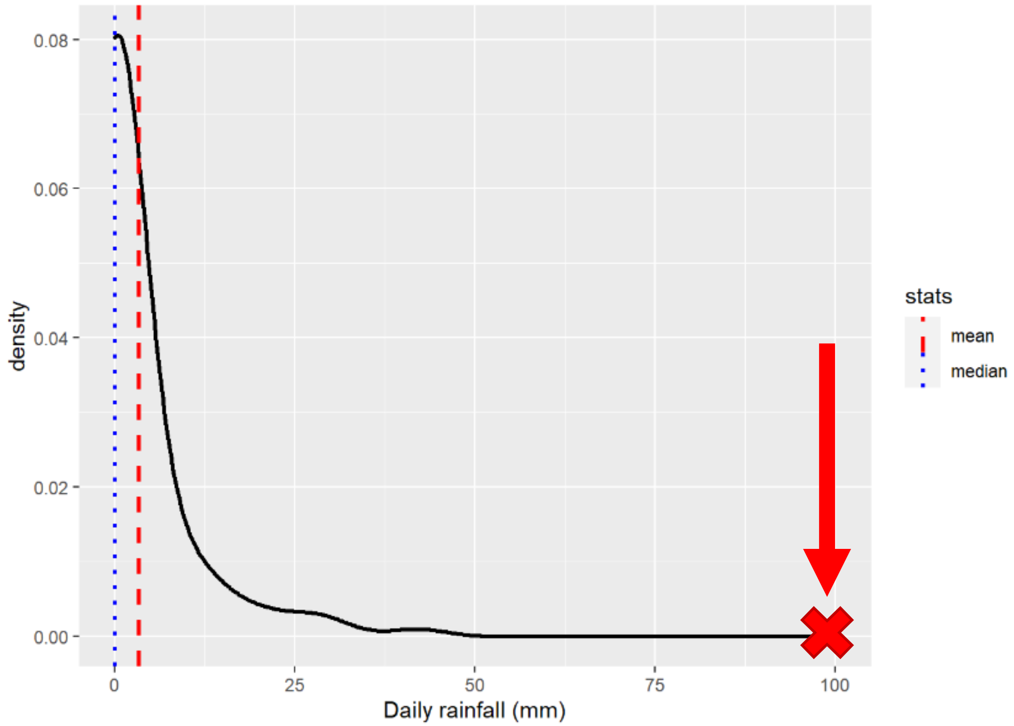
The sample median

The sample median is the middle value after sorting the values by numerical order.



Outliers

An outlier is a value in a data set which differs substantially from other values.



There is no standard definition! – can be related to distance from the median or mean.

Outliers

There are two different types of outliers we can encounter in practice:

1. A faithful representation of genuinely anomalous event

   e.g. A day of extremely unusual torrential rainfall.

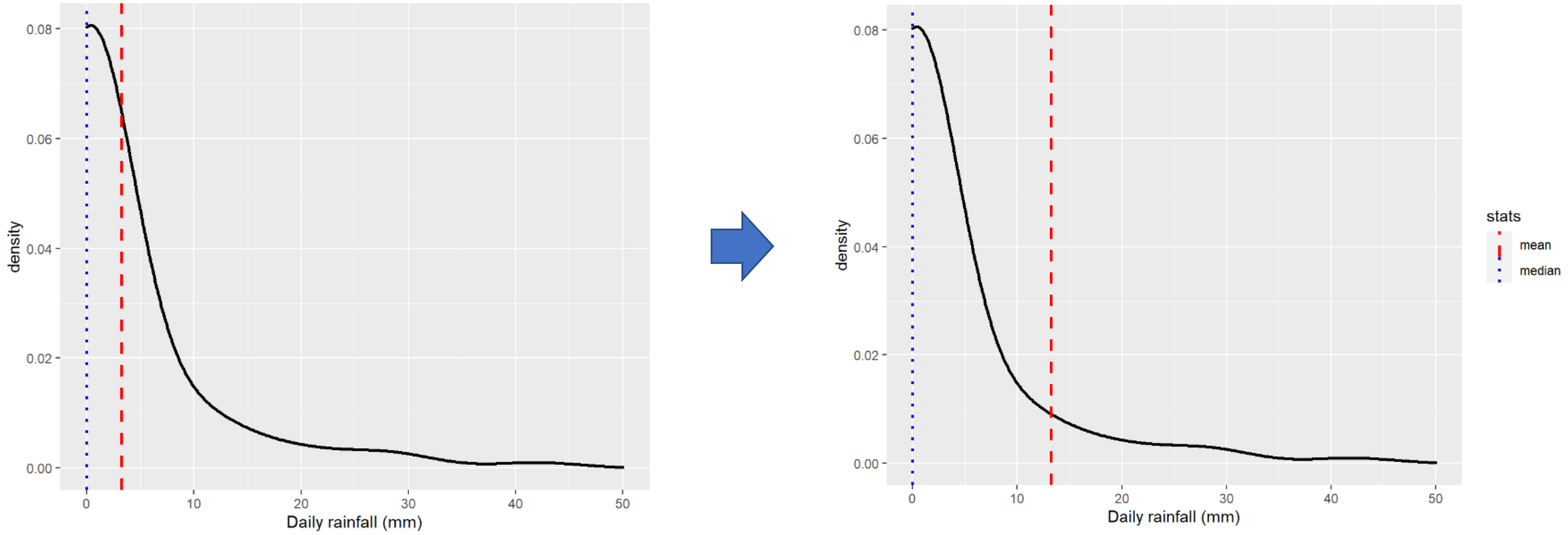
2. An error in the data resulting from problems in measurement, recording etc.

Robustness of the sample median

A major advantage of the median over the mean is that it is robust to small corruptions in the data set.

Uncorrupted data

A single corrupted value



Comparing the sample median and the sample mean

✓ The sample median is robust to small corruptions in the data set, unlike the mean.

✗ The sample median effectively ignores a large section of the data set, unlike the mean.

This makes it difficult to aggregated medians from multiple sources.

Example: The sample median might do a poor job of distinguishing regions with very different rainfall levels.

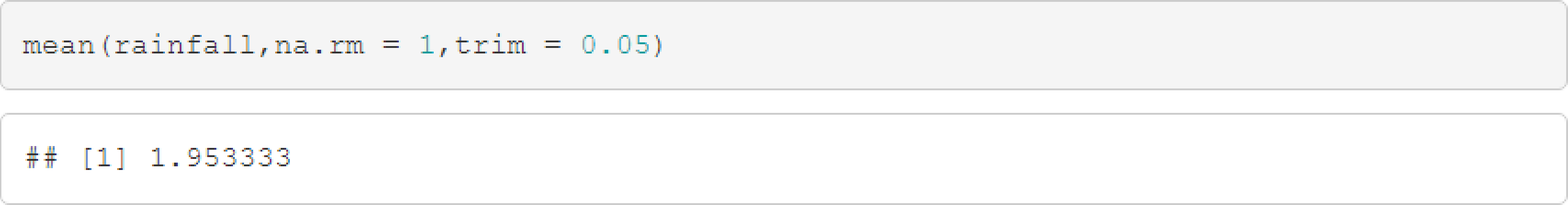
The trimmed sample mean

The trimmed sample mean is the mean computed after removing a prescribed fraction of the data.

The trimmed sample mean with trim fraction $q \in (0, 1/2]$ is computed as follows:

Starting with $X_1, \cdots, X_n$ we apply a sorting algorithm to obtain $X^{(1)} \leq X^{(2)} \leq \cdots \leq X^{(n)}$.

$$\text{Trimmed sample mean} := \frac{1}{n - 2 \cdot \lfloor q \cdot n \rfloor} \cdot \sum_{i=\lfloor q \cdot n \rfloor + 1}^{n - \lfloor q \cdot n \rfloor + 1} X^{(i)}$$



```
mean(rainfall,na.rm = 1,trim = 0.05)
```

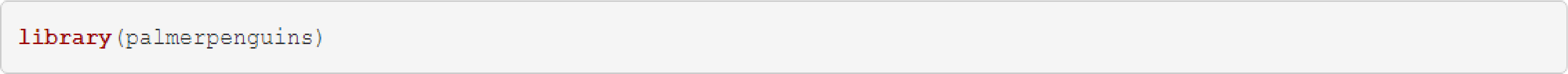
```
## [1] 1.953333
```

The trimmed sample mean is more robust to outliers than the mean but more sensitive than the median.

Estimates of location with penguins data

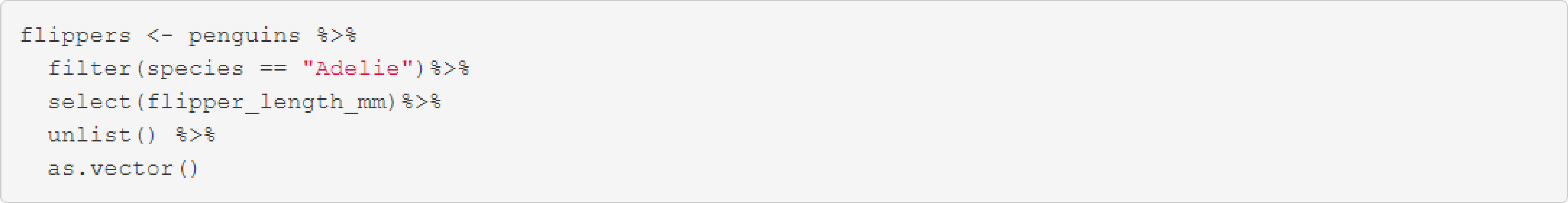
We use of the Palmer penguins data set introduced by Alison Hill, Allison Horst, Kristen Gorman.

First load the palmer penguins library.

