Introduction to Data

Owen G. Ward

2021-05-03

Introduction

A natural question if you have never really studied statistics before is what exactly do we use it for? What is the point of statistics?

Thankfully, there are lots of good examples for this! One immediate example where statistics is used is in developing a Covid vaccine. If you create a new vaccine which you think will help stop Covid, how do you convince people it works, or that it is even safe to take? All of these questions are answered using statistics.

One historic example is the development of the first Polio vaccine in the 1950's. A subset of the data from this study is shown below.

```
##        Group Population Paralytic
## 1 Vaccinated     200745        33
## 2    Placebo     201229       115
```

It seems quite clear that less people became Paralytic among the group who received the vaccine, but how can we say this with any certainty? To do that, the we need to use the language of probability and statistics.

Statistics allows us to make decisions with data. There are two general types of decisions we can use statistics for.

Inferential statistics allows us to infer something unobservable. For example, how much would someone's income increase if they got a college degree. Or, are you more likely to get Covid if you don't get vaccinated.

Predictive statistics allow us to estimate future or unknown quantities. For example, how much can you expect to make 5 years after taking this class?

So, on a very simple level, we want to be able to get some data and use it to understand relationships. But that raises even more questions. What do we even mean by data? How do we get data?

Data

For the previous polio example, the original data consisted of hundreds of thousands of children being recruited for the study. For each child we know whether they were vaccinated or got a placebo (a vaccine with nothing in it). Then, we know eventually if these children became paralytic or not.

So the data above is really a summary of the raw data, a sample of which would look something like this.

```
## # A tibble: 4 x 2
##   Group      Paralytic
##   <chr>      <chr>
## 1 Vaccinated No
## 2 Placebo    No
## 3 Placebo    Yes
## 4 Vaccinated No
```

Here we have rows of data, with each row being a single **observation** (patient in the study). For each observation we observe two things, whether they received the vaccine or not, and whether they became paralytic. These are the **variables** we observe.

Here the two variables we observe are **binary**, they can each only take on two values. In general we can have lots of different types of variables.

For example, here is some data which was collected about penguins. Here researchers went to Antarctica and collected this directly.

```
## # A tibble: 6 x 8
##   species island bill_length_mm bill_depth_mm flipper_length_~ body_mass_g sex
##   <fct>   <fct>           <dbl>         <dbl>            <int>       <int> <fct>
## 1 Adelie  Torge~           39.1          18.7              181        3750 male
## 2 Adelie  Torge~           39.5          17.4              186        3800 fema~
## 3 Adelie  Torge~           40.3          18                195        3250 fema~
## 4 Adelie  Torge~           NA            NA                 NA          NA <NA>
## 5 Adelie  Torge~           36.7          19.3              193        3450 fema~
## 6 Adelie  Torge~           39.3          20.6              190        3650 male
## # ... with 1 more variable: year <int>
```

We have several variables here which display the different types of variables we can observe.

- Numerical, which can be either continuous or discrete (int or dbl).
- Categorical, which can be either nominal (unordered) or ordinal (ordered).

Visualizing and Summarizing Data

While looking at the raw data is often important to explore and understand properties of the data, we often want to summarize the data in a more compact way.

How we summarize the data will depend on what properties of the data we think are important.

For example, we could use a numeric summary of the data to capture some overall general properties, while a graphical visualisation may be more useful to capture other properties.

Numeric Summaries

For the Polio data, we could summarize the raw data by simply getting the counts of how many people who got the vaccine were Paralytic and how many weren't. This is because we are only considering two variables which can only take on two values each. So there are only 4 possible values for each observation.

But what can we do if we want to summarise the weight of a penguin? Counting all the penguins who weigh 3750g or 3250g maybe doesn't make much sense. This would lead to a massive table, not a useful summary.

For numeric data we often want to reduce the amount of information we have to consider. If we have the weight of several hundred penguins, and we want to give someone a general summary that describes this data, we want to reduce this down to considerably fewer numbers.

- We are often interested in reducing this information down to one or two numbers which capture important properties of the variable.

The maximum reduction in information would be to use a single number to summarise this variable. There are several possibilities, but one commonly used is the average, also known as the **mean**.

If we have weights $x_1, x_2, \ldots, x_n$ then this is computed by

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

If we wanted to ask "Do male Adelie penguins weigh more than female Adelie penguins?", we need to be able to compare these numbers.

We could get the mean of each

```
## # A tibble: 2 x 2
##   sex    avg_weight
##   <fct>       <dbl>
## 1 female      3369.
## 2 male        4043.
```

An alternative one number summary of data such as this would be the **median** weight.

The median splits the data in half, such that 50% of the data is less than that value and 50% is greater than it.

In general this is not equal to the mean value.

What if a single variable is not enough. Could have two datasets which have the same mean but are vastly different.

For example consider the datasets

$$X = \{0, 0, 0, 0, 0, 1, -1\}, Y = \{4, -10, -10, 10, 6, 5, -3, -2\}.$$

The mean of of both these datasets is 0 but at the same time they are very different. Simply saying they have the same mean is not really sufficient. We need a more informative summary.

To do this we need to consider further properties of the data, such as the **variance**.

This is closely related to the mean, captures how the data is spread about the average value, often using the notation σ^2.

