Exploratory data analysis

Describing and summarizing data with fundamental statistical quantities

Statistical Computing and Empirical Methods
Unit EMATM0061, Data Science MSc

Rihuan Ke
rihuan.ke@bristol.ac.uk

Teaching Block 1, 2024

University of BRISTOL

What we will 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 the overall variability of a feature in a data set.

We will explore methods for estimating how connected two features in a data set are.

Exploratory data analysis

Suppose that we have a new data set (also referred to as a data sample).

Before performing a formal data analysis process, we can explore it and gain a preliminary understanding, by carrying out exploratory data analysis:

1) Generate questions about the data
2) Find answers by inspecting your data, with modelling and visualisation techniques
3) Based on the understanding you gained, generate new questions or refine your questions, and go to step 2)

Typical exploratory data analysis processes:

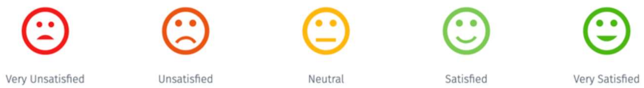
1) Understanding the meaning and data type of each of the variables aka. features.
2) Computing sample statistics (mean, median, variance, etc.) to understand the main characteristic of the data
3) Using visualisation to efficiently identify the shape of distributions and key relationships.

A taxonomy of data types

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

Common data types:

1) **Continuous**: Data that can take any value on an interval e.g. bill length in mm
2) **Discrete**: Data with a minimum distance between possible values e.g. year, number of restaurant meals in a month
3) **Categorical**: Data that can take on only a specific set of values representing distinct categories e.g. brand, species, island.
4) **Binary**: Categorical data with exactly two categories e.g. pass or fail a driving test.
5) **Ordinal**: Categorical data with an ordering e.g. "How was your meal?" on a Likert scale.



Example: Palmer penguins data set

```
## # A tibble: 6 × 8
##   species island    bill_length_mm bill_depth_mm flipper_l…¹ body_…² sex    year
##   <fct>   <fct>              <dbl>         <dbl>       <int>   <int> <fct> <int>
## 1 Adelie  Torgersen           39.1          18.7         181    3750 male   2007
## 2 Adelie  Torgersen           39.5          17.4         186    3800 fema…  2007
```

Sample statistics

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

- mean, median, etc. of the sample (they are functions of the sample)

Typical statistics that we will cover next:

1) Sample mode
2) Sample mean
3) Sample median
4) Trimmed sample mean
5) Sample quantiles and sample percentiles
6) Sample variance and sample standard deviation
7) Sample median absolute deviation
8) Sample range
9) Interquartile range
10) Sample covariance and sample correlation

1. Sample mode

Estimates of location

- **Question**: which single value is most representative or typical?

For **categorical data**, the natural answer is the sample mode.

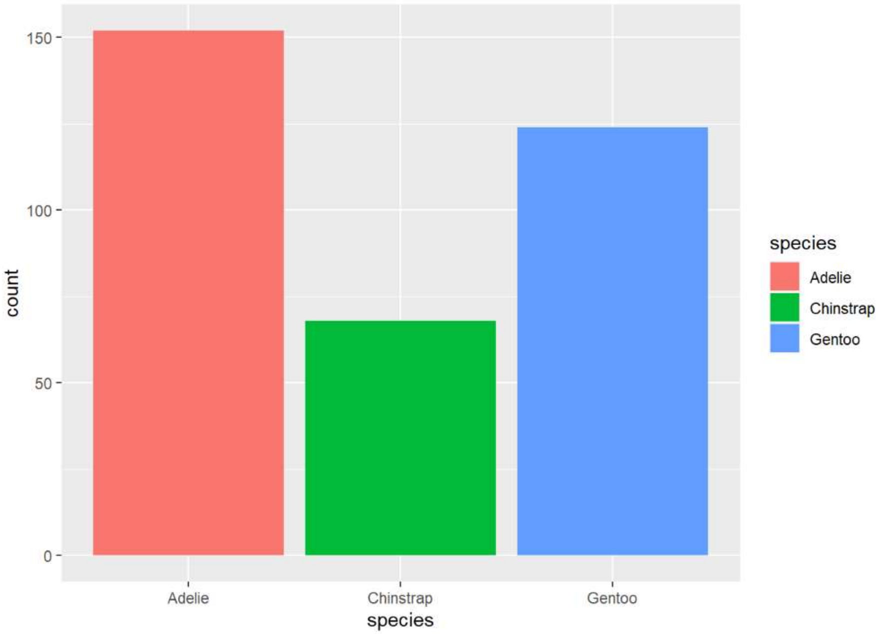
Definition: the sample mode is the value which occurs with the highest frequency for a feature within a data set

The sample mode can be computed by the function mfv1 (from the **modeest** package).
Example (the penguins dataset):

```
library(palmerpenguins) # Load the Palmer penguins data set
library(modeest)
mfv1(penguins$species) # sample mode for the species column
```

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

```
ggplot(penguins, aes(x=species, fill=species )) + geom_bar()
```



2. Sample mean

Estimates of location

- **Question**: which single value is most representative or typical?

For numeric type data (e.g., continues, discrete variables), the most well-known estimate of location is the sample mean (the arithmetic mean)

Definition. Suppose that the variable of interest has values $x_1, x_2, \cdots, x_n$, then the sample mean is given by

$$\text{sample mean} := \frac{1}{n}(x_1 + x_2 + \cdots + x_n)$$

Example:

Sample: 1 2 3 4 10

Then sample mean = (1+2+3+4+10)/5 = 4

Example

Suppose we have the following daily rainfall data for San Martino for the first 200 days of 1985:

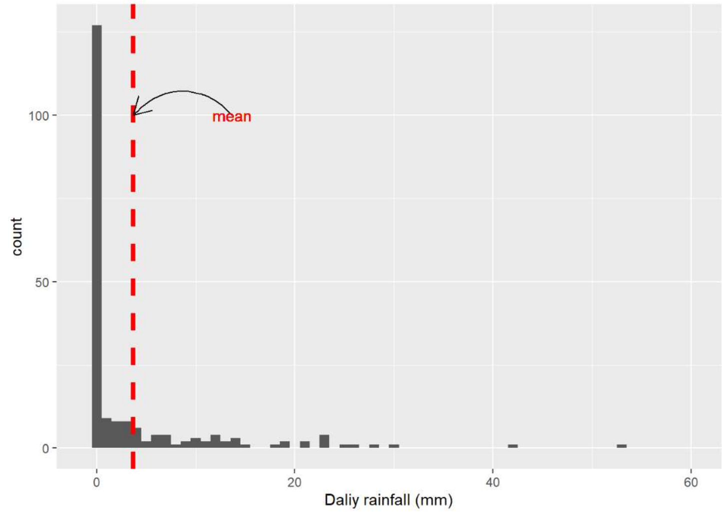
rainfall = (0.0 0.0 … 12.6 20.6 0.0 0.0)

The mean of rainfall can be computed in R using the mean() function:

```
rainmean <- mean(rainfall, na.rm=TRUE)
rainmean
```

```
## [1] 3.671642
```

This is the location of the sample mean in the histogram plot of the rainfall data:



3. Sample median

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

Definition. Suppose that the variable of interest has values $x_1, x_2, \cdots, x_n$, and $x_1 \leq x_2 \leq \cdots \leq x_n$ then the sample median is given by

$$\text{sample median} := \frac{1}{2}(x_{\lfloor (n+1)/2 \rfloor} + x_{\lceil (n+1)/2 \rceil}).$$

Here $\lfloor \cdot \rfloor$ is the floor function, $\lceil \cdot \rceil$ is the ceiling function. Eg.

$$\lfloor 3.4 \rfloor = 3, \quad \text{and} \quad \lceil 3.4 \rceil = 4.$$

Example:

Sample: 1 2 3 4 10

Then sample median = (3+3)/2 = 3

Example

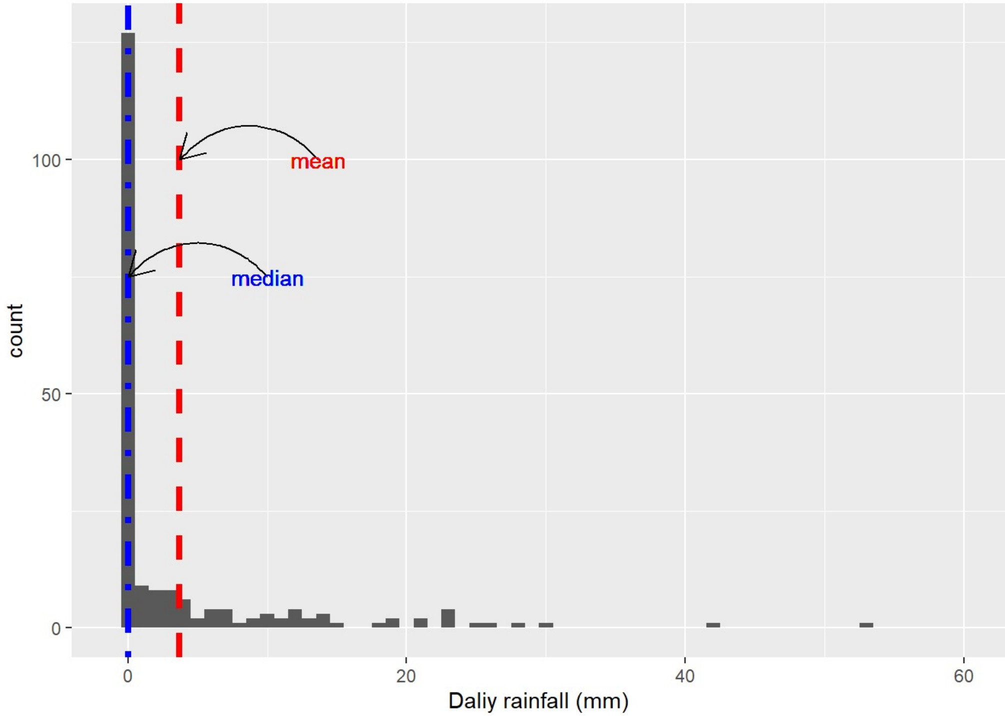
Let's use the rainfall data again as examples

The sample median can be computed in R using the median function:

```
rainmedian <- median(rainfall)
rainmedian
```

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

This is the location of the sample median in the histogram plot of the rainfall data:



The sample mean is in general not equal to the sample median

-- For distributions with a heavy right tail (e.g., the small subset of numbers with large values), the sample mean can be much bigger than the median

Sample mean and outliers

An outlier is a value in a data set which differs substantially from other values.
-- for example, a value is much bigger than the rest of the values