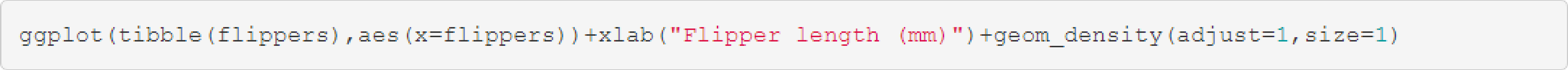


```r
library(palmerpenguins)
```

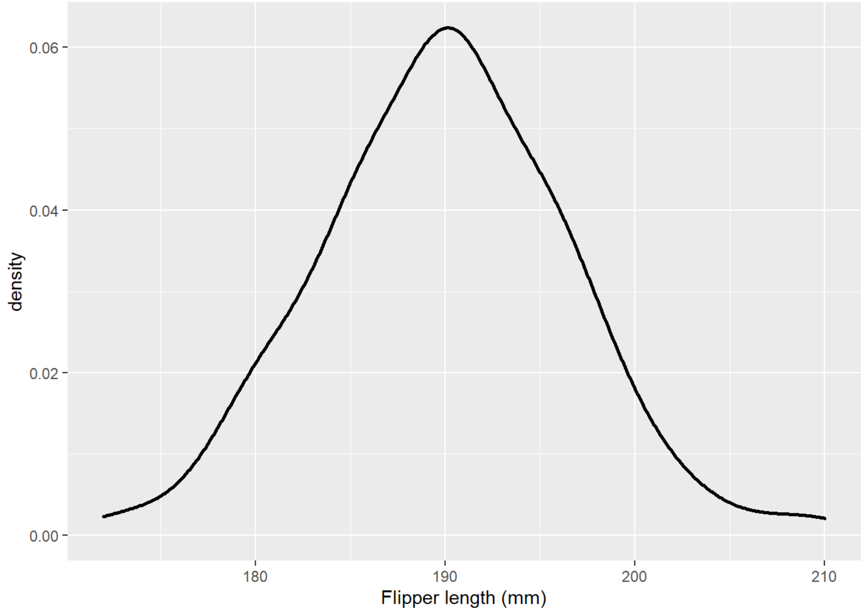
Estimates of location



```
flippers <- penguins %>%
  filter(species == "Adelie")%>%
  select(flipper_length_mm)%>%
  unlist() %>%
  as.vector()
```

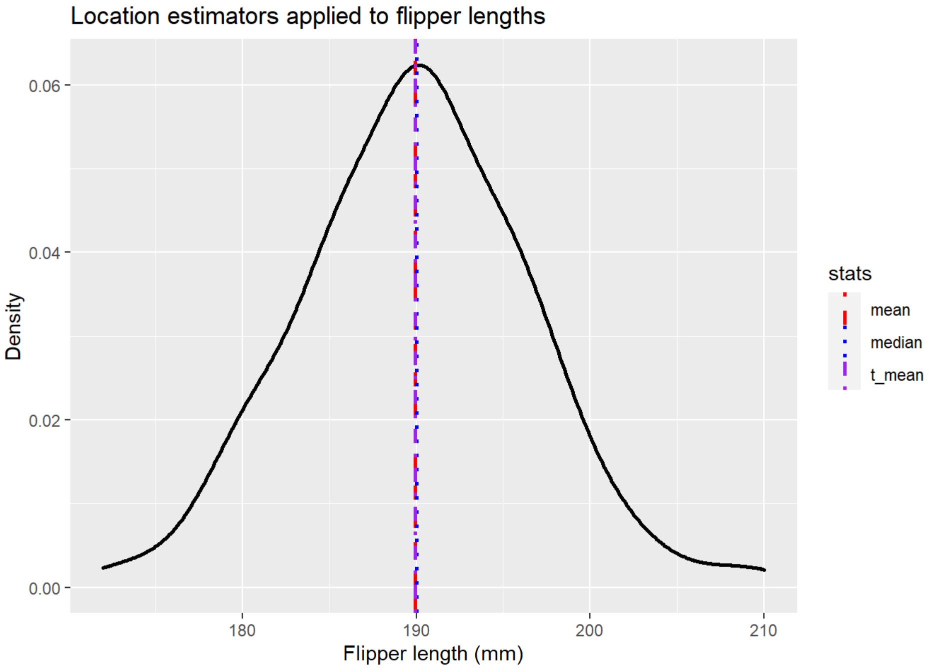


```
ggplot(tibble(flippers),aes(x=flippers))+xlab("Flipper length (mm)")+geom_density(adjust=1,size=1)
```

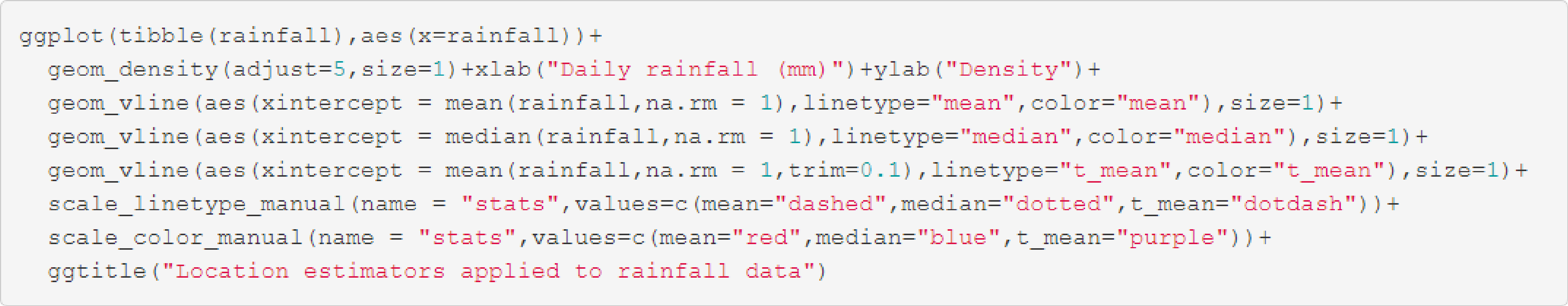


Estimates of location

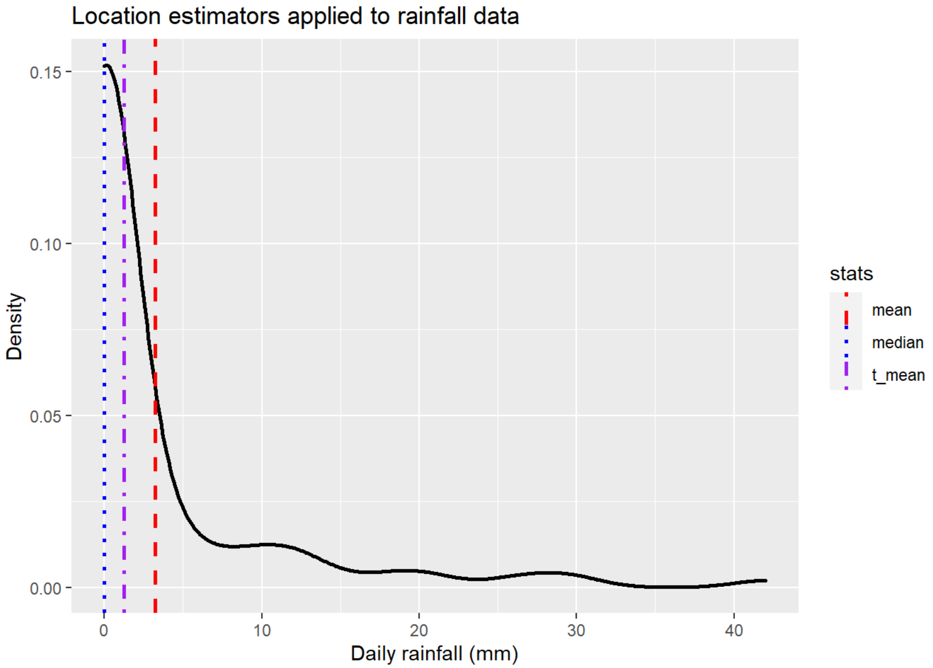
```
ggplot(tibble(flippers),aes(x=flippers))+
  geom_density(adjust=1,size=1)+xlab("Flipper length (mm)")+ylab("Density")+
  geom_vline(aes(xintercept = mean(flippers,na.rm = 1),linetype="mean",color="mean"),size=1)+
  geom_vline(aes(xintercept = median(flippers,na.rm = 1),linetype="median",color="median"),size=1)+
  geom_vline(aes(xintercept = mean(flippers,na.rm = 1,trim=0.1),linetype="t_mean",color="t_mean"),size=1)+
  scale_linetype_manual(name = "stats",values=c(mean="dashed",median="dotted",t_mean="dotdash"))+
  scale_color_manual(name = "stats",values=c(mean="red",median="blue",t_mean="purple"))+
  ggtitle("Location estimators applied to flipper lengths")
```



Estimates of location



```
ggplot(tibble(rainfall),aes(x=rainfall))+
  geom_density(adjust=5,size=1)+xlab("Daily rainfall (mm)")+ylab("Density")+
  geom_vline(aes(xintercept = mean(rainfall,na.rm = 1),linetype="mean",color="mean"),size=1)+
  geom_vline(aes(xintercept = median(rainfall,na.rm = 1),linetype="median",color="median"),size=1)+
  geom_vline(aes(xintercept = mean(rainfall,na.rm = 1,trim=0.1),linetype="t_mean",color="t_mean"),size=1)+
  scale_linetype_manual(name = "stats",values=c(mean="dashed",median="dotted",t_mean="dotdash"))+
  scale_color_manual(name = "stats",values=c(mean="red",median="blue",t_mean="purple"))+
  ggtitle("Location estimators applied to rainfall data")
```

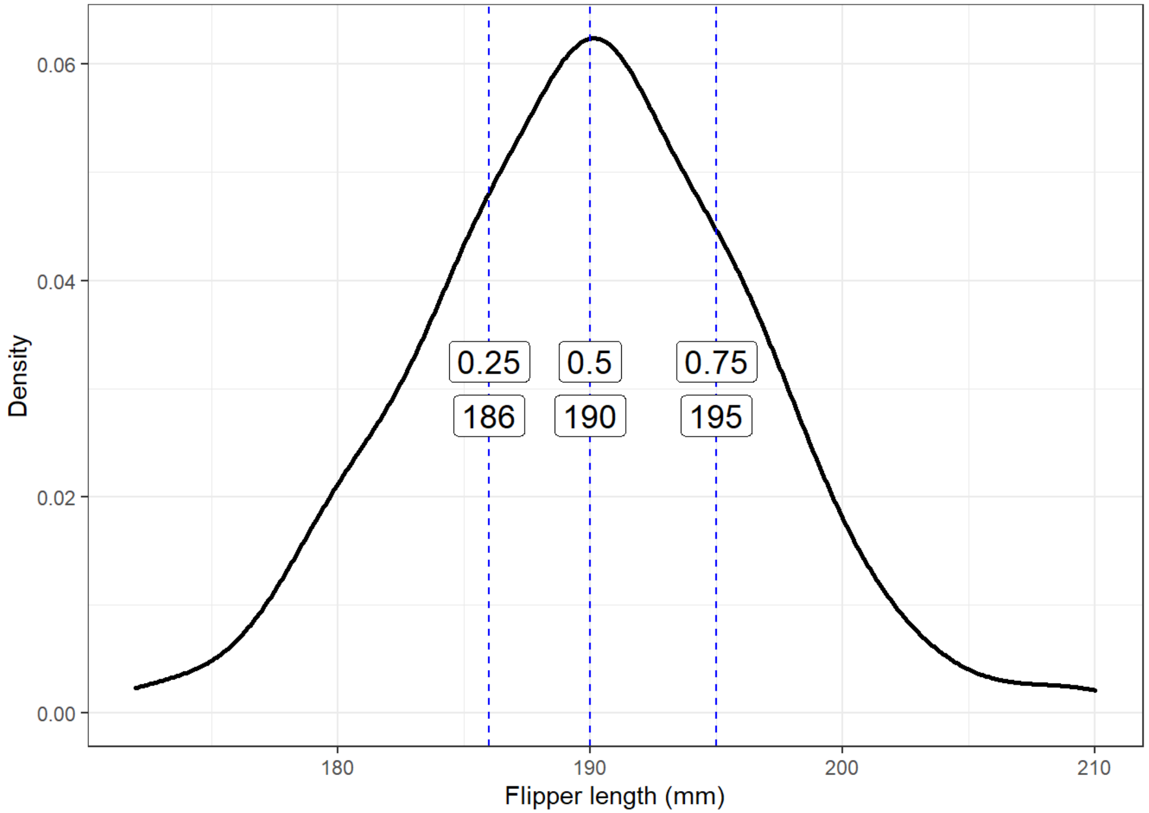


Now take a break!

