

Intro and descriptive statistics

The ASTA team

Contents

1	Software	1
1.1	Rstudio	1
1.2	R extensions	1
1.3	R help	2
2	Data	2
2.1	Data example	2
2.2	Data types	2
3	Graphics for quantitative variables	3
3.1	Scatterplot	3
3.2	Histogram	5
4	Summaries of quantitative variables	6
4.1	Percentiles	6
4.2	Boxplot	7
4.3	Measures of center of data: Mean and median	8
4.4	Measures of variability of data: range, standard deviation and variance	8

1 Software

1.1 Rstudio

- Make a folder on your computer where you want to keep files to use in **Rstudio**. **Do NOT use Danish characters æ, ø, å** in the folder name (or anywhere in the path to the folder).
- Set the working directory to this folder: **Session** -> **Set Working Directory** -> **Choose Directory** (shortcut: Ctrl+Shift+H).
- Make the change permanent by setting the default directory in: **Tools** -> **Global Options** -> **Choose Directory**.

1.2 R extensions

- The functionality of **R** can be extended through libraries or packages (much like plugins in browsers etc.). Some are installed by default in **R** and you just need to load them.
- To install a new package in **Rstudio** use the menu: **Tools** -> **Install Packages**
- You need to know the name of the package you want to install. You can also do it through a command:

```
install.packages("mosaic")
```

- When it is installed you can load it through the **library** command:

```
library(mosaic)
```

- This loads the `mosaic` package which has a lot of convenient functions for this course (we will get back to that later). It also prints a lot of info about functions that have been changed by the `mosaic` package, but you can safely ignore that.

1.3 R help

- You get help via `?<command>`:

```
?sum
```

- Use `tab` to make **Rstudio** guess what you have started typing.
- Search for help:

```
help.search("plot")
```

- You can find a cheat sheet with the **R** functions we use for this course here.

2 Data

2.1 Data example

We use data about penguins from the R package `palmerpenguins`

```
pingviner <- palmerpenguins::penguins
pingviner
```

```
## # A tibble: 344 x 8
##   species island  bill_length_mm bill_depth_mm flipper_length_mm body_mass_g
##   <fct>   <fct>      <dbl>         <dbl>         <int>         <int>
## 1 Adelie  Torgersen    39.1          18.7           181           3750
## 2 Adelie  Torgersen    39.5          17.4           186           3800
## 3 Adelie  Torgersen    40.3           18            195           3250
## 4 Adelie  Torgersen    NA             NA              NA             NA
## 5 Adelie  Torgersen    36.7          19.3           193           3450
## 6 Adelie  Torgersen    39.3          20.6           190           3650
## 7 Adelie  Torgersen    38.9          17.8           181           3625
## 8 Adelie  Torgersen    39.2          19.6           195           4675
## 9 Adelie  Torgersen    34.1          18.1           193           3475
## 10 Adelie Torgersen    42            20.2           190           4250
## # i 334 more rows
## # i 2 more variables: sex <fct>, year <int>
```

- What is fundamentally different about the the variables (columns) `species` and `body_mass_g`?

2.2 Data types

2.2.1 Quantitative variables

- The measurements have numerical values.
- Quantative data often comes about in one of the following ways:
 - **Continuous variables**: measurements of time, length, size, age, mass, etc.
 - **Discrete variables**: counts of e.g. words in a text, hits on a webpage, number of arrivals to a queue in one hour, etc.
- Measurements like this have a well-defined scale and in **R** they are stored as the type **numeric**.
- It is important to be able to distinguish between discrete count variables and continuous variables, since this often determines how we describe the uncertainty of a measurement.