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Note: This is always a non negative number.

Can also consider $\sigma = \sqrt{\sigma^2}$, known as the **standard deviation**.

For the above example, the X data has variance 0.333 while the Y data has variance 55.71.

Note that for small data it is easy enough to compute these statistics by hand, but when we have real data (eg, 350 penguins), you want to use a computer to do these things.

A related notion of the variability of data is the **Inter Quartile Range** (IQR), which is based on **percentiles** of the data.

The 25th percentile, or first quartile, is a number Q_1 such that 25% of the data are less than that number.

The 75th percentile, or third quartile, is a number Q_3 such that 75% of the data are less than that number.

Then, the IQR is

$$IQR = Q_3 - Q_1$$

A large IQR value indicates larger variability in the data.

All the numeric summaries we have seen so far deal with a single variable. What if we want to summarize something more complex, such as the relationship between the height and weight

Want a numerical way of quantifying the relationship between two variables.

One statistic which has been developed to do this is the **covariance** between two variables, where for $x_1, x_2, \ldots, x_n, y_1, y_2, \ldots, y_n$

$$\sigma_{x,y} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}).$$

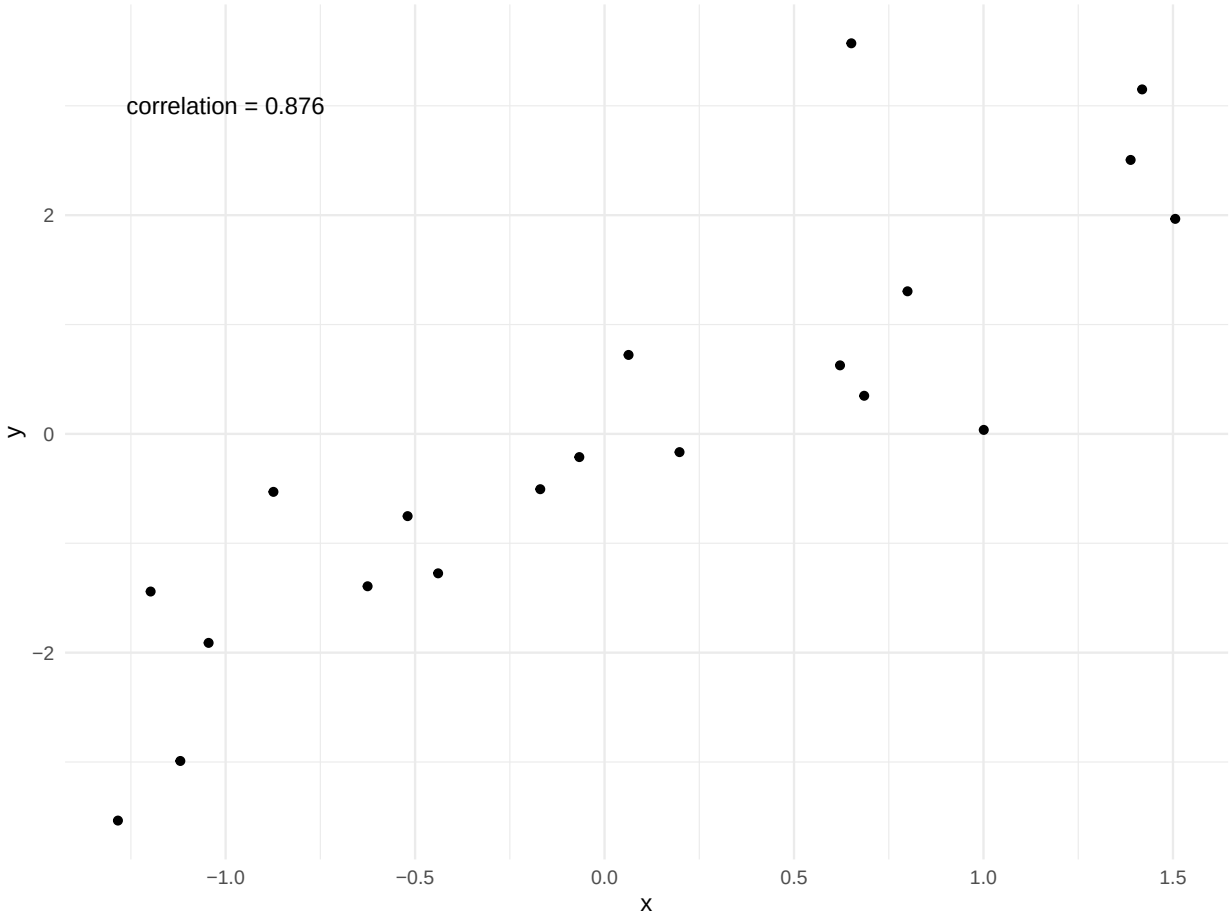
This can take on positive and negative values.