Sample quantiles and sample percentiles

A sample median can be seen as a point which is halfway through your data set.

Sample quantiles extend this notion to other fractions e.g. which point is a 1/4 of the way through your data?

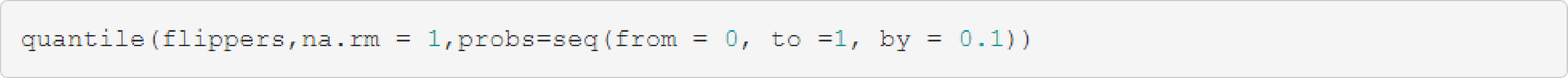


Sample quantiles and sample percentiles

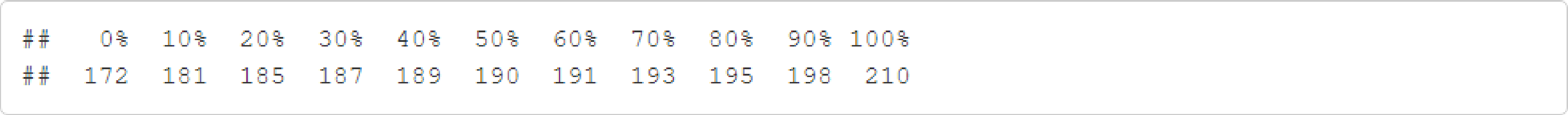
Given $X_1, \cdots, X_n$ and $q \in (0,1]$ the corresponding sample q - quantile is of the following form:

First sort the data to obtain $X^{(1)} \leq X^{(2)} \leq \cdots \leq X^{(n)}$ then take

$$X^{(\max\{\lfloor qn \rfloor, 1\})} \leq \text{ sample q-quantile } < X^{(\lceil qn \rceil)}$$



```r
quantile(flippers,na.rm = 1,probs=seq(from = 0, to =1, by = 0.1))
```



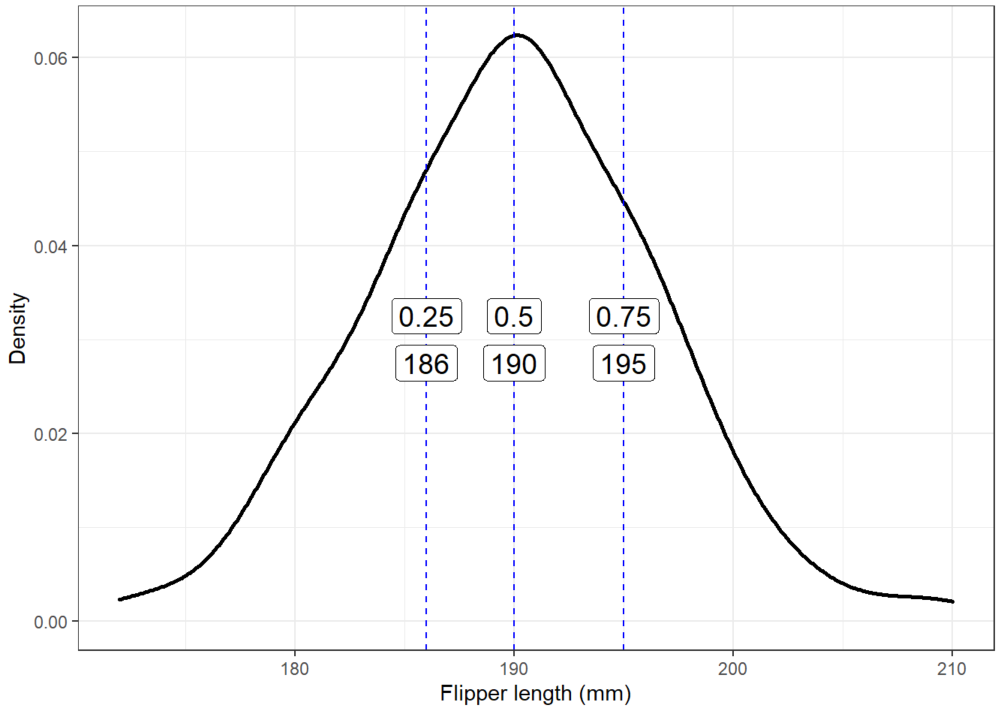
```
##    0%  10%  20%  30%  40%  50%  60%  70%  80%  90% 100%
##   172  181  185  187  189  190  191  193  195  198  210
```

The sample $100 \cdot q$ -th percentile is precisely the same as the sample q - quantile.

Sample quantiles and sample percentiles

```
probabilities<-c(0.25,0.5,0.75)
quantiles<-quantile(flippers,probs=probabilities,na.rm=1)
```

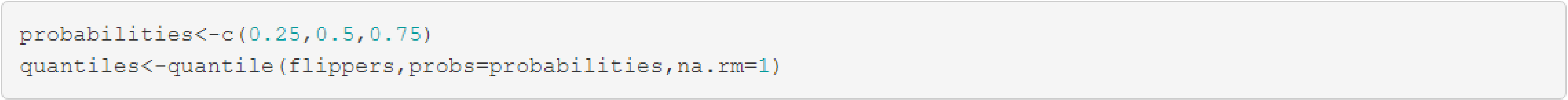
```
ggplot(tibble(flippers),aes(x=flippers))+theme_bw()+
  geom_density(adjust=1,size=1)+xlab("Flipper length (mm)")+ylab("Density")+
  geom_vline(xintercept = quantiles,linetype="dashed",color="blue")+
  annotate("label", x = quantiles, y = 0.0325,size=5, fill="white", label = probabilities)+
  annotate("label", x = quantiles, y = 0.0275,size=5, fill="white", label = quantiles)
```



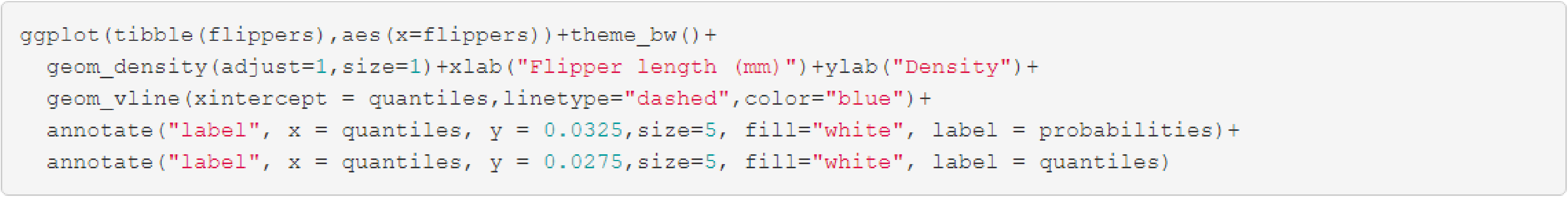
```
quantile(flippers,probs=c(0.25,0.5,0.75),na.rm=1)
```

```
## 25% 50% 75%
## 186 190 195
```

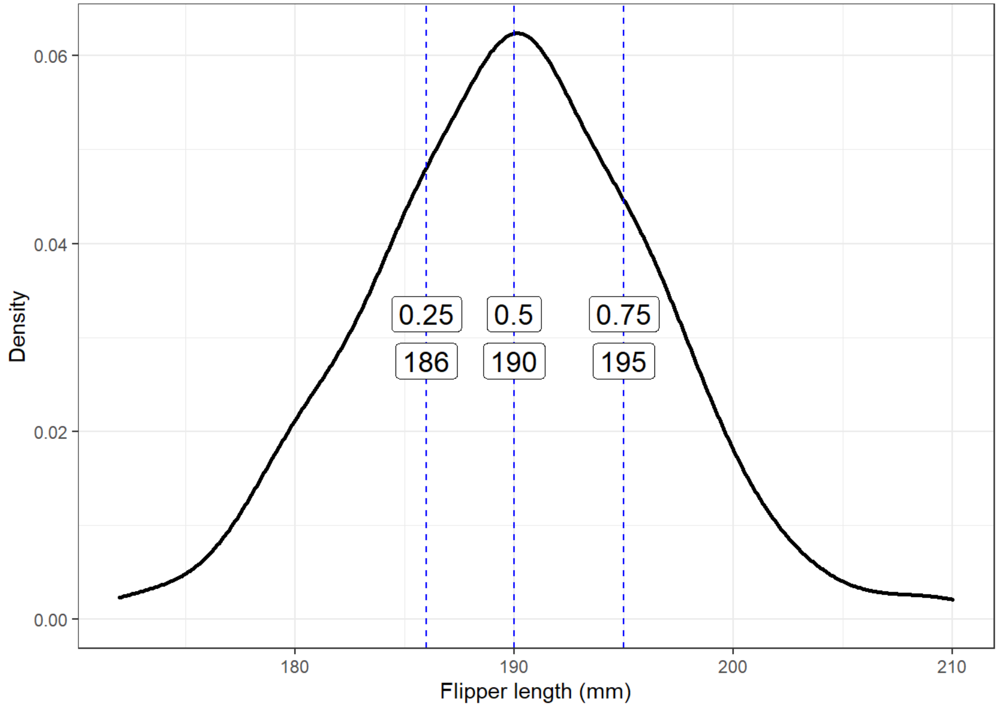
Sample quantiles and sample percentiles



```
probabilities<-c(0.25,0.5,0.75)
quantiles<-quantile(flippers,probs=probabilities,na.rm=1)
```



```
ggplot(tibble(flippers),aes(x=flippers))+theme_bw()+
  geom_density(adjust=1,size=1)+xlab("Flipper length (mm)")+ylab("Density")+
  geom_vline(xintercept = quantiles,linetype="dashed",color="blue")+
  annotate("label", x = quantiles, y = 0.0325,size=5, fill="white", label = probabilities)+
  annotate("label", x = quantiles, y = 0.0275,size=5, fill="white", label = quantiles)
```