- Are any of the measurements in our data set quantitative?
-

2.2.2 Categorical/qualitative variables

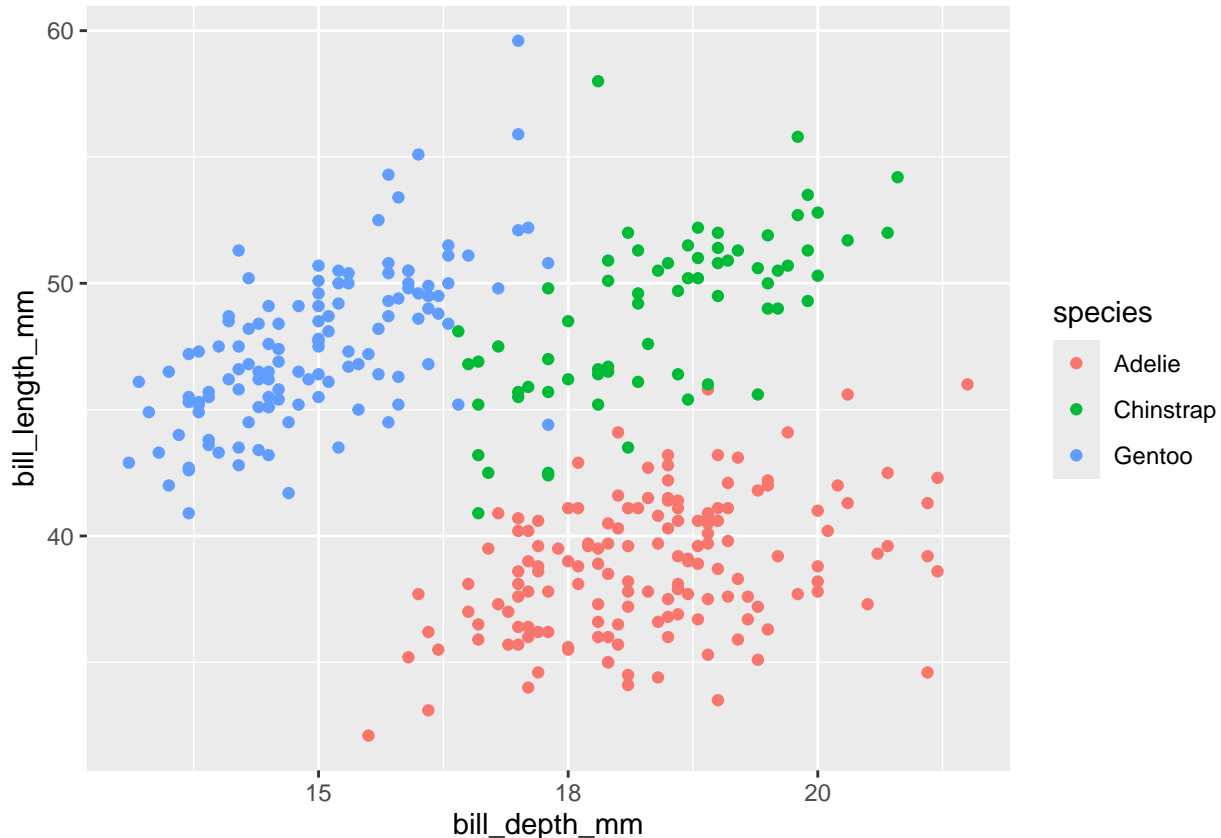
- The measurement is one of a set of given categories, e.g. sex (male/female), social status, satisfaction score (low/medium/high), etc.
- The measurement is usually stored (which is also recommended) as a **factor** in **R**. The possible categories are called **levels**. Example: the levels of the factor “sex” is male/female.
- Factors have two so-called scales:
 - **Nominal scale**: There is no natural ordering of the factor levels, e.g. sex and hair color.
 - **Ordinal scale**: There is a natural ordering of the factor levels, e.g. social status and satisfaction score. A factor in **R** can have a so-called **attribute** assigned, which tells if it is ordinal.
- Are any of the measurements in our data set categorical/qualitative?