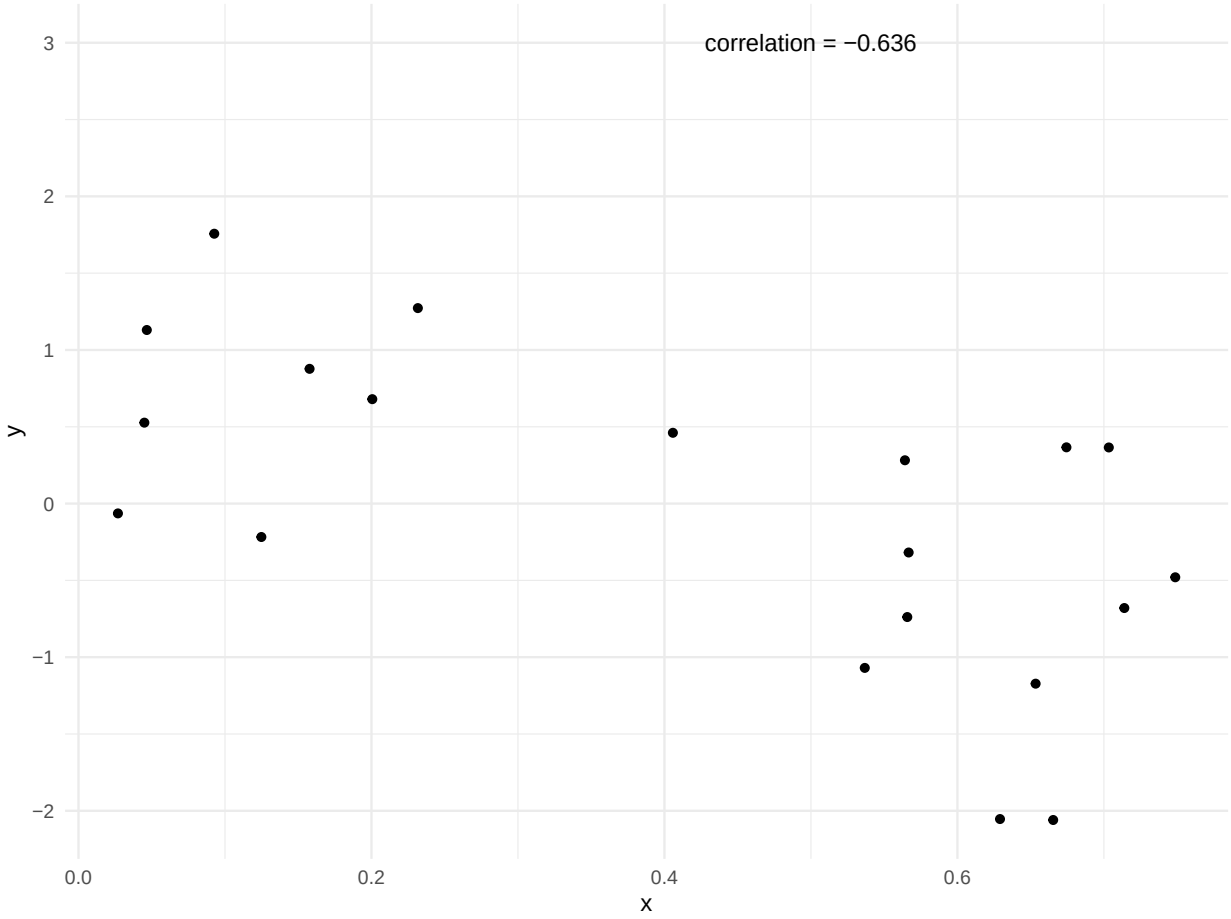
We can also look at the **correlation** of X, Y where

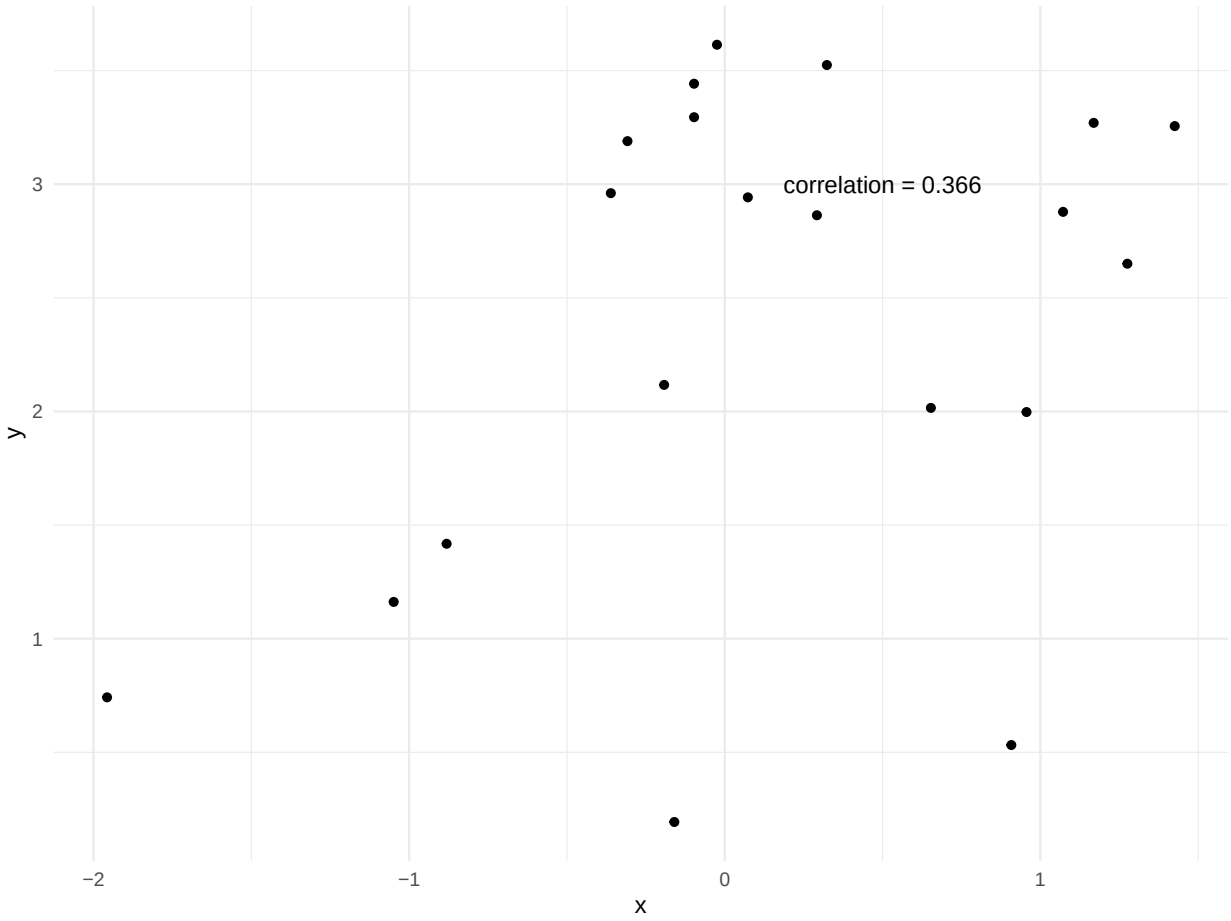
$$corr_{x,y} = \frac{\sigma_{x,y}}{\sigma_x \sigma_y}.$$