0.25-quantile = 25th -percentile = 1st quartile

0.50-quantile = 50th -percentile = 2nd quartile = median

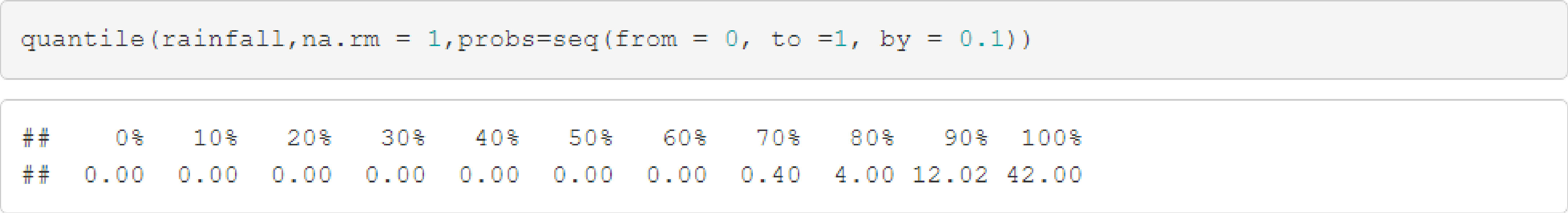
0.75-quantile = 75th -percentile = 3rd quartile.

Sample quantiles and sample percentiles

Given $X_1, \cdots, X_n$ and $q \in (0,1]$ the corresponding sample q - quantile is of the following form:

First sort the data to obtain $X^{(1)} \leq X^{(2)} \leq \cdots \leq X^{(n)}$ then take

$$X^{(\max\{\lfloor qn \rfloor, 1\})} \leq \text{ q-quantile } < X^{(\lceil qn \rceil)}$$



```
quantile(rainfall,na.rm = 1,probs=seq(from = 0, to =1, by = 0.1))
```

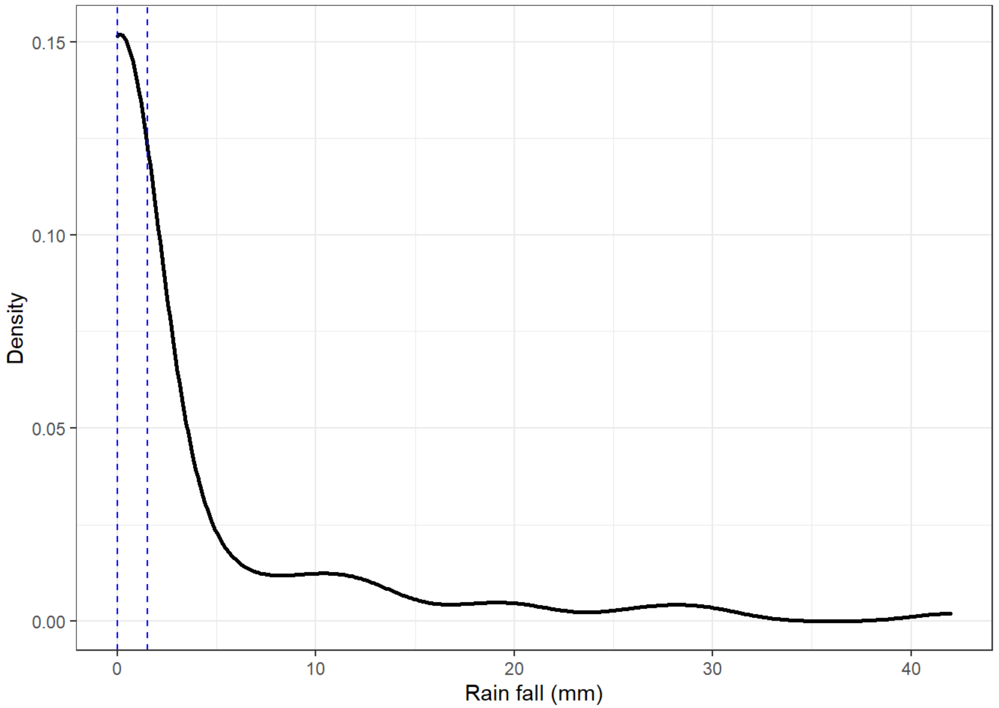
```
##    0%   10%   20%   30%   40%   50%   60%   70%   80%   90%  100% 
##  0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.40  4.00 12.02 42.00
```

The $100 \cdot q$ -th percentile is precisely the same as the q - quantile.

Sample quantiles and sample percentiles

```
probabilities<-c(0.25,0.5,0.75)
quantiles<-quantile(rainfall,probs=probabilities,na.rm=1)
```

```
ggplot(tibble(rainfall),aes(x=rainfall))+theme_bw()+
  geom_density(adjust=5,size=1)+xlab("Rain fall (mm)")+ylab("Density")+
  geom_vline(xintercept = quantiles,linetype="dashed",color="blue")
```



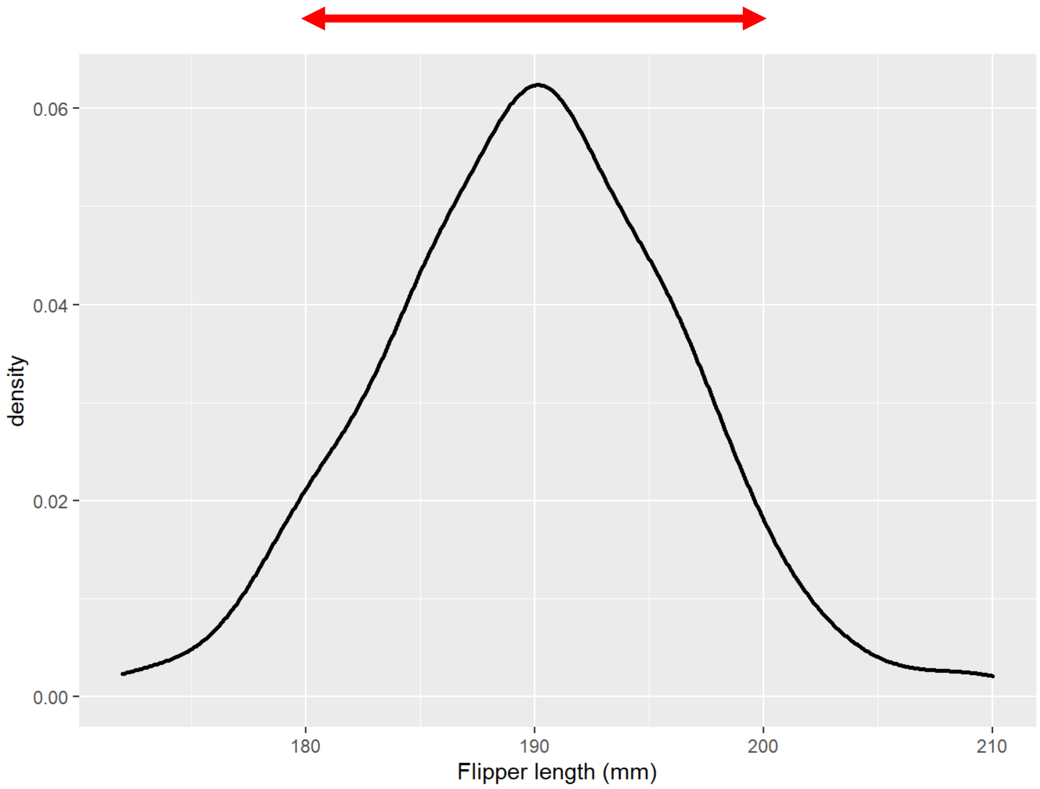
```
quantile(rainfall,probs=c(0.25,0.5,0.75),na.rm=1)
```

```
## 25% 50% 75%
## 0.0 0.0 1.5
```

Estimates of variability

Location is just one aspect of a feature in a data set.

Another crucial aspect of a feature in a data set is its variability or dispersion.



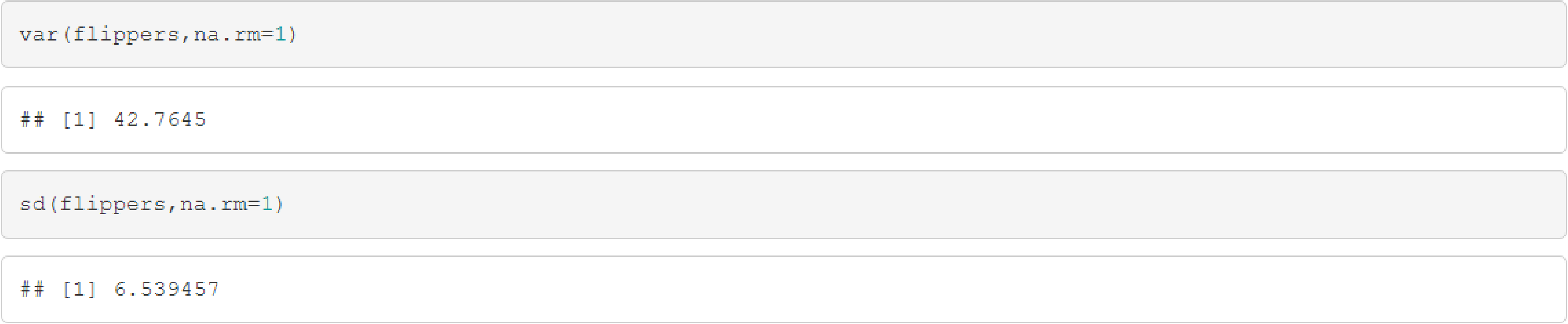
The sample variance and sample standard deviation

The classical measures of variability are the sample variance and sample standard deviation.

Given $X_1, \cdots, X_n$ the sample variance and sample standard deviation are defined as follows:

$$\text{Sample-variance} := \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \text{Sample-mean})^2$$

$$\text{Sample-standard-deviation} := \sqrt{\text{Sample-variance}}$$



```
var(flippers,na.rm=1)
```

```
## [1] 42.7645
```

```
sd(flippers,na.rm=1)
```

```
## [1] 6.539457
```

The sample median absolute deviation

The median absolute deviation is a robust alternative to the standard deviation.