3 Graphics for quantitative variables

3.1 Scatterplot

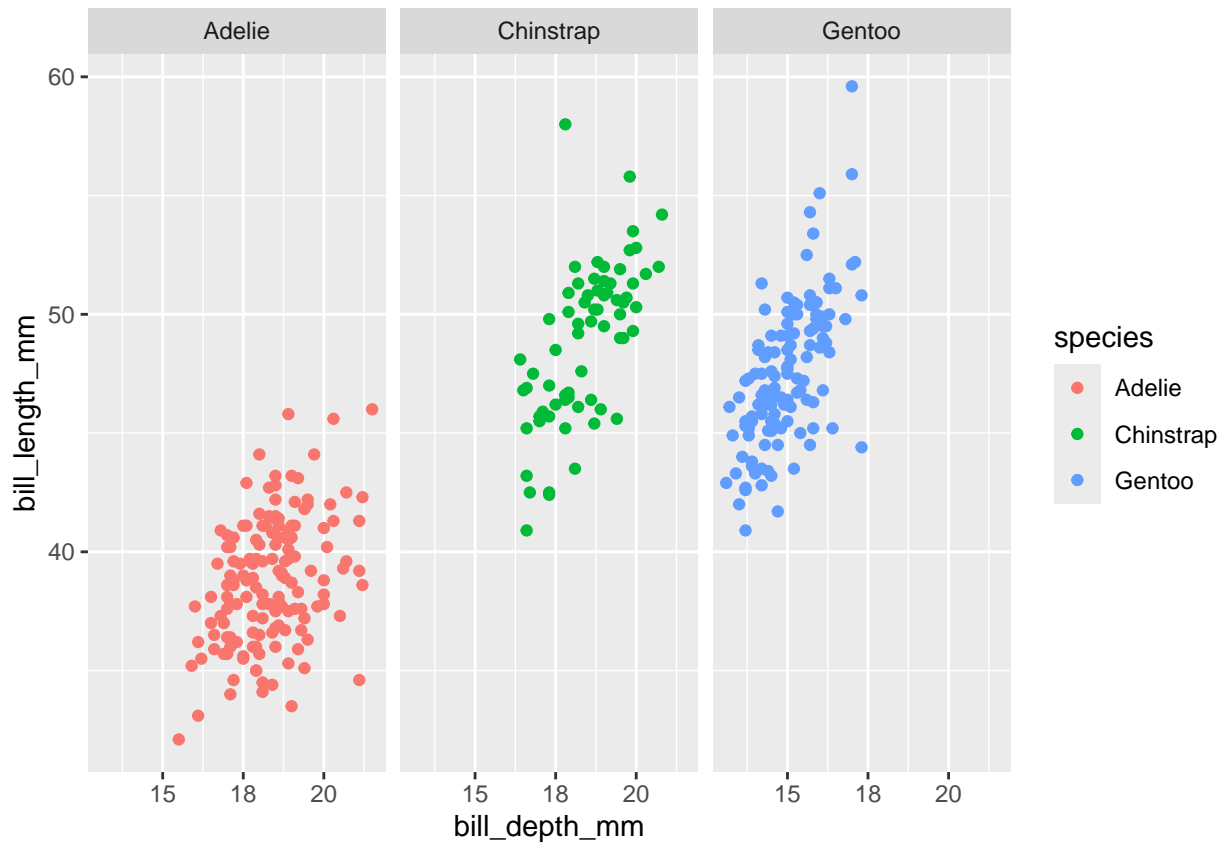
- To study the relation between two quantitative variables a scatterplot is used:

```
gf_point(bill_length_mm ~ bill_depth_mm, color = ~ species, data = penguin)
```



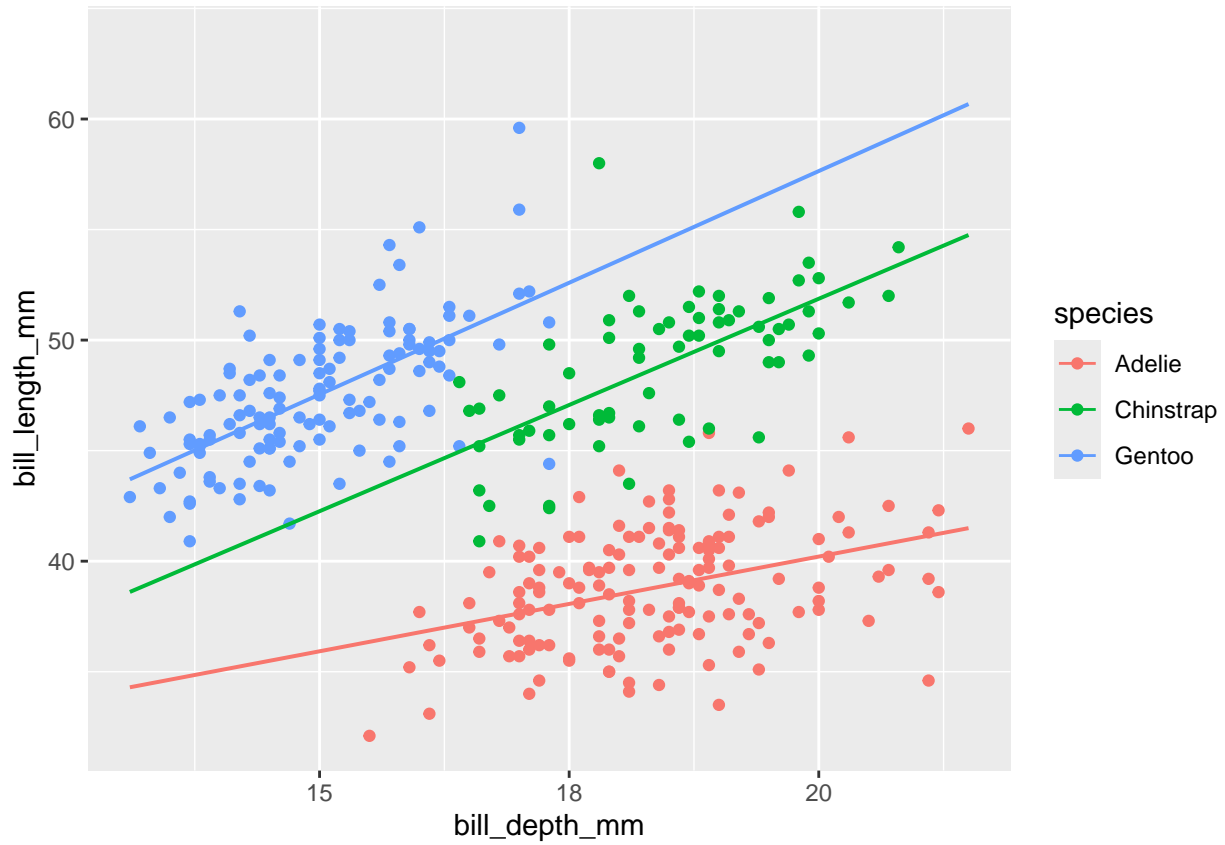
- We could also draw the graph for each species:

```
gf_point(bill_length_mm ~ bill_depth_mm | species, color = ~ species, data = pingviner)
```