This is the covariance divided by the individual standard deviations. This scales the covariance so that we always have $-1 \leq corr_{x,y} \leq 1$.

Covariance and correlation will always have the same sign.




correlation = −0.636
3
2
1
0
−1
−2
y
0.0
0.2
0.4
0.6
x



We can compute the correlation between the body mass and flipper length of the penguins.

```
## # A tibble: 1 x 1
##   cor_mass_flip
##           <dbl>
## 1         0.871
```

Correlation just means there is **some** relationship between two variables, it doesn't tell you if there is a **causal** relationship (more later).

Graphical Summaries

Instead of trying to reduce all the information for one variable down into one or two numbers, we could instead visualize the data as a way of summarising it.

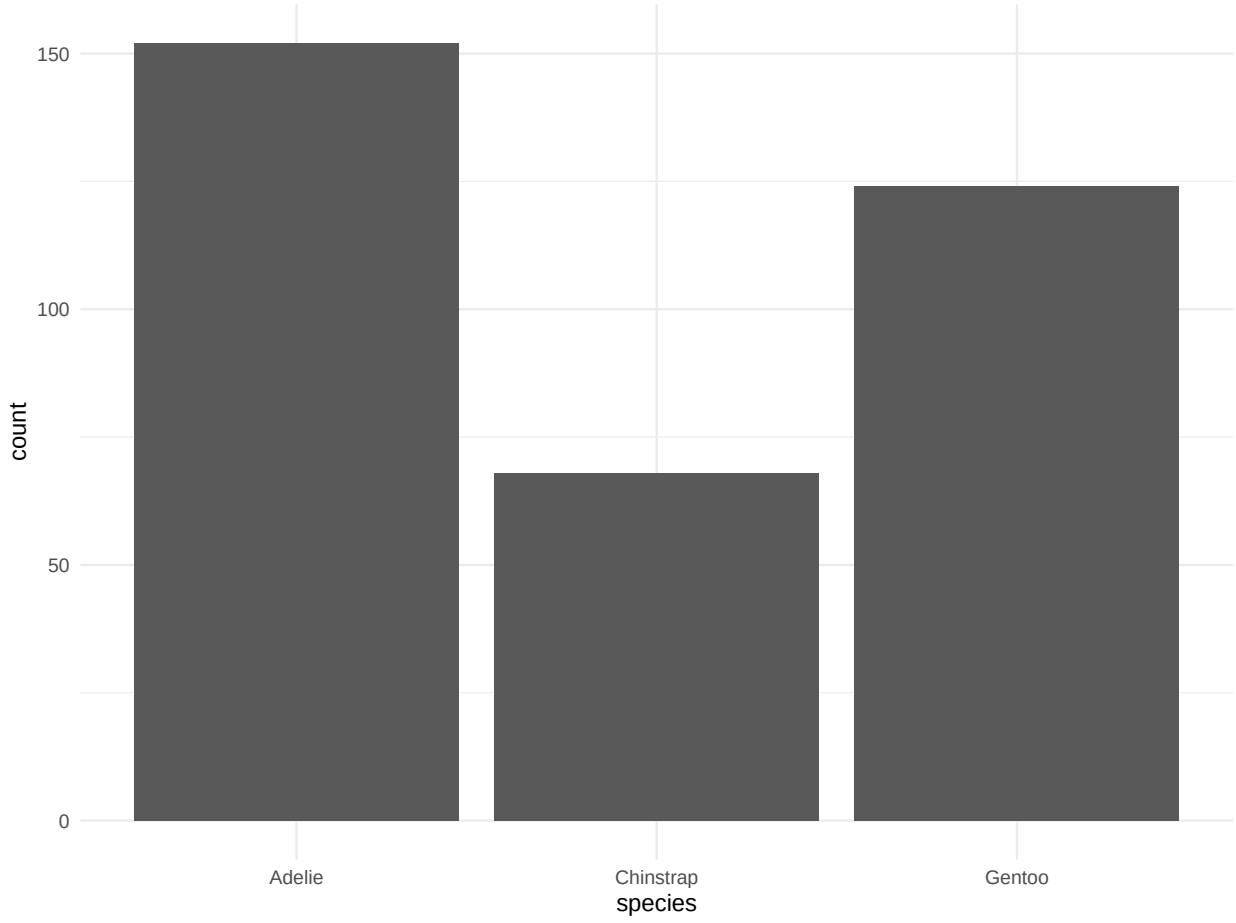
While this is not as succinct as using the mean and variance, it provides vastly more information and can help to identify important components of the data.