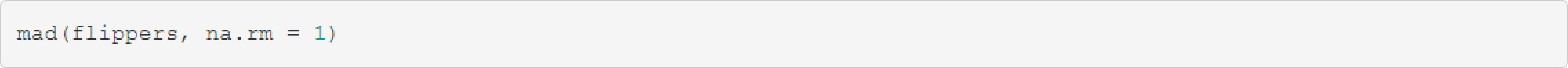
Given $X_1, \cdots, X_n$ first compute the sample median $\mathrm{MEDIAN}\left(X_1, \cdots, X_n\right)$

Now for each $i = 1, \cdots, n$ compute the absolute deviations from the sample median

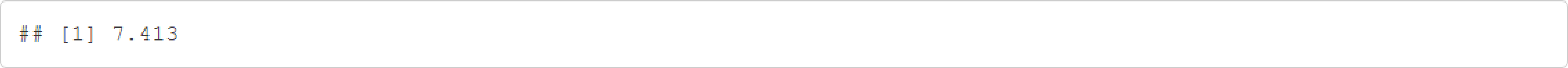
$$D_i := \left|X_i - \mathrm{MEDIAN}\left(X_1, \cdots, X_n\right)\right|$$

The sample median absolute deviation is given by

$$\mathrm{MAD}\left(X_1, \cdots, X_n\right) := 1.4826 * \mathrm{MEDIAN}\left(D_1, \cdots, D_n\right)$$



```
mad(flippers, na.rm = 1)
```



```
## [1] 7.413
```

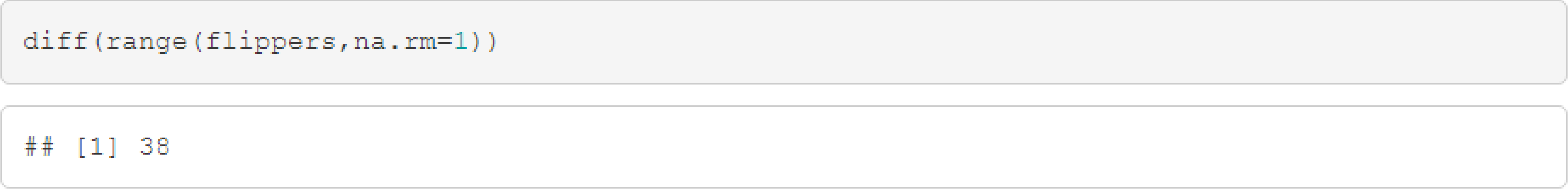
The sample range

The simplest estimate of variability is the sample range.

Given $X_1, \cdots, X_n$ we compute the range as follows:

First sort the data to obtain $X^{(1)} \leq X^{(2)} \leq \cdots \leq X^{(n)}$ then compute

$$\text{Range} := X^{(n)} - X^{(1)}$$



```
diff(range(flippers,na.rm=1))
```

```
## [1] 38
```

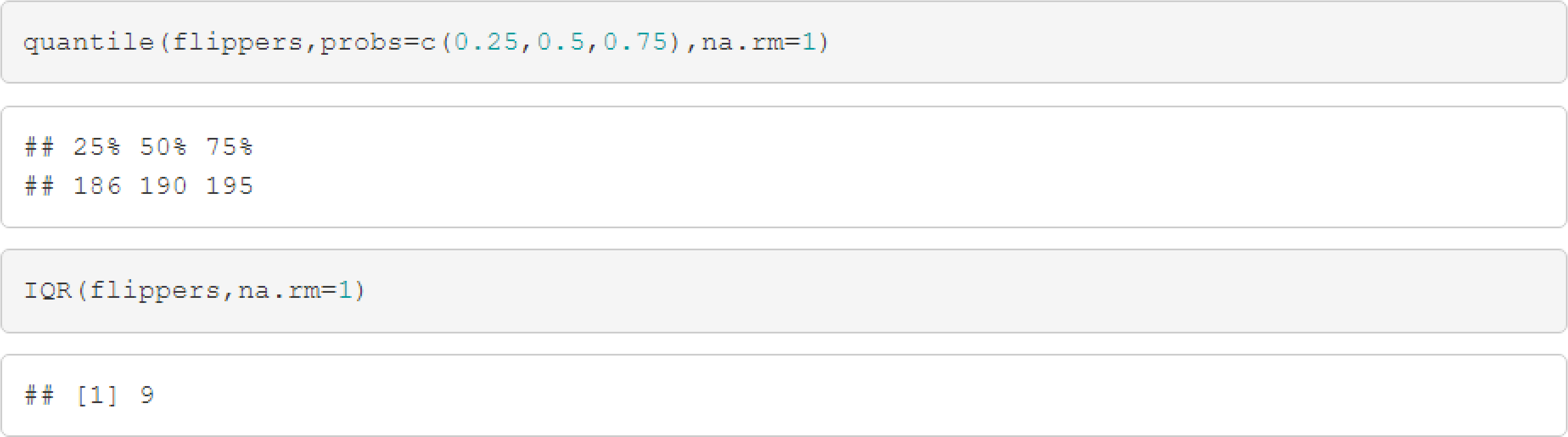
The range has the major drawback of being extremely sensitive to outliers.

The interquartile range

The concept of quantiles can be used to give a more robust estimate of variability.

Given $X_1, \cdots, X_n$ the interquartile-range is defined as follows:

$$\text{Interquartile range} := 0.75\text{-quantile} - 0.25\text{-quantile}.$$



```
quantile(flippers,probs=c(0.25,0.5,0.75),na.rm=1)
```

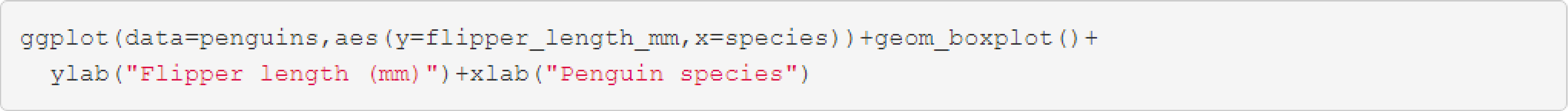
```
## 25% 50% 75%
## 186 190 195
```

```
IQR(flippers,na.rm=1)
```

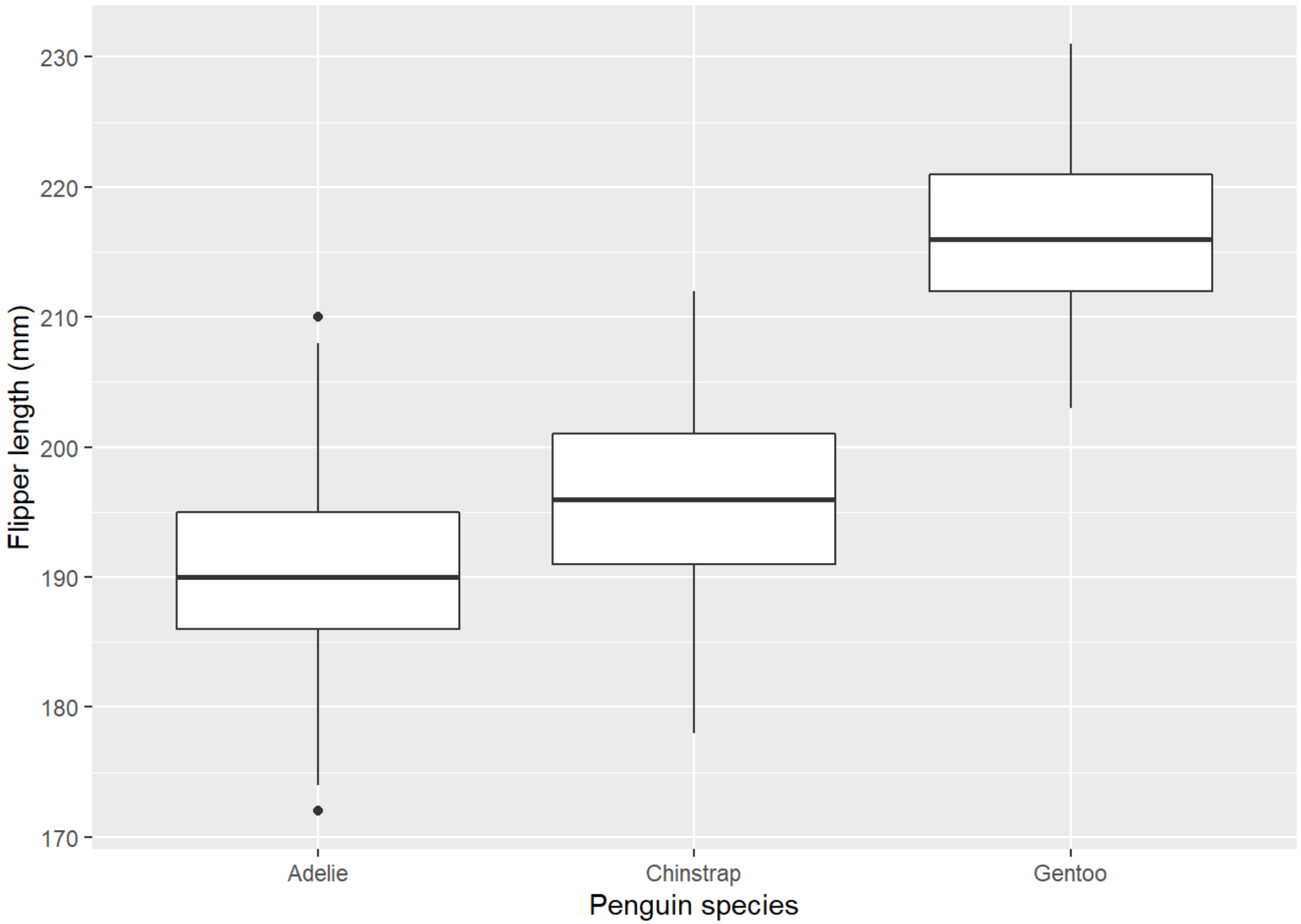
```
## [1] 9
```

Now take a break!