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

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

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

2. A faithful representation of a genuinely anomalous event e.g. a day of extremely unusual torrential rainfall.

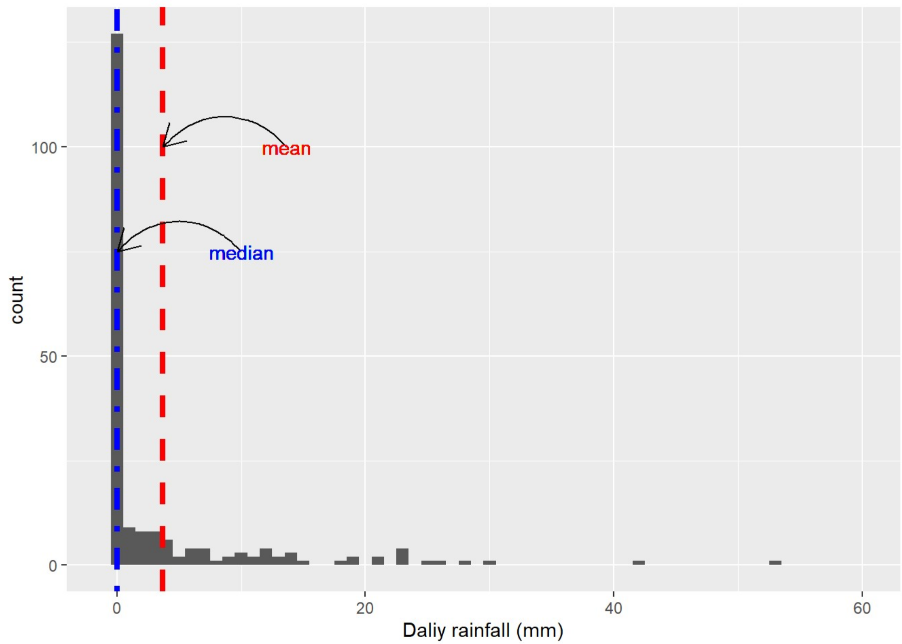
measurement error

anomalous event

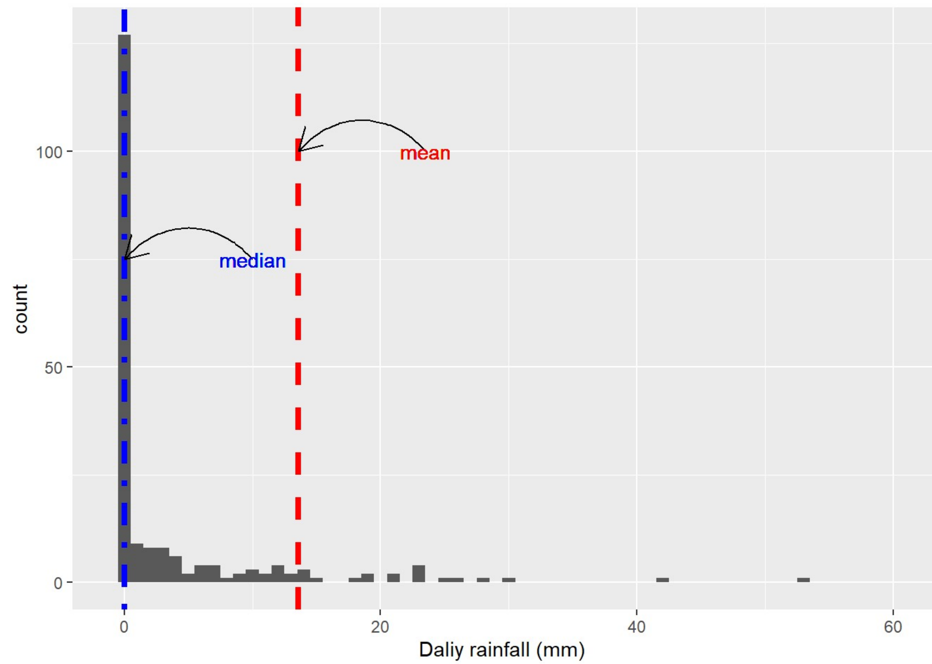
Robustness of the sample median

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

Uncorrupted data



Corrupted data with an outlier



The sample mean is significantly changed, while the sample median is not

Comparing the sample median and the sample mean

The sample medians have advantages and disadvantages:

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

2. The sample median effectively ignores a large section of the data set, unlike the mean. This makes it difficult to aggregate medians from multiple sources.

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

4. Trimmed sample mean

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

Definition. Suppose that the variable of interest has values $x_1, x_2, \cdots, x_n$, and $x_1 \leq x_2 \leq \cdots \leq x_n$. The trimmed sample mean with trim fraction $q \in (0, 1/2]$ is computed as follows:

$$\text{trimmed sample mean} := \frac{1}{n - 2 \cdot \lfloor q \cdot n \rfloor} \sum_{i=\lfloor q \cdot n \rfloor + 1}^{n - \lfloor q \cdot n \rfloor} x_i.$$

Example:

Sample: 1 2 3 4 10

Trimmed sample mean (with q=1/4) = (2+3+4)/3 = 3

Recall that: the sample mean = (1+2+3+4+10)/5 = 4

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

Example

Let's use the rainfall data again as examples

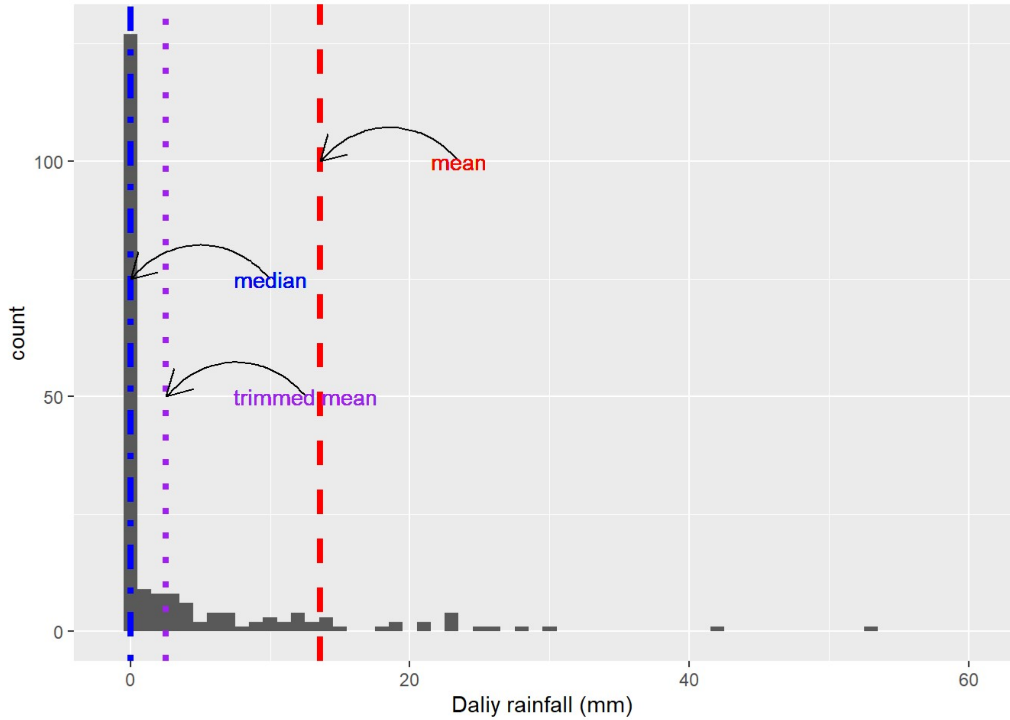
The trimmed sample mean can be computed in R using the mean function:

```
rainmean_trim <- mean(rainfallv2, na.rm=TRUE, trim=0.05)
print(rainmean_trim)
```

```
## [1] 2.556044
```