- If we want a regression line along with the points we can do:

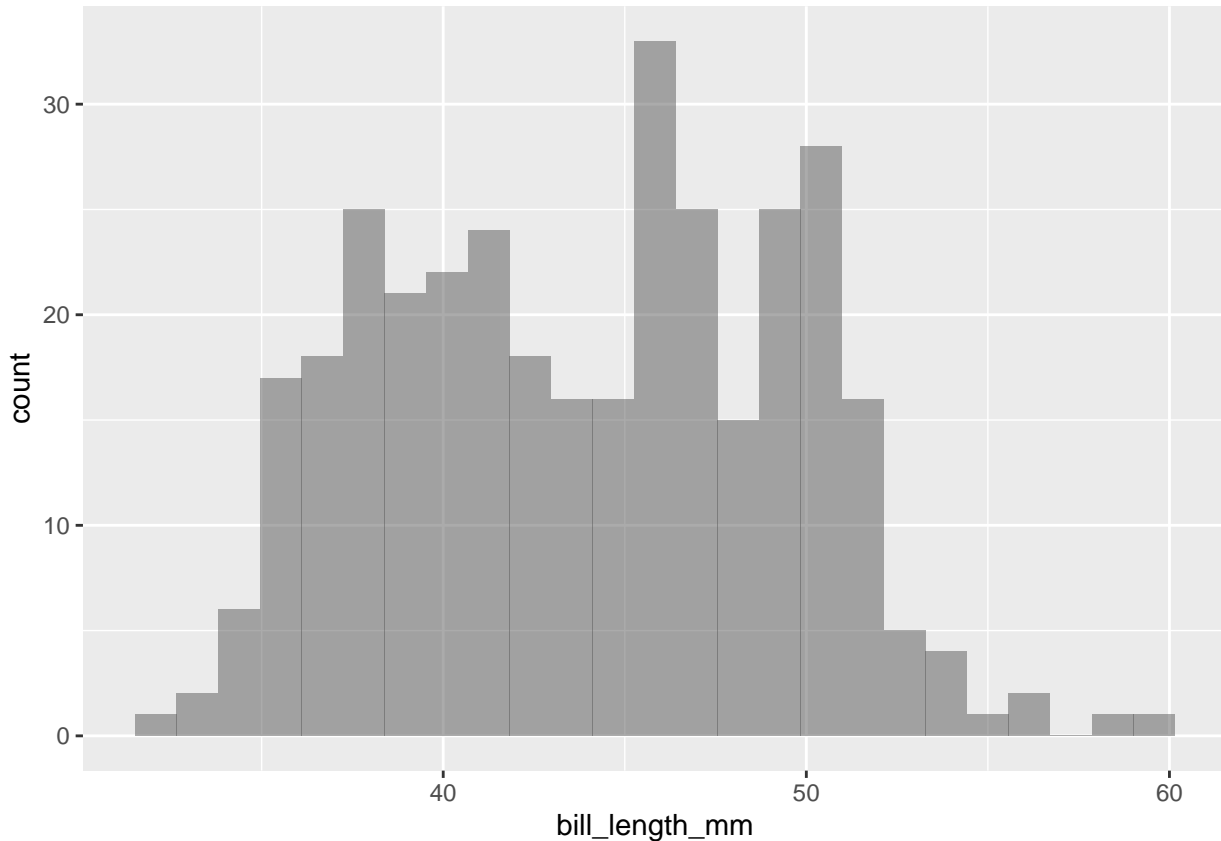
```
gf_point(bill_length_mm ~ bill_depth_mm, color = ~ species, data = pingviner) %>%  
gf_lm()
```