Understanding box plots



```
ggplot(data=penguins,aes(y=flipper_length_mm,x=species))+geom_boxplot()+
  ylab("Flipper length (mm)")+xlab("Penguin species")
```



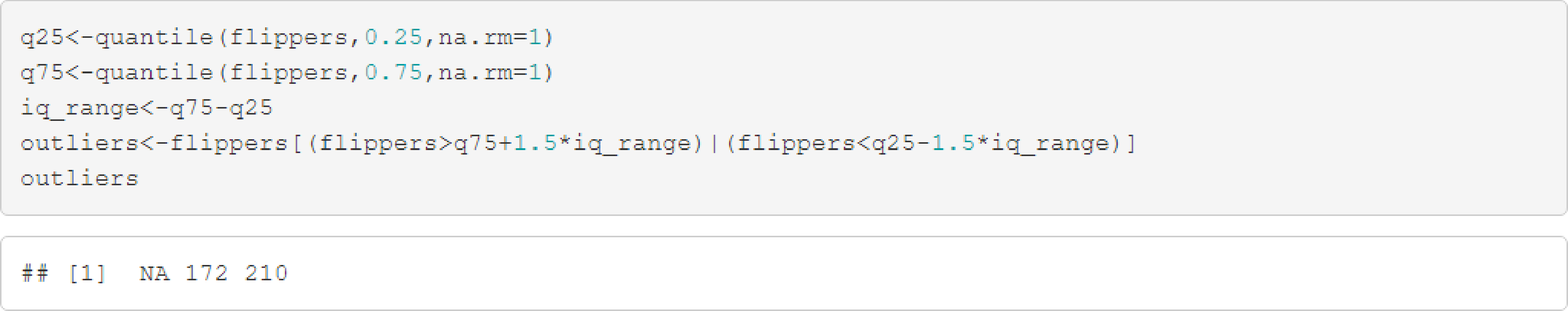
Interquartile range and outliers

An outlier is a value in a data set which differs substantially from other values.

Quantitative formulation:

$$X_i > 0.75\text{-quantile} + 1.5 \times \text{Interquartile-range}$$

or $$X_i < 0.25\text{-quantile} - 1.5 \times \text{Interquartile-range}$$

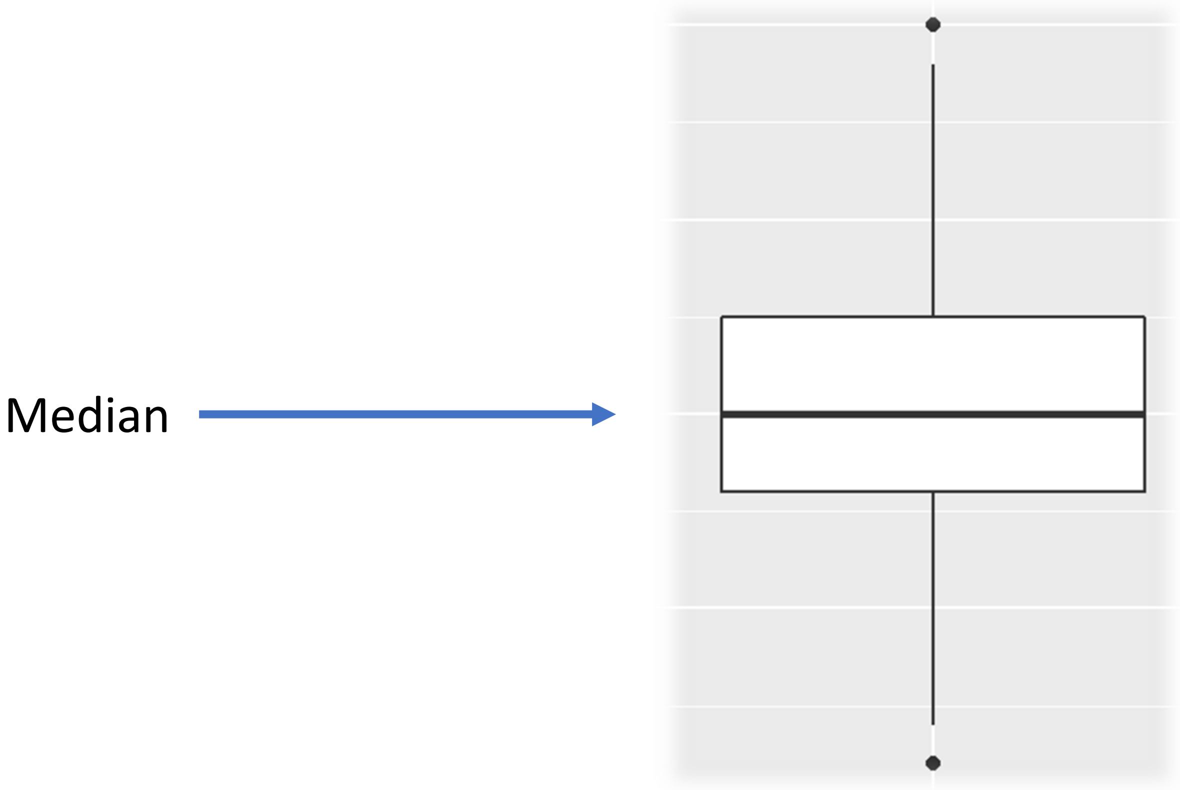


```
q25<-quantile(flippers,0.25,na.rm=1)
q75<-quantile(flippers,0.75,na.rm=1)
iq_range<-q75-q25
outliers<-flippers[(flippers>q75+1.5*iq_range)|(flippers<q25-1.5*iq_range)]
outliers
```

```
## [1]  NA 172 210
```

Understanding box plots

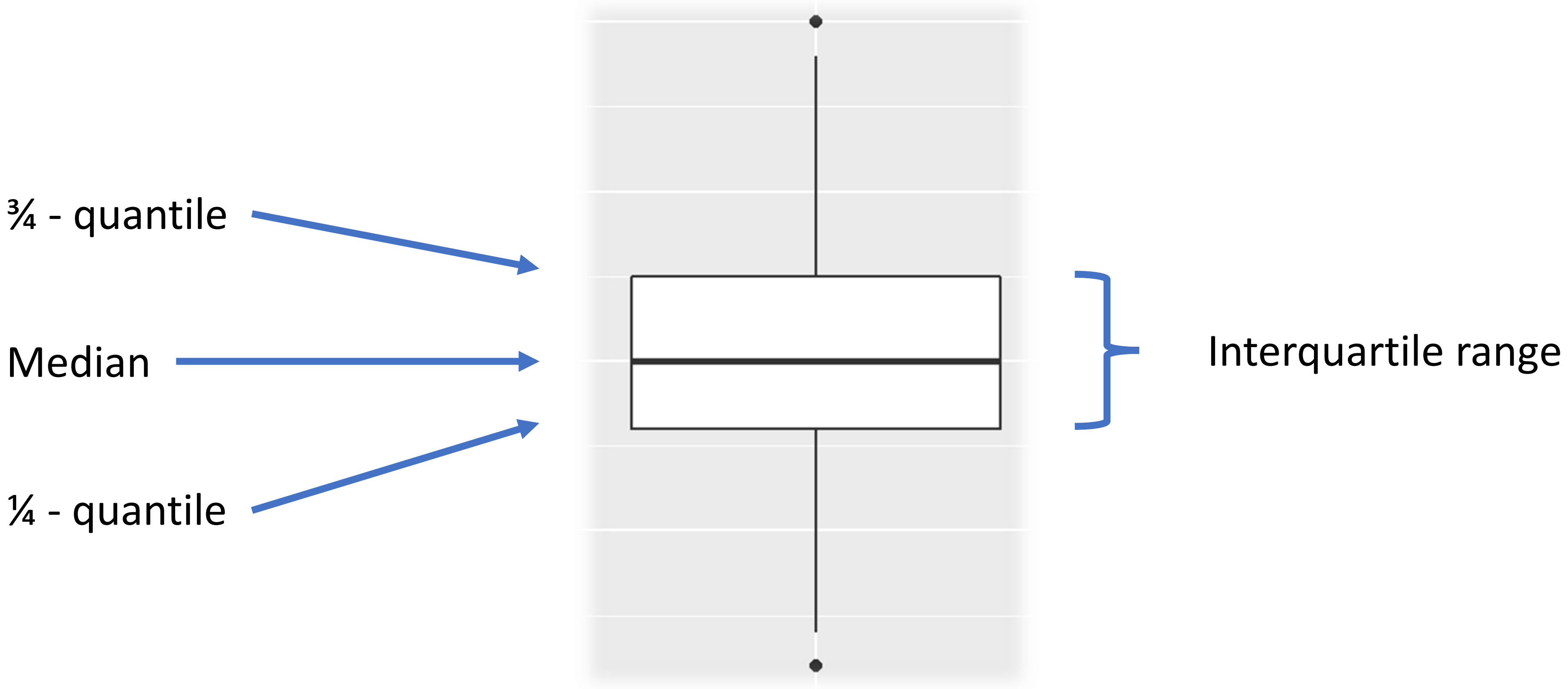
How do we interpret box plots?



Median

Understanding box plots

How do we interpret box plots?



¾ - quantile

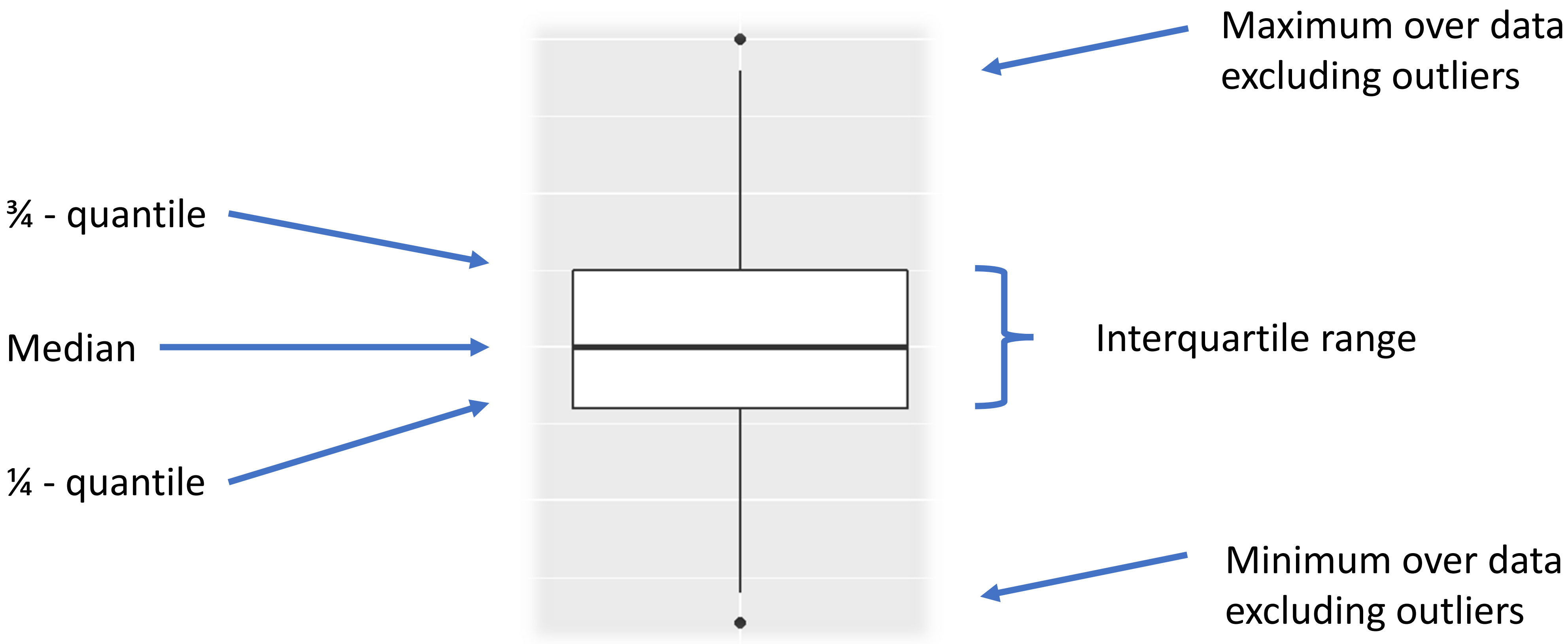
Median

¼ - quantile

Interquartile range

Understanding box plots

How do we interpret box plots?