This is the location of the trimmed sample mean in the histogram plot of the rainfall data:

Corrupted data with an outlier



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

Another example: Palmer Penguins Dataset

With the Palmer Penguins Dataset, compute the sample mean, median, and trimmed mean of the flipper length of the Adelie species. Demonstrate the value of the three statistics in a histogram plot (the result is on the next slide)

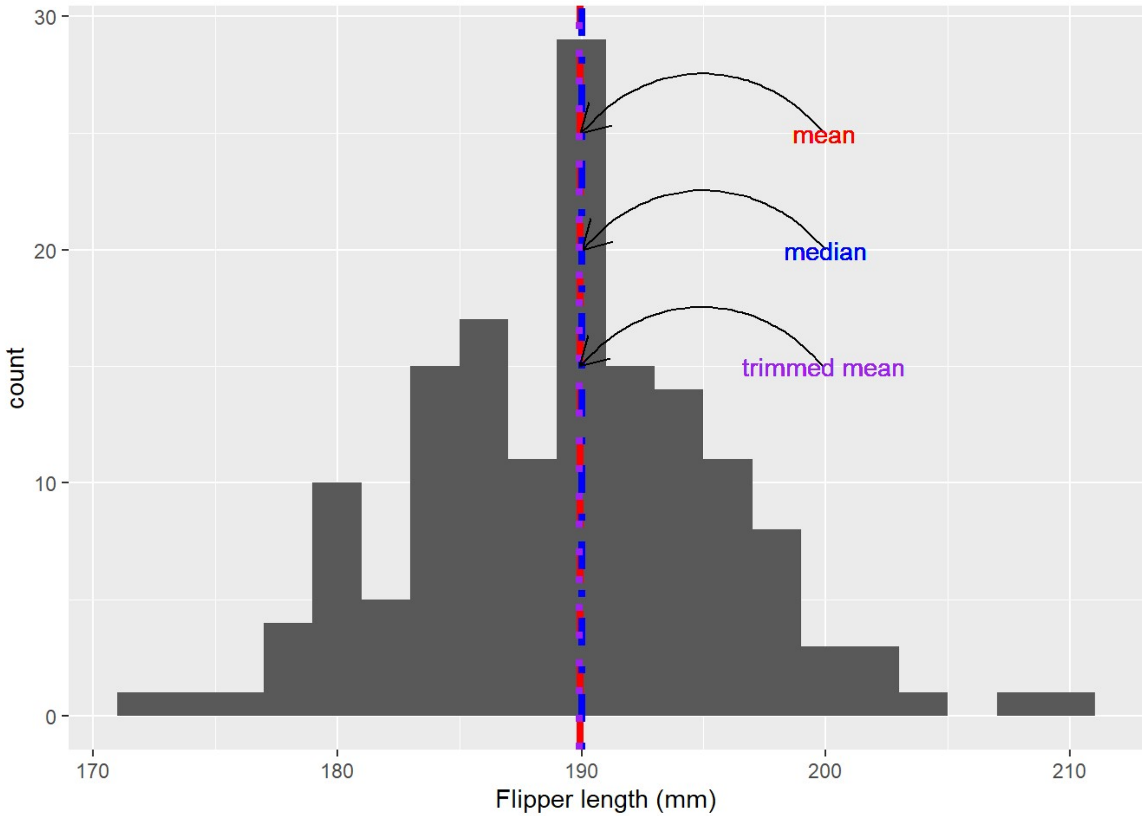
```
flippers <- penguins %>% filter(species == 'Adelie') %>%
  select(flipper_length_mm) %>% unlist() %>% as.vector() #flipper data for the Adelie species
f_mean <- mean(flippers, na.rm=TRUE) # sample mean
f_median <- median(flippers, na.rm=TRUE) # sample median
f_mean_trim <- mean(flippers, na.rm=TRUE, trim=0.05) # trimmed sample mean
```

```
# a function for adding arrow & annotation
vline_w_anno <- function(plot_object, value, linetype, color, y, label){
  plot_object2 <- plot_object +
    geom_vline(xintercept=value, linetype=linetype, color=color, size=1.5) +
    geom_curve(x=value+10, xend=value, y=y, yend=y, arrow=arrow(length=unit(0.5,'cm'))) +
    geom_text(x=value+10, y=y, label=label, color=color)
  return (plot_object2)
}

penguins_plot <- ggplot(tibble(flippers), aes(x=flippers)) +
  xlab('Flipper length (mm)') + geom_histogram(binwidth = 2) # histogram plot
penguins_plot %>%
  vline_w_anno(f_mean, 'dashed', 'red', 25, 'mean') %>% # annotation for the mean value
  vline_w_anno(f_median, 'dotdash', 'blue', 20, 'median') %>% # annotation for the median value
  vline_w_anno(f_mean_trim, 'dotted', 'purple', 15, 'trimmed mean') # annotation for the median value
```

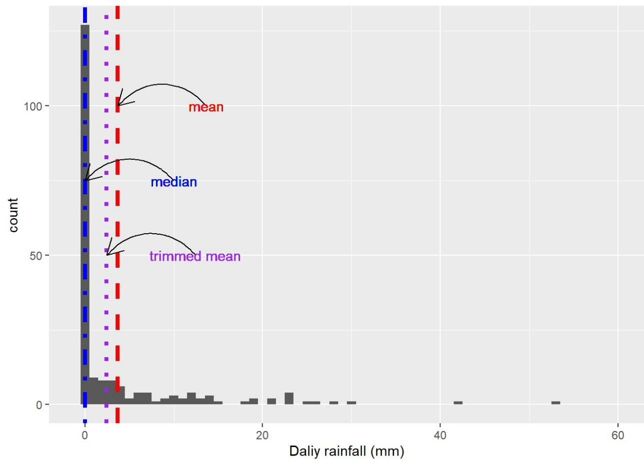
Another example: Palmer Penguins Dataset

The result looks like this:



For data with normal (Gaussian) distribution, the three statistics (sample mean, median, trimmed mean) tends to have similar values

In contrast, the previous rainfall data has a heave tail, their values are different



5. Sample quantiles and sample percentiles

A sample median can be seen as a point that ranks in the middle of the data set

Sample quantiles extend this notion to other fractions. e.g. which point ranks 1/4 in your data?