- Visualizations can help identify problems with our data.

- Visualizations are more informative than numeric summaries when we wish to summarise more variables or the relationship between multiple variables.

The simplest such example is a bar plot for categorical data, which is equivalent to a table.

For example, we can do a bar plot of the penguin species.

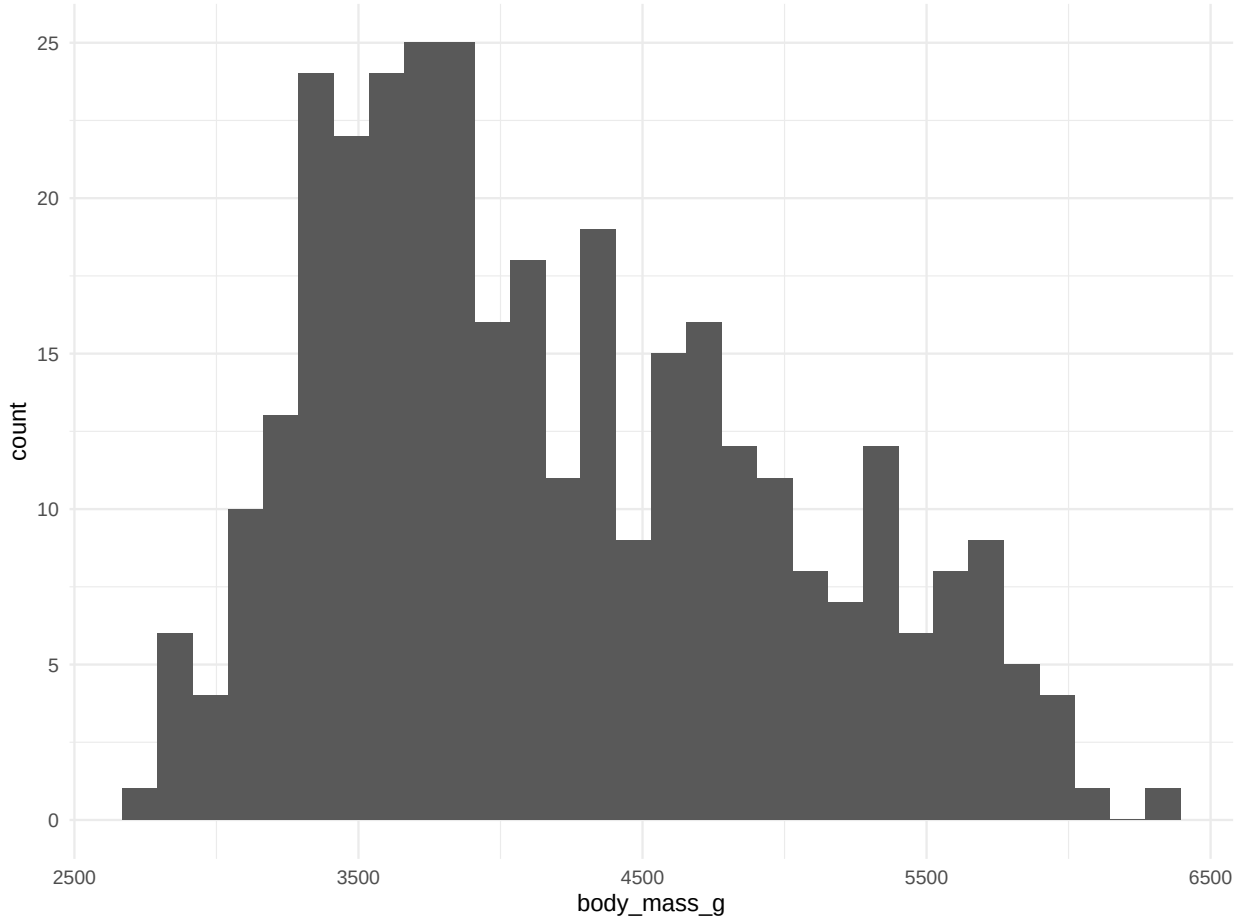


Single continuous variable

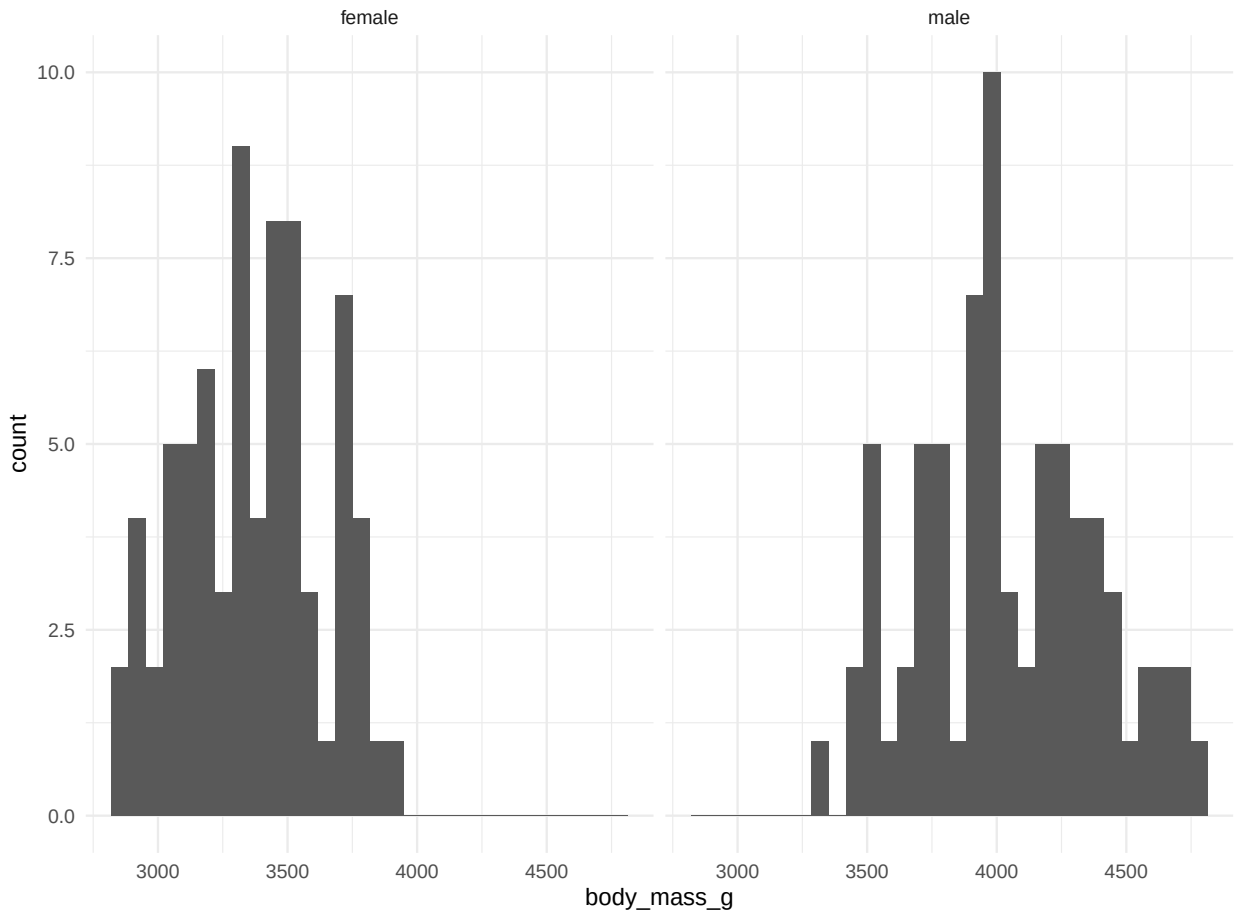
For a single continuous variable we can construct a **histogram** of the data.

To do this we break the data into several bins, plot how many values occur in each bin.

This gives us a measure of the overall spread of the data. Can also be used to identify unusual values.