¾ - quantile

Median

¼ - quantile

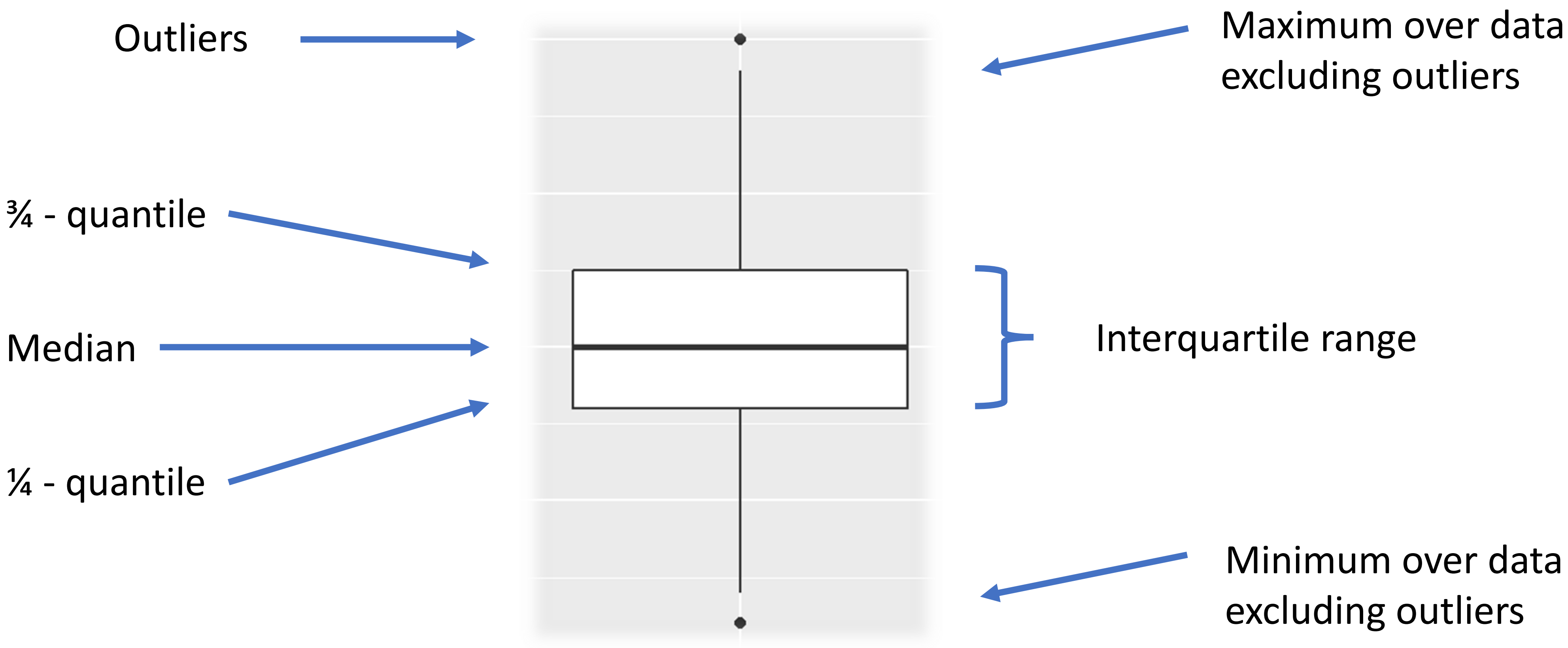
Maximum over data excluding outliers

Interquartile range

Minimum over data excluding outliers

Understanding box plots

How do we interpret box plots?