Definition. Suppose that the variable of interest has values $x_1, x_2, \cdots, x_n$, and $x_1 \leq x_2 \leq \cdots \leq x_n$. For $q \in (0,1]$, the q-quantile is of the following form:

$$x_{\max(\lfloor qn \rfloor, 1)} \leq \text{q-quantile} \leq x_{\lceil qn \rceil}.$$

Example:

Sample: 1 2 3 4 10

Then 2 is a 0.25-quantile of the sample; 3 is a 0.5-quantile of the sample

5. Sample quantiles and sample percentiles

Sample percentiles are similar to sample quantiles

Definition: For $q \in [0, 100]$, the sample q-th percentile is precisely the same as the sample (0.01q)-quantile

So 25^{th} percentile = 0.25 quantile, and 78^{th} percentile = 0.78 quantile

Example:

Sample: 1 2 3 4 10

Then 2 is the 25^{th} percentile of the sample; 3 is the 50^{th} percentile of the sample

Quartile:

1 quartile = 25^{th} percentile (also 0.25-quartile)

2 quartile = 50^{th} percentile (also 0.50-quartile)

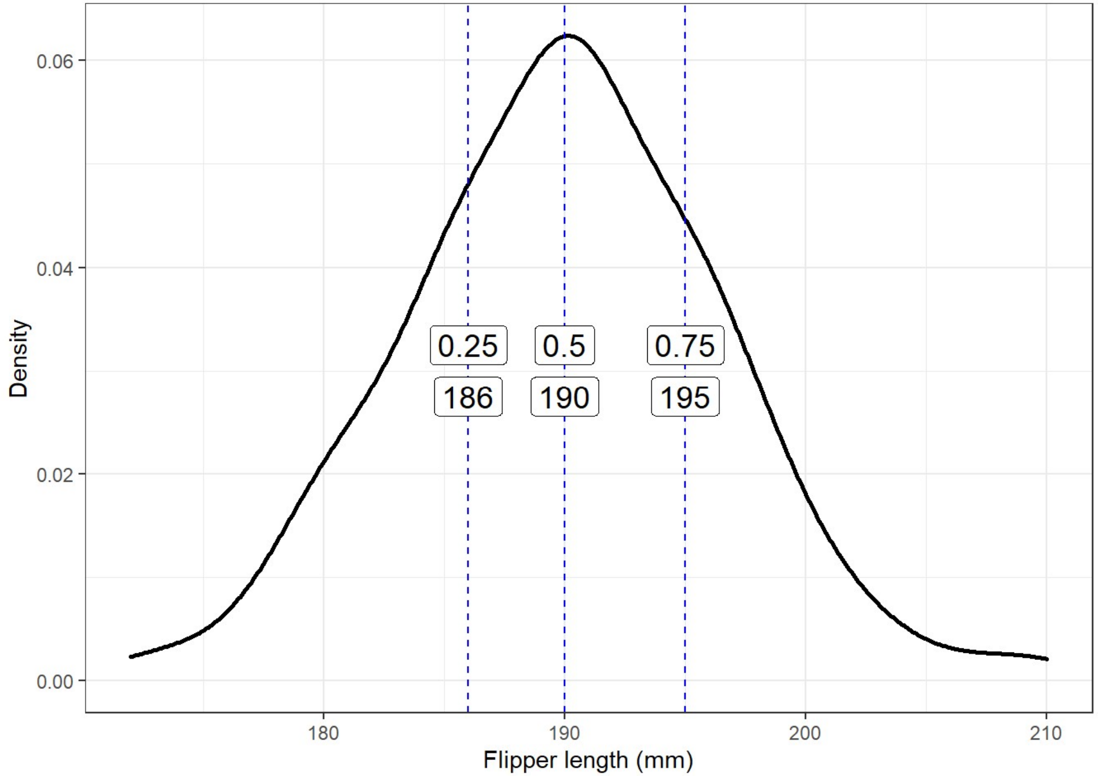
3 quartile = 75^{th} percentile (also 0.75-quartile)

