# An introduction to statistics and data science

Owen Ward

2021-02-26

A question that comes to mind pretty quickly when you study something, particularly if it is something you don't like, is "Why do we need to study this, what's the point?" Statistics and data science is no different. What is the point of statistics? What can we use it for?

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.

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
##   <fct>   <fct>              <dbl>         <dbl>
## 1 Adelie  Torgersen           39.1          18.7
## 2 Adelie  Torgersen           39.5          17.4
## 3 Adelie  Torgersen           40.3          18
## 4 Adelie  Torgersen           NA            NA
## 5 Adelie  Torgersen           36.7          19.3
## 6 Adelie  Torgersen           39.3          20.6
## # ... with 4 more variables: flipper_length_mm <int>,
## #   body_mass_g <int>, sex <fct>, 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
- 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.

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.

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.

For numeric data we often want 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.
```

Similarly, we can compare the weight of each.