Outliers

¾ - quantile

Median

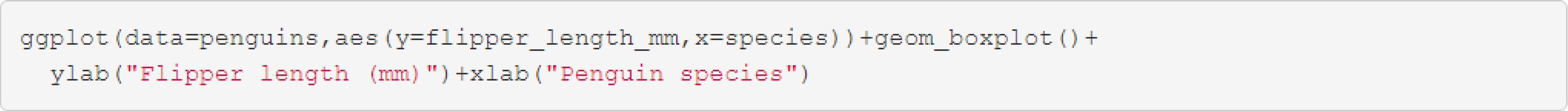
¼ - quantile

Maximum over data excluding outliers

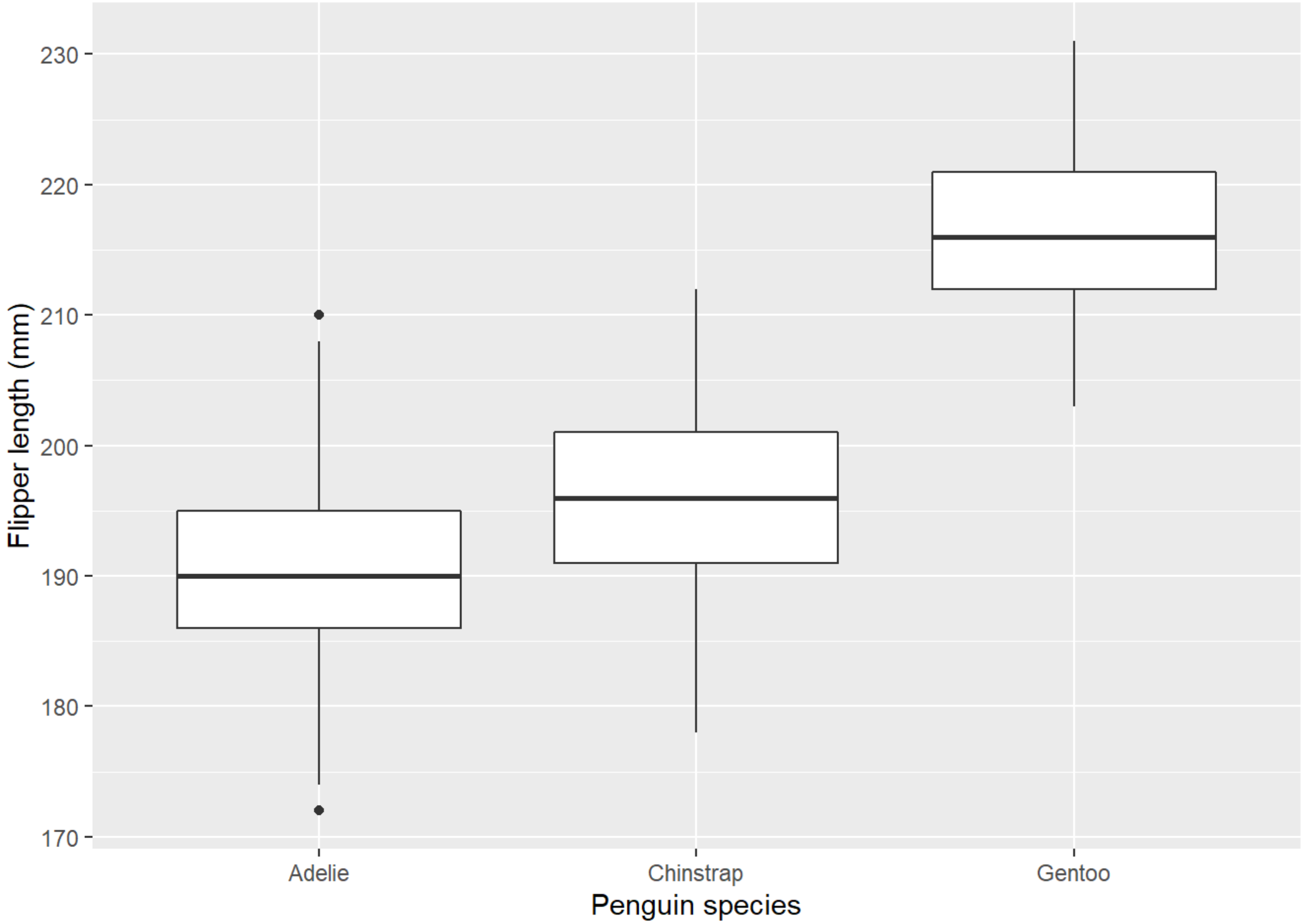
Interquartile range

Minimum over data excluding outliers

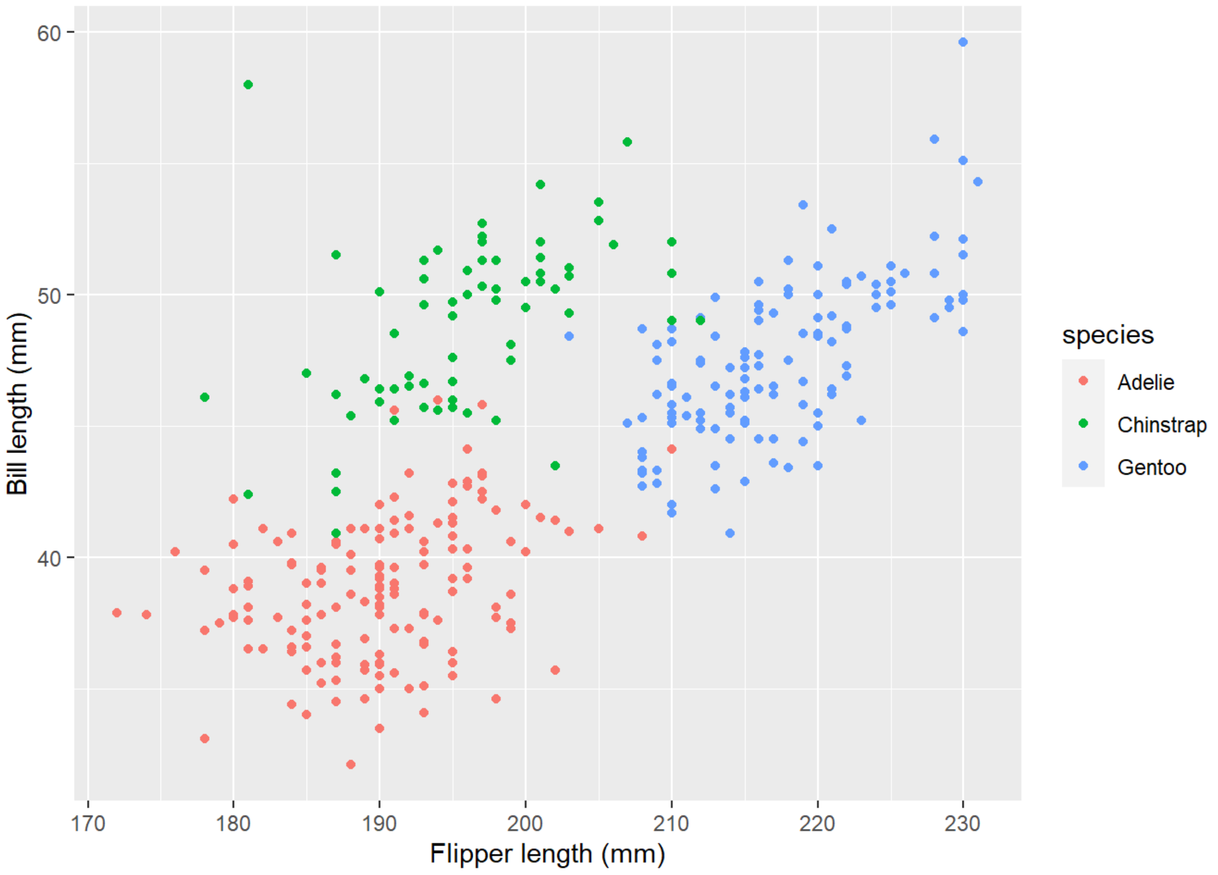
Understanding box plots



```
ggplot(data=penguins,aes(y=flipper_length_mm,x=species))+geom_boxplot()+
  ylab("Flipper length (mm)")+xlab("Penguin species")
```



Relating variables via sample covariance and sample correlation



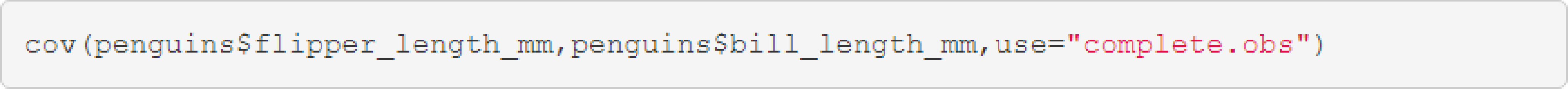
The covariance and correlation give us ways to see how connected two continuous variables.

Relating variables via sample covariance and sample correlation

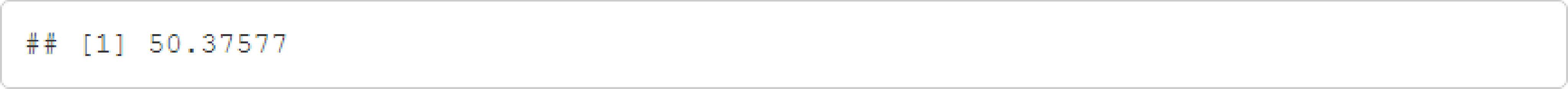
The **sample covariance** give us ways to see how connected two variables or features.

Given two variables $X_1, \cdots, X_n$ and $Y_1, \cdots, Y_n$ these can be computed as follows:

$$\text{COVAR}\left((X_i)_{i=1}^n, (Y_i)_{i=1}^n\right) := \frac{1}{n-1}\sum_{i=1}^n \left(X_i - \text{MEAN}\left((X_i)_{i=1}^n\right)\right) \cdot \left(Y_i - \text{MEAN}\left((Y_i)_{i=1}^n\right)\right)$$



```r
cov(penguins$flipper_length_mm,penguins$bill_length_mm,use="complete.obs")
```



```
## [1] 50.37577
```

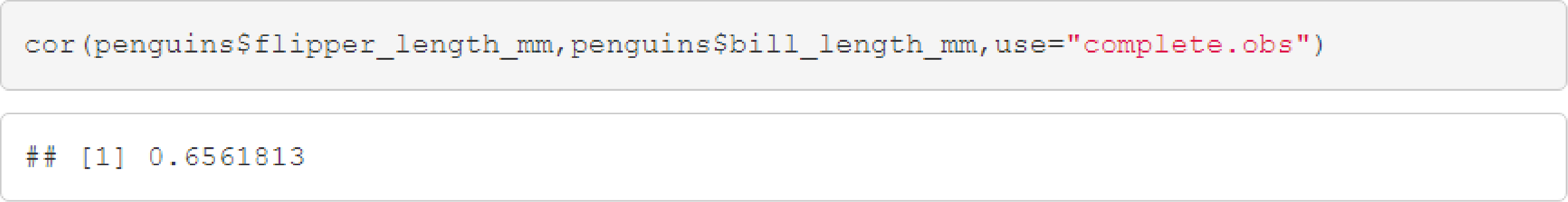
Relating variables via sample covariance and sample correlation

The **sample correlation** give us a way to see how connected two variables or features are.

Given two variables $X_1, \cdots, X_n$ and $Y_1, \cdots, Y_n$ the sample correlation is computed as follows:

$$\mathrm{COVAR}\left((X_i)_{i=1}^n, (Y_i)_{i=1}^n\right) := \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \mathrm{MEAN}\left((X_i)_{i=1}^n\right)\right) \cdot \left(Y_i - \mathrm{MEAN}\left((Y_i)_{i=1}^n\right)\right)$$

$$\mathrm{CORR}\left((X_i)_{i=1}^n, (Y_i)_{i=1}^n\right) := \frac{\mathrm{COVAR}\left((X_i)_{i=1}^n, (Y_i)_{i=1}^n\right)}{\mathrm{S\_DEV}\left((X_i)_{i=1}^n\right) \cdot \mathrm{S\_DEV}\left((Y_i)_{i=1}^n\right)}$$



```
cor(penguins$flipper_length_mm,penguins$bill_length_mm,use="complete.obs")
```

```
## [1] 0.6561813
```

Relating variables via sample covariance and sample correlation

$$\mathrm{CORR}\left((X_i)_{i=1}^n, (Y_i)_{i=1}^n\right) := \frac{\mathrm{COVAR}\left((X_i)_{i=1}^n, (Y_i)_{i=1}^n\right)}{\mathrm{S\_DEV}\left((X_i)_{i=1}^n\right) \cdot \mathrm{S\_DEV}\left((Y_i)_{i=1}^n\right)}$$

Two features are **positively correlated** if one tends to be higher than average when the other is.

Example The height and weight of an animal are positively correlated.

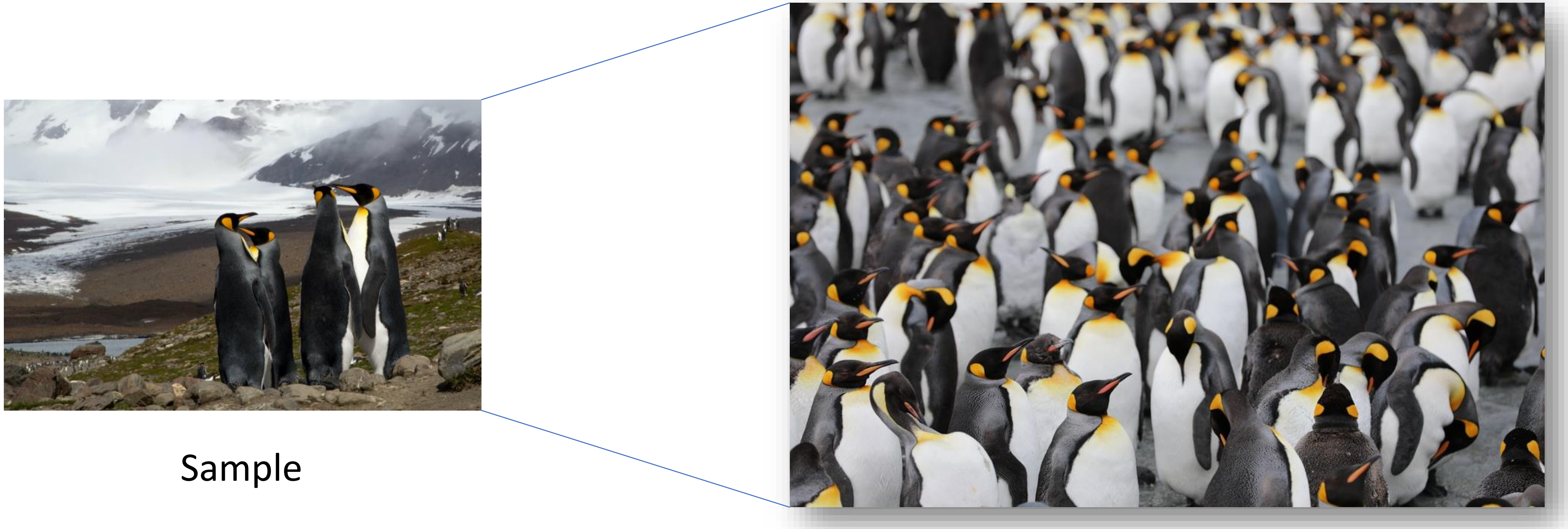
Two features are **negatively correlated** if one variable tends to be higher than average when the other variable is lower than average.

Example Hours per week of cardio-vascular exercise and resting heart rate.

Sample vs. population quantities

Why do we refer to the “sample mode” and “sample mean” rather than just the “mode” and “mean”?

We view the data set as a sample from a much larger population of penguins.



Sample

Population

Sample vs. population quantities

We view the data set as a sample from a much larger population of penguins.

The data set is often referred to as a data sample, or just sample.

A statistic (aka sample statistics or summary statistic) is any function of the sample.

Whilst we compute sample statistics based on the data …

…. our true interests often lie in the associated population quantity.

Example: We compute the sample mode penguin species, but our real interest is in the population mode i.e. the most common penguin in the population.

Making inferences about the underlying population based on the sample lies at the heart of statistics.

What have we covered?

- We gave a taxonomy of the different types of data

- We discussed a wide variety of location estimators

- We introduced the concepts of sample quantiles, percentiles and quartiles.

- We also considered several estimators of variability.

- We concluded by introducing correlation as a measure of interdependency between two variables.

Thanks for listening!

Henry W J Reeve

henry.reeve@bristol.ac.uk

Unit: EMATM0061