Example: Palmer Penguins Dataset

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

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

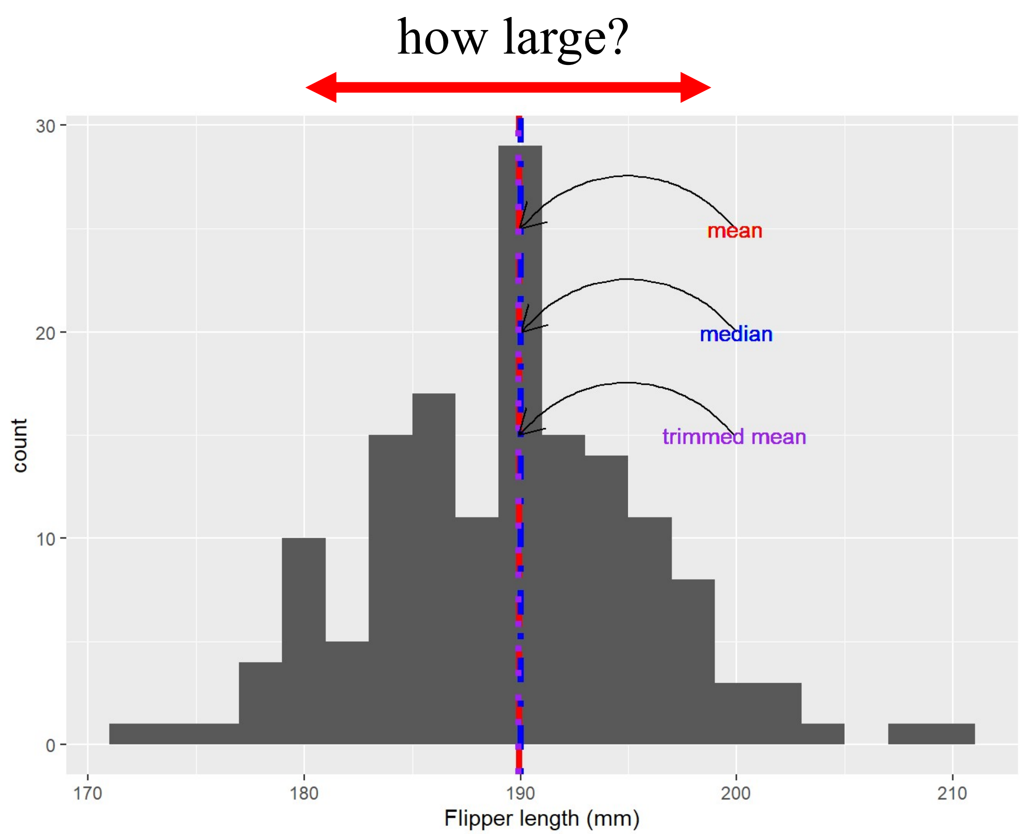
```r
ggplot(tibble(flippers), aes(x=flippers)) + theme_bw() +
  geom_density(adjust=1, size=1) + xlab('Flipper length (mm)') + ylab('Density') +
  geom_vline(xintercept = quantiles, linetype='dashed', color='blue') +
  annotate('label', x=quantiles, y=0.0325, size=5, fill='white', label=probabilities)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



6. Sample variance and sample standard deviation

The statistics like mean and median are about location, which 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.



6. Sample variance and sample standard deviation

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

Definition. Suppose that the variable of interest has values $x_1, x_2, \cdots, x_n$, then the sample variance and sample standard deviation are given by

$$\text{sample-variance} := \frac{1}{n-1}\sum_{i}^{n}(x_i - \text{sample-mean})^2,$$

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

Example:

Sample: 1 2 3 4 10

Sample-variance = ((1-4)^2+(2-4)^2+(3-4)^2+(4-4)^2+(10-4)^2)/4 = 12.5

Example (using R function var() and sd()):

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

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

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

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

7. Sample median absolute deviation

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

Definition. Suppose that the variable of interest has values $x_1, x_2, \cdots, x_n$, then the sample median absolute deviation is computed by

$$D_i := |x_i - \text{Median}(x_1, x_2, \cdots, x_n)|, \quad i = 1, 2, \cdots, n,$$
$$\text{sample-median-absolute-deviation} := 1.4826 \cdot \text{Median}(D_1, D_2, \cdots, D_n).$$

Here Median() is the function for computing sample medians.

Example (using R function mad()):

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

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

8. Sample range

Another simple estimate of variability is the sample range

Definition. Suppose that the variable of interest has values $x_1, x_2, \cdots, x_n$, and $x_1 \leq x_2 \leq \cdots \leq x_n$. The sample range is given by:

$$\text{sample range} := x_n - x_1.$$

So the sample range is the largest value subtracted by the smallest value

Example (computing sample range using R)**:**

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

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

Note: the range() function computes the smallest and largest values

```
range(flippers, na.rm=TRUE)
```

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

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

9. Interquartile range

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

The interquartile range is the range of the sample after removing the largest/smallest values

Definition. Suppose that the variable of interest has values $x_1, x_2, \cdots, x_n$, then the interquartile range is computed by

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

Example (computing the interquartile range using R):

```
quantiles=quantile(flippers, prob=c(0.25, 0.5, 0.75), na.rm=TRUE)
print(quantiles)
```

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

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

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

Interquartile range and outliers

Recall that:

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

-- for example, a value is much bigger than the rest of the values

One way to find outliers is based on quantitative formulation:

Suppose that the variable of interest has values $x_1, x_2, \cdots, x_n$, then x_i is an outlier if

$$x_i > 0.75\text{-quantile} + 1.5 \times \text{Interquartile-range} \quad \text{or}$$
$$x_i < 0.25\text{-quantile} - 1.5 \times \text{Interquartile-range}$$