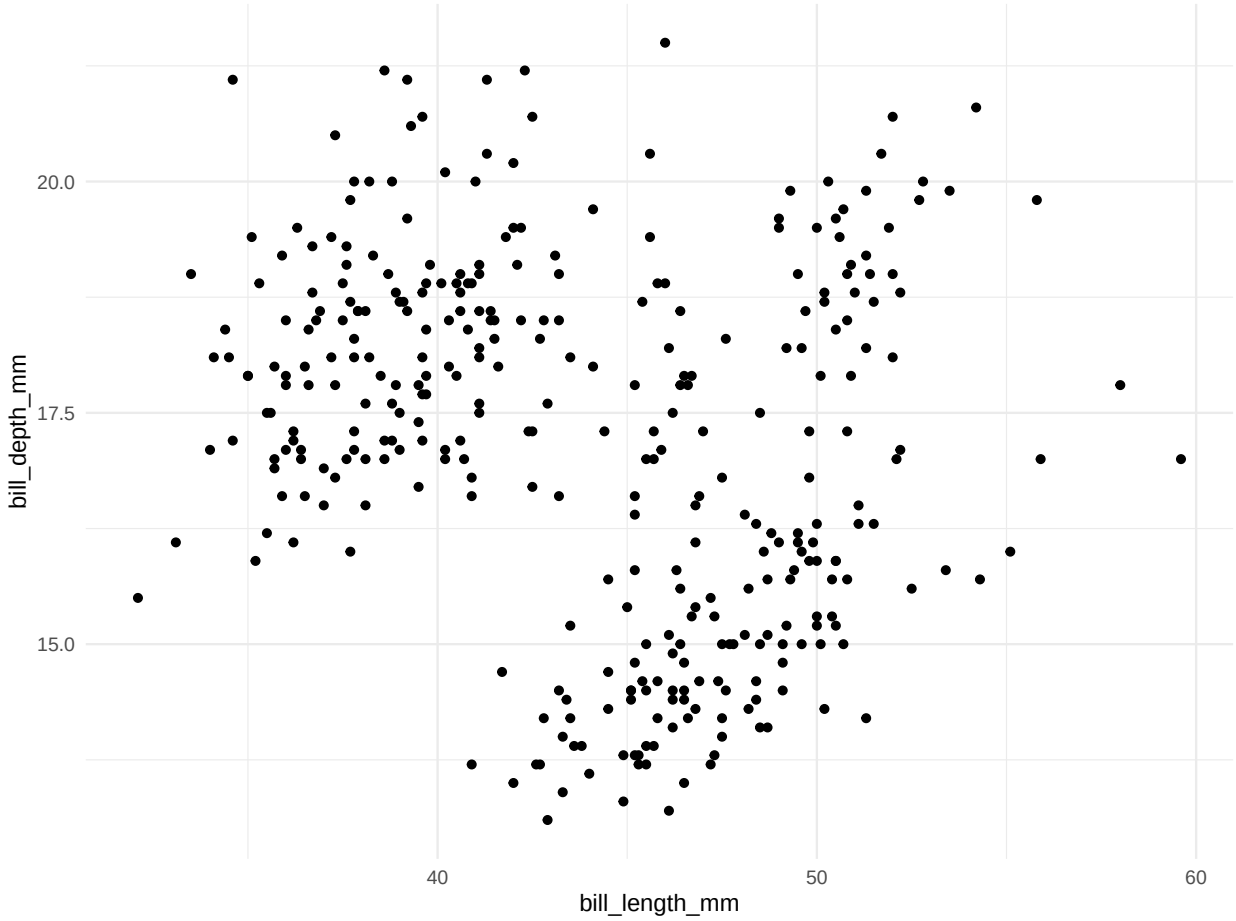
Similarly, we can compare the weight of male and female penguins by looking at a histogram for each.



Two Variables

For examining the relationship between two variables we can use a scatter plot (seen above), where we plot one variable on each axis and use a point to illustrate each observation.

This gives us some insight into how both variables vary together.