3.2 Histogram

- For a single quantitative variable a histogram offers more details:

```
gf_histogram(~ bill_length_mm, data = pingviner)
```



- How to make a histogram for some variable x :
 - Divide the interval from the minimum value of x to the maximum value of x in an appropriate number of equal sized sub-intervals.
 - Draw a box over each sub-interval with the height being proportional to the number of observations in the sub-interval.

4 Summaries of quantitative variables

4.1 Percentiles

- The p th percentile is a value such that at least $p\%$ of the sample lies below or at this value and at least $(100 - p)\%$ of the sample lies above or at the value.

```
Q <- quantile(bill_length_mm ~ species, data = pingviner, na.rm = TRUE)
Q
```

```
##   species 0% 25% 50% 75% 100%
## 1  Adelie 32 37 39 41 46
## 2 Chinstrap 41 46 50 51 58
## 3  Gentoo 41 45 47 50 60
```

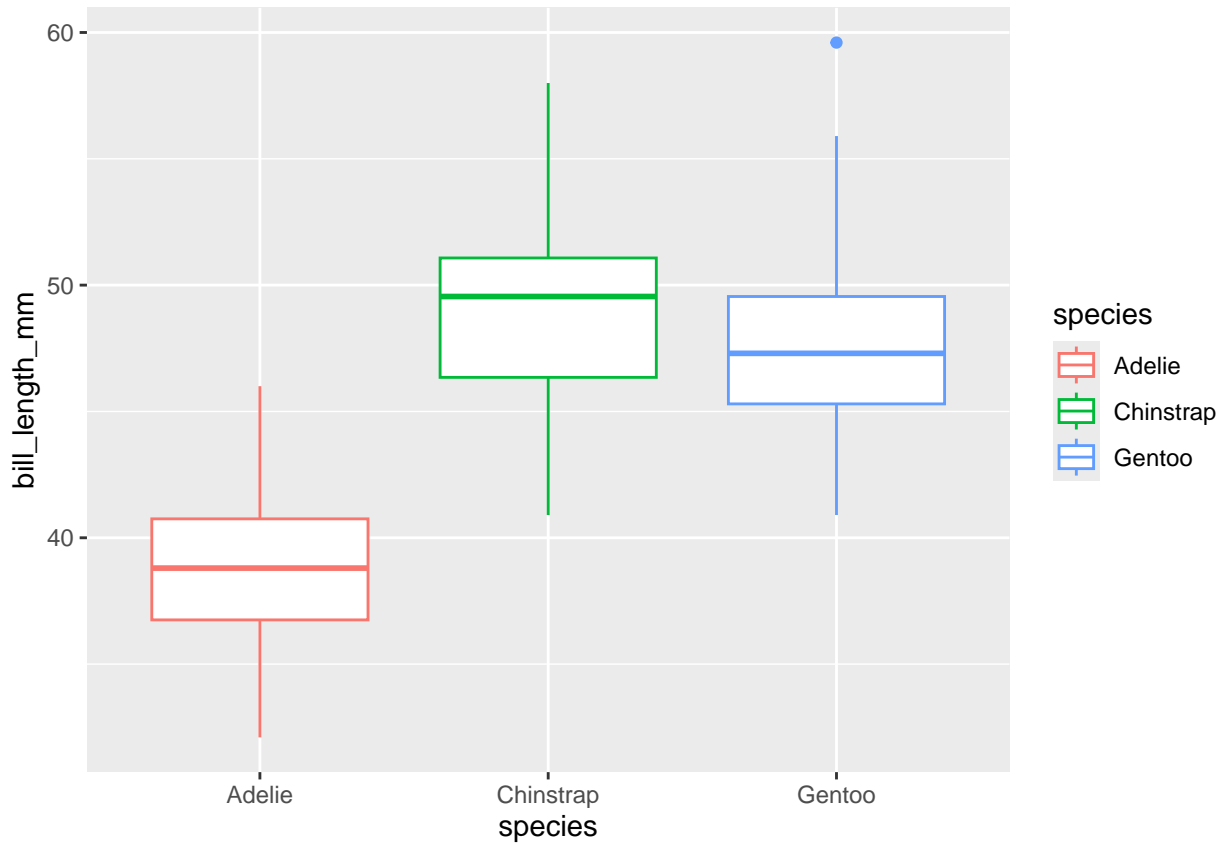
- 50-percentile is the **median** and it is a measure of the center of data as the number of data points below the median is the same as the number above the median.
- 0-percentile is the **minimum** value.
- 25-percentile is called the **lower quartile** (Q1). Median of lower 50% of data.
- 75-percentile is called the **upper quartile** (Q3). Median of upper 50% of data.
- 100-percentile is the **maximum** value.