```
quantile25 <- quantile(flippers, 0.25, na.rm=TRUE)
quantile75 <- quantile(flippers, 0.75, na.rm=TRUE)
iq_range <- quantile75 - quantile25 # Interquantile-range
outliers <- flippers[(flippers>quantile75+1.5*iq_range) | (flippers<quantile25-1.5*iq_range) ]
outliers
```

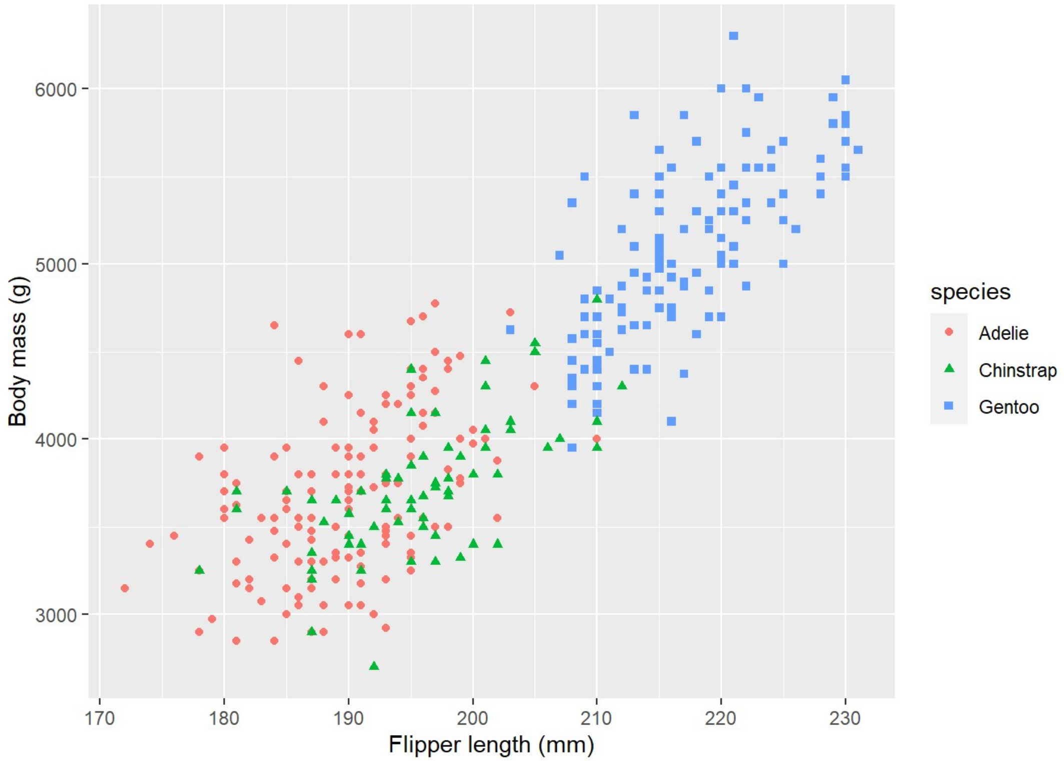
```
## [1]  NA 172 210
```

10. Relating variables via sample covariance and sample correlation

The sample mean, median, …, variance, and range are basic statistics for a single variable

If we have more than one variable, how can we describe the relationship among them?

A multivariate plot for the three penguin species:



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

Sample covariance and sample correlation

The sample covariance gives us ways to see how connected two variables or features are.

Definition. Suppose that two variables have values $x_1, \cdots, x_n$, and $y_1, \cdots,$ y_n. The sample covariance can be computed as

$$\mathrm{Covar}(\{x_i\}_i^n, \{y_i\}_i^n) := \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y}),$$

where $\bar{x}$ and $\bar{y}$ are the sample means of $\{x_i\}_i^n$ and $\{y_i\}_i^n$, respectively.

Example (computing the covariance of flipper length and bill length)**:**

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

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

NB: The sample covariance takes values in $(-\infty, +\infty)$.

Sample correlation

The sample correlation is a normalized version of the sample covariance.

Definition. Suppose that two variables have values $x_1, \cdots, x_n$, and $y_1, \cdots,$ y_n. The sample correlation can be computed as

$$\mathrm{Corr}(\{x_i\}_i^n, \{y_i\}_i^n) := \frac{\mathrm{Covar}((x_i)_{i=1}^n, (y_i)_{i=1}^n)}{\text{Standard-deviation}((x_i)_{i=1}^n) \cdot \text{Standard-deviation}((y_i)_{i=1}^n)}.$$

Recall that $\mathrm{Covar}(\{x_i\}_i^n, \{y_i\}_i^n) := \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}) \cdot (y_i - \bar{y})$.