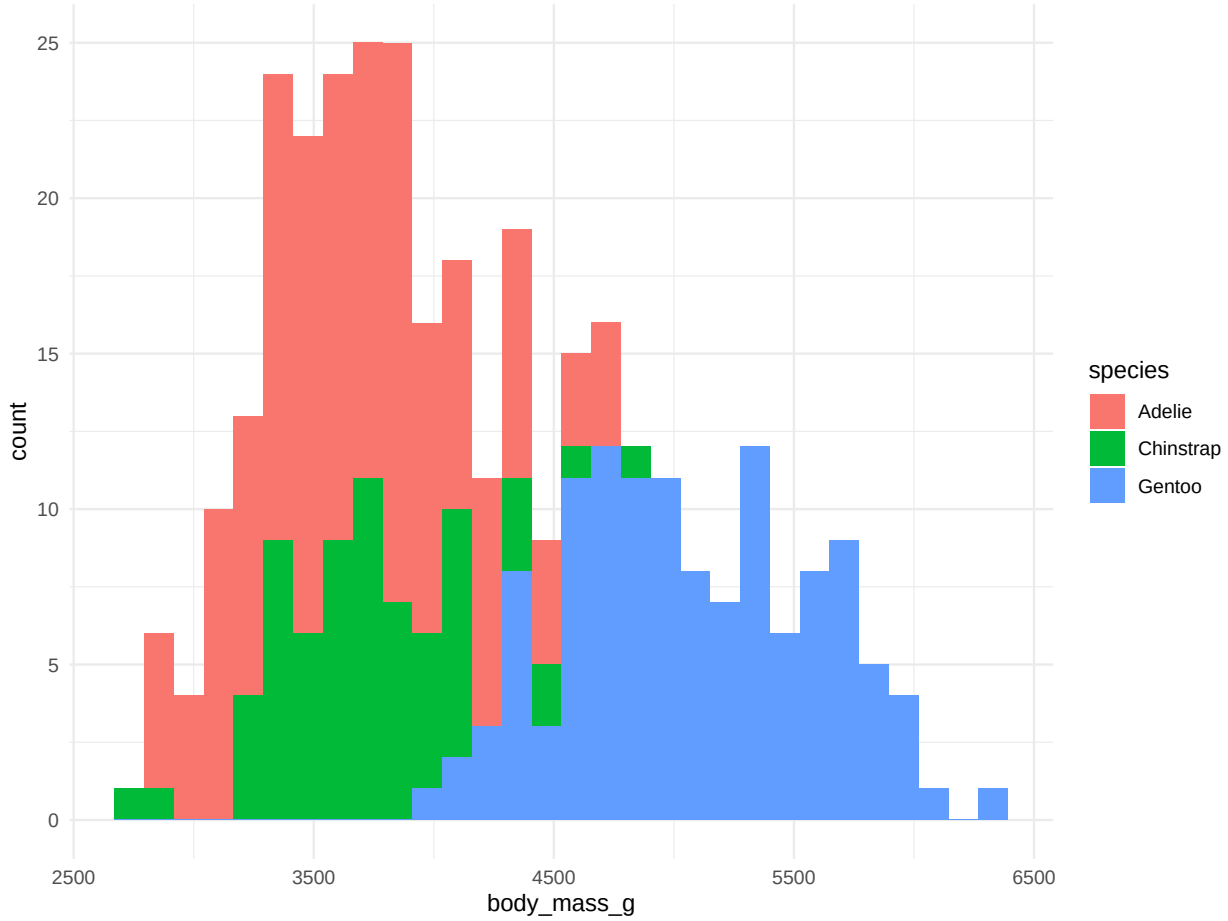


Continuous and categorical

Often, we will want to explore these numeric relationships across some categorical groups. For example, we know there are multiple penguin species in this data. How can we include that in these plots?

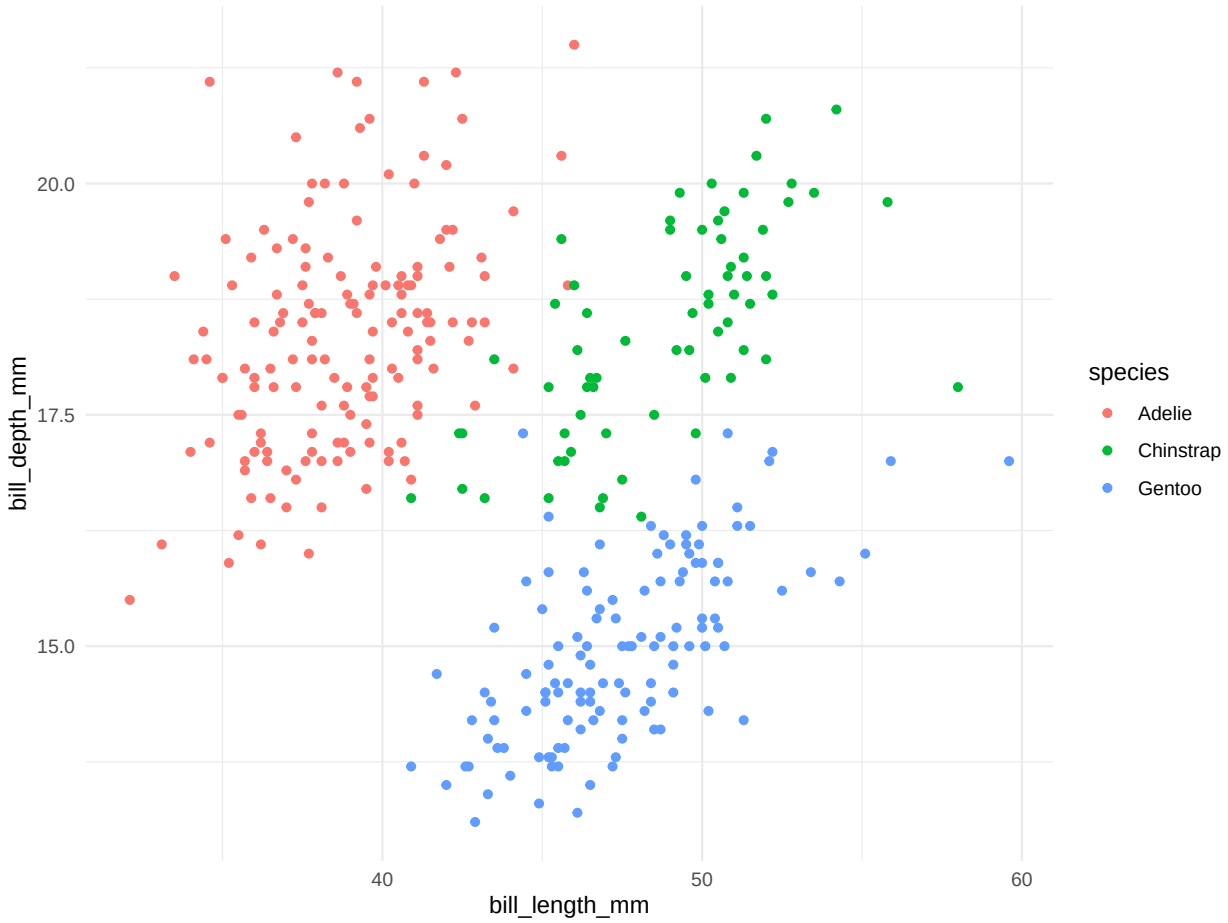
Could create a separate plot for each species or indicate the different species on the one plot, eg using colour.

Let's see what happens with the histogram.



For the histogram it is difficult to deal with the overlap in a single variable, so 3 separate plots might be clearer.

However, using colour works well for the scatterplot.



Visualizing data can be difficult and it is important to strike the balance between detail and clarity. We will see how to make these sorts of plots and compute numeric summaries later.