- **Interquartile Range (IQR)**: a measure of variability given by the difference of the upper and lower quartiles.

4.2 Boxplot

Boxplot can be good for comparing groups (notice we put the values on the y-axis here as it is more conventional for boxplots):

```
gf_boxplot(bill_length_mm ~ species, color = ~ species, data = pingviner)
```

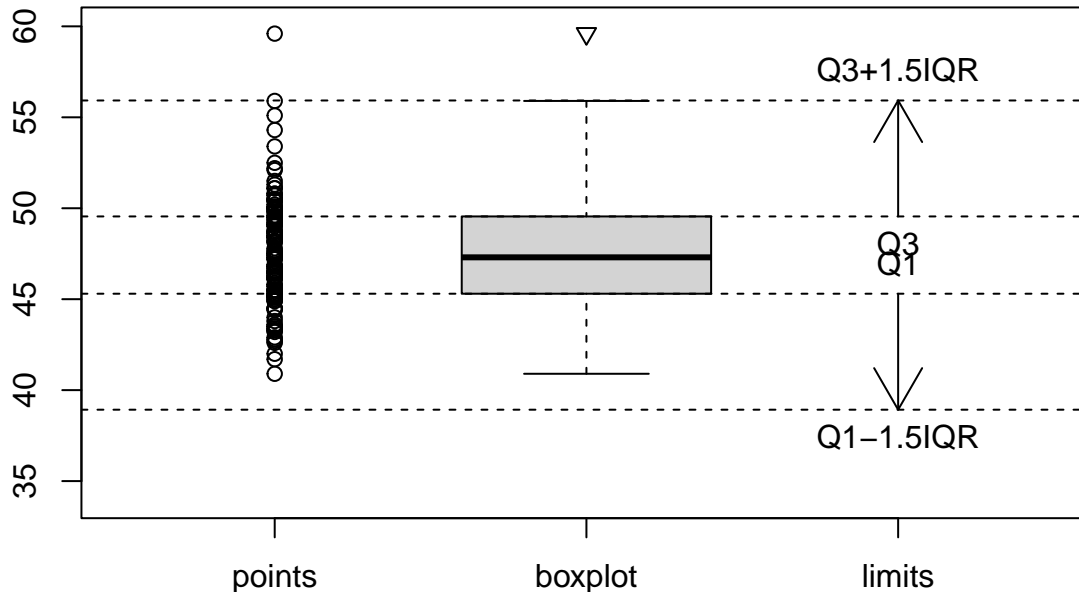


4.2.1 How to draw a box plot

- Box:
 - Calculate the median, lower and upper quartiles.
 - Plot a line by the median and draw a box between the upper and lower quartiles.
- Whiskers:
 - Calculate interquartile range and call it IQR.
 - Calculate the following values:
 - * $L = \text{lower quartile} - 1.5 \cdot \text{IQR}$
 - * $U = \text{upper quartile} + 1.5 \cdot \text{IQR}$
 - Draw a line from lower quartile to the smallest measurement, which is larger than L .
 - Similarly, draw a line from upper quartile to the largest measurement which is smaller than U .
- Outliers: Measurements smaller than L or larger than U are drawn as circles.