Example (computing the correlation of flipper length and bill length)**:**

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

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

NB: The sample correlation takes values in $(-1, 1)$.

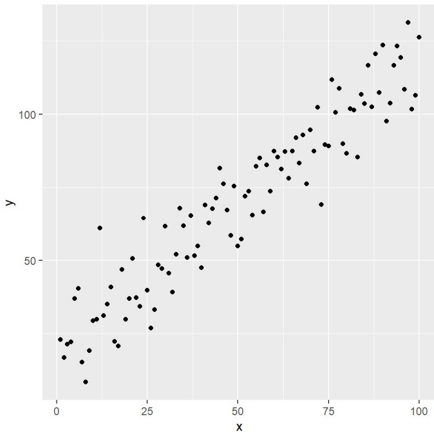
Positive correlation and negative correlation

Recall that $\text{Corr}(\{x_i\}_i^n, \{y_i\}_i^n) := \dfrac{\text{Covar}((x_i)_{i=1}^n, (y_i)_{i=1}^n)}{\text{Standard-deviation}((x_i)_{i=1}^n) \cdot \text{Standard-deviation}((y_i)_{i=1}^n)}$.

NB: The sample correlation takes values in $(-1, 1)$.

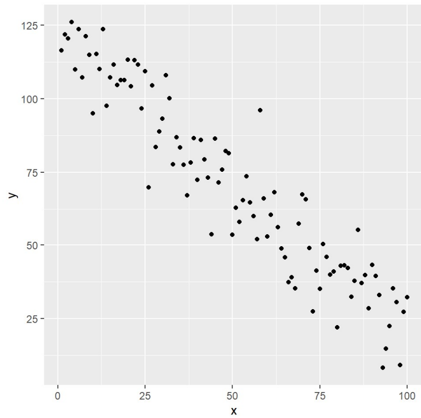
Two variables are positively correlated if one tends to be higher than average when the other is (i.e., $\text{Corr}(\{x_i\}_i^n, \{y_i\}_i^n) > 0$).

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



Two variables are negatively correlated if one tends to be higher than average when the other is lower than average (i.e., $\text{Corr}(\{x_i\}_i^n, \{y_i\}_i^n) < 0$).

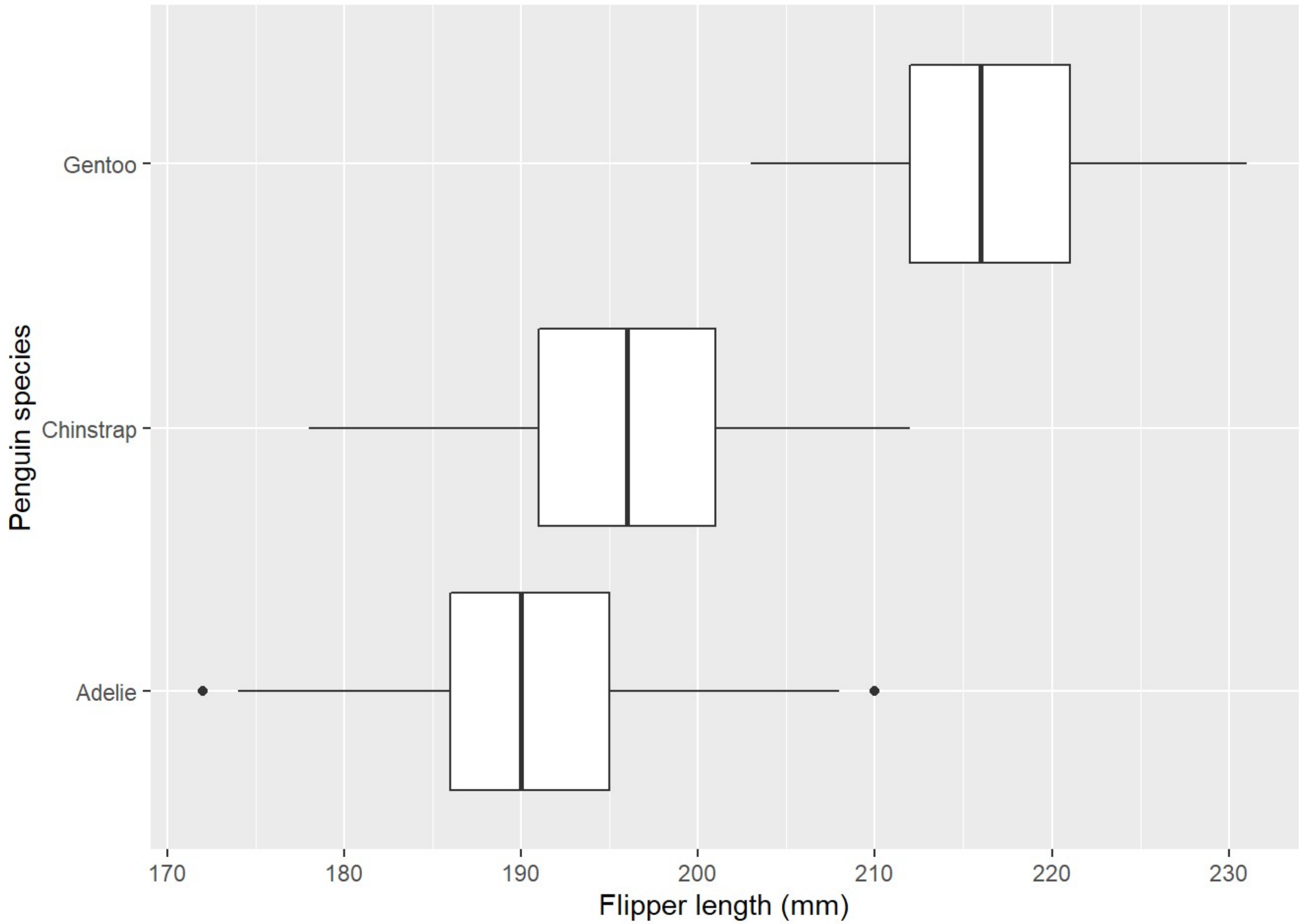
Example: The driving speed and number of car accidents are negatively correlated.



11. Understanding box plots

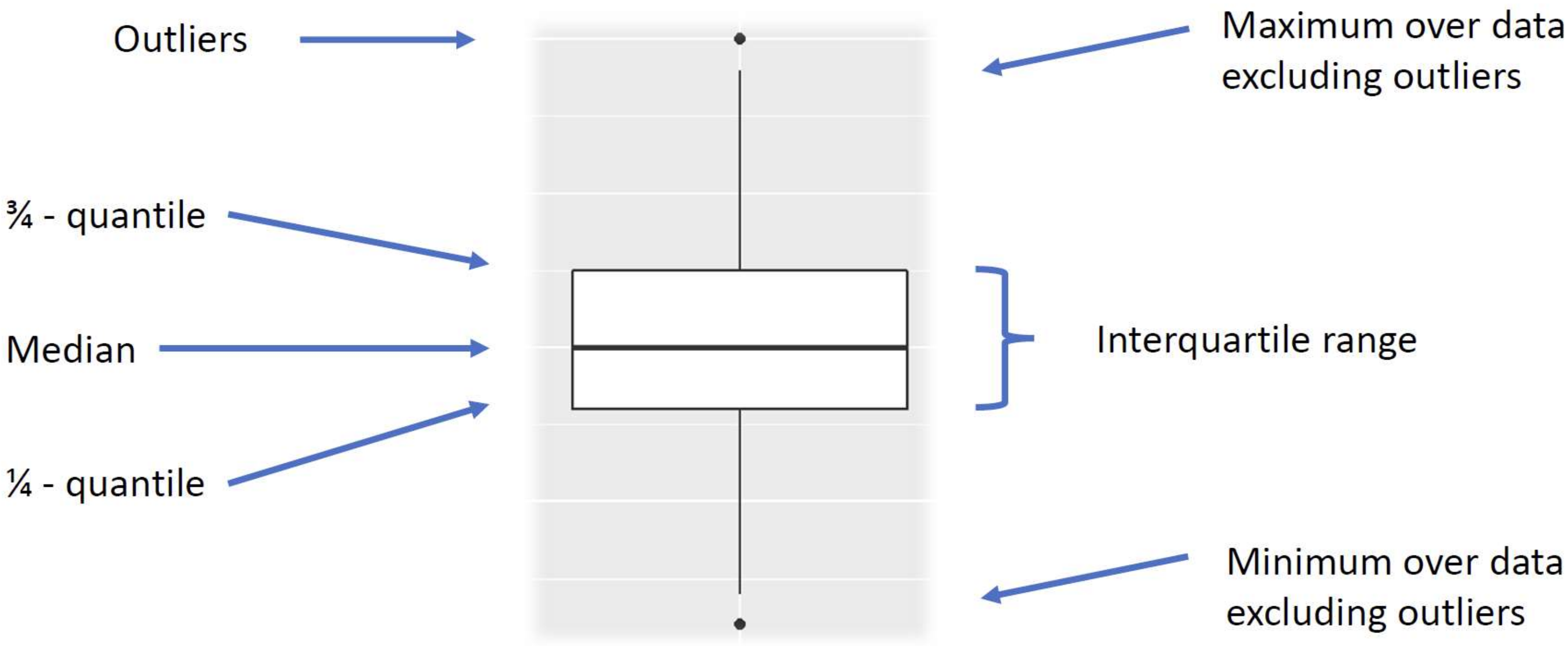
Recall that we used boxplots to study the relationship between two variables:

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



Understanding box plots

How do we interpret box plots?



What have we covered?

We gave a taxonomy of the different types of data

We discussed a wide variety of location estimators (sample mean, median, etc)

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

We also considered several estimators of variability (variance, standard deviation etc).

We learned about how to interpret a boxplot

We introduced correlation as a measure of interdependency between two variables.

University of BRISTOL

Thanks for listening!

Dr. Rihuan Ke
rihuan.ke@bristol.ac.uk

Statistical Computing and Empirical Methods
Unit EMATM0061, MSc Data Science