Note: Whiskers are minimum and maximum of the observations that are not deemed to be outliers.

Gentoo bill length



4.3 Measures of center of data: Mean and median

- A number of numerical summaries can be retrieved using the `favstats` command:

```
favstats(bill_length_mm ~ species, data = pingviner)
```

```
##   species min Q1 median Q3 max mean sd  n missing
## 1  Adelie  32 37   39 41  46  39 2.7 151      1
## 2 Chinstrap 41 46   50 51  58  49 3.3  68      0
## 3  Gentoo  41 45   47 50  60  48 3.1 123      1
```

- The observed values of `bill_length_mm` are $y_1 = 46.1, y_2 = 50, \dots, y_n = 49.9$, where there are a total of $n = 123$ values.

As previously defined this constitutes a **sample**.

- mean** = 48 is the **average** of the sample, which is calculated by

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i.$$

We may also call \bar{y} the **(empirical) mean** or the **sample mean**. It is calculated using `mean()` in **R**.

- median** = 47 is calculated using `median()` in **R**.
- An important property of the **mean** and the **median** is that they have the same unit as the observations (e.g. millimeter).

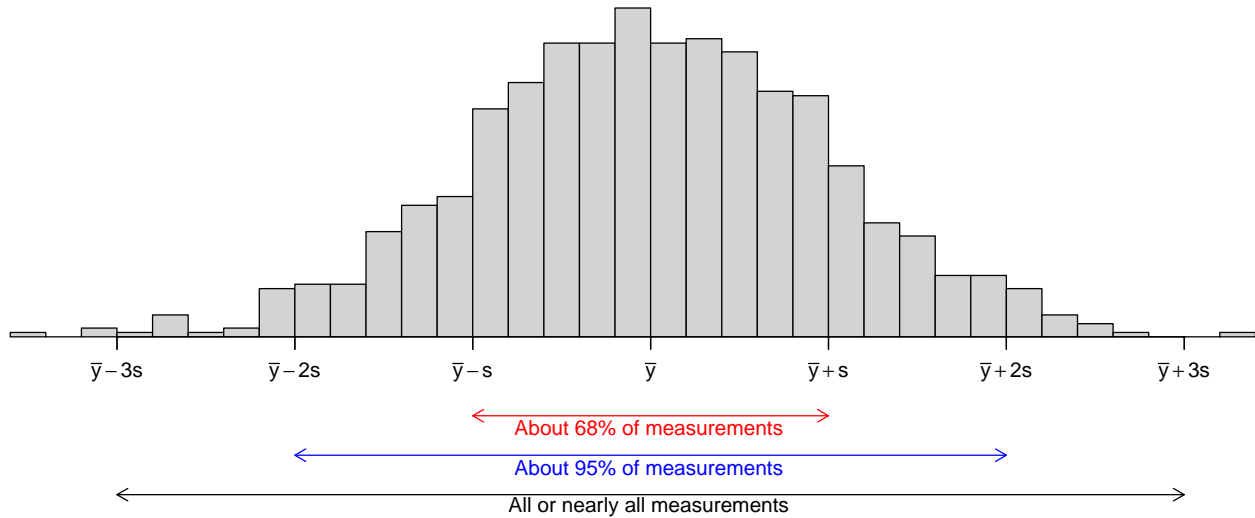
4.4 Measures of variability of data: range, standard deviation and variance

- The **range** is the difference of the largest and smallest observation (`range()` in **R**).
- The **(empirical) variance** (`var()` in **R**) is the average of the squared deviations from the mean:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2.$$

- **sd = standard deviation** = $s = \sqrt{s^2}$ (**sd()** in **R**).
- Note: If the observations are measured in mm, the **variance** has unit mm^2 which is hard to interpret. The **standard deviation** on the other hand has the same unit as the observations.
- The standard deviation describes how much data varies around the (empirical) mean.

4.4.1 The empirical rule



If the histogram of the sample looks like a bell shaped curve, then

- about 68% of the observations lie between $\bar{y} - s$ and $\bar{y} + s$.
- about 95% of the observations lie between $\bar{y} - 2s$ and $\bar{y} + 2s$.
- All or almost all (99.7%) of the observations lie between $\bar{y} - 3s$ and $\bar{y} + 3s$.