Data Analysis

Steph de Silva

11/11/2017

In order to really learn about data analysis, my view is that it should be something *interesting* you’re learning it on. The yhat blog put out an article awhile back (http://blog.yhat.com/posts/7-funny-datasets.html) called *7 Datasets You’ve Likely Never Seen Before* and it’s a goldmine of *interesting*.

One of the datasets suggested in that post is the Medieval English Crop Yields (1211-1491) by Bruce Campbell (2007). You can find out about this dataset here. (http://www.cropyields.ac.uk/index.php)

What’s more interesting than Medieval crop yields?

The Bishop of Exeter’s Clyst

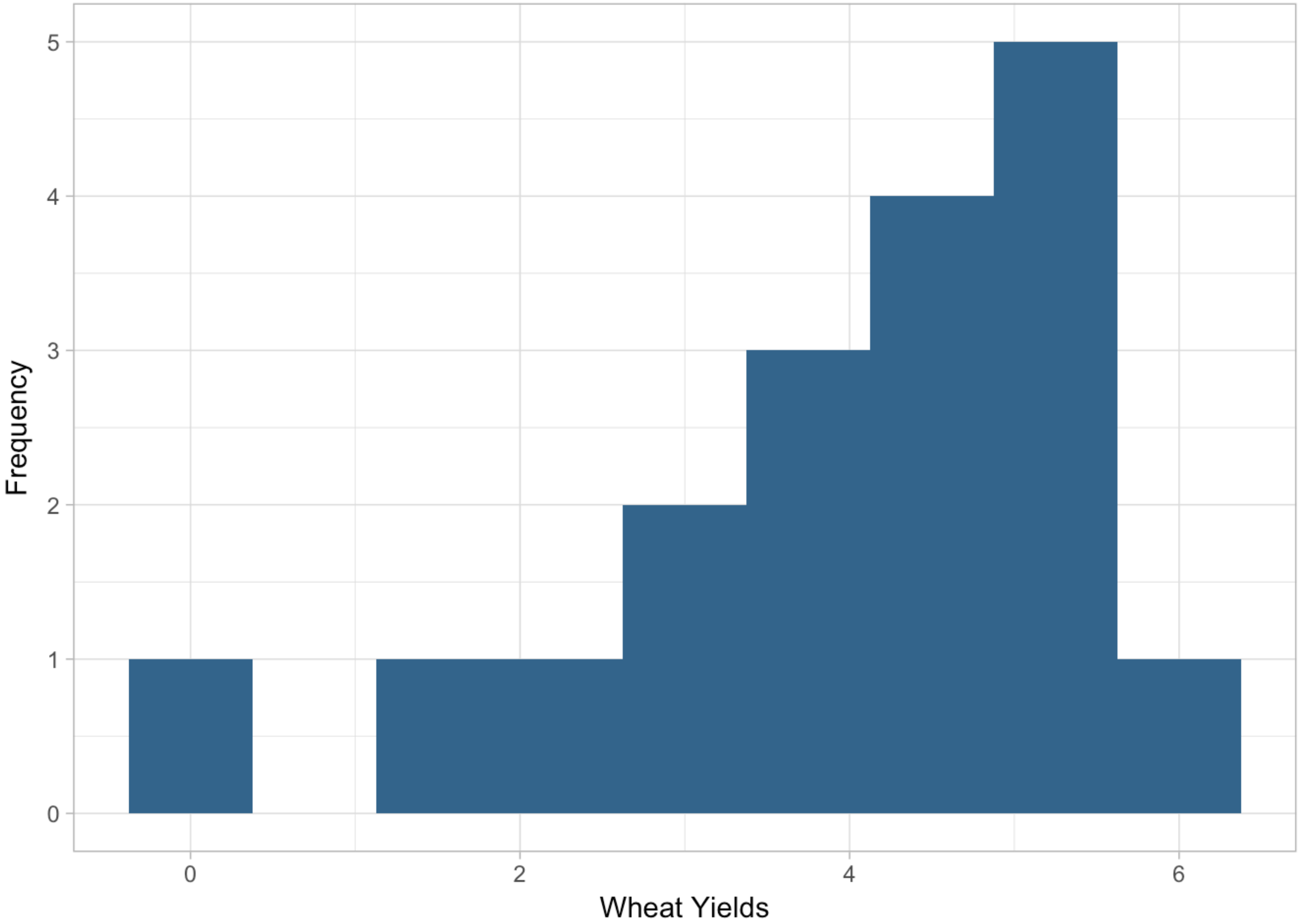
In Devonshire, England, during the 14th and 15th centuries, meticulous records of crop yields were kept. The manor planted wheat, rye, barley and oats and the yields of each (e.g. how much the harvested compared to how much the sowed) were recorded in many years. Let’s take a look at that data.

```
##   County          Manor           Estate Year Wheat  Rye Barley Oats
## 1  Devon Bishop`s Clyst Bishop of Exeter 1372  0.00 5.01   5.33 3.35
## 2  Devon Bishop`s Clyst Bishop of Exeter 1373  2.73 1.62   0.10 1.15
## 3  Devon Bishop`s Clyst Bishop of Exeter 1374    NA   NA     NA   NA
## 4  Devon Bishop`s Clyst Bishop of Exeter 1377  4.26 3.57   2.47 3.74
## 5  Devon Bishop`s Clyst Bishop of Exeter 1378    NA   NA     NA   NA
## 6  Devon Bishop`s Clyst Bishop of Exeter 1380  3.79 4.78     NA 2.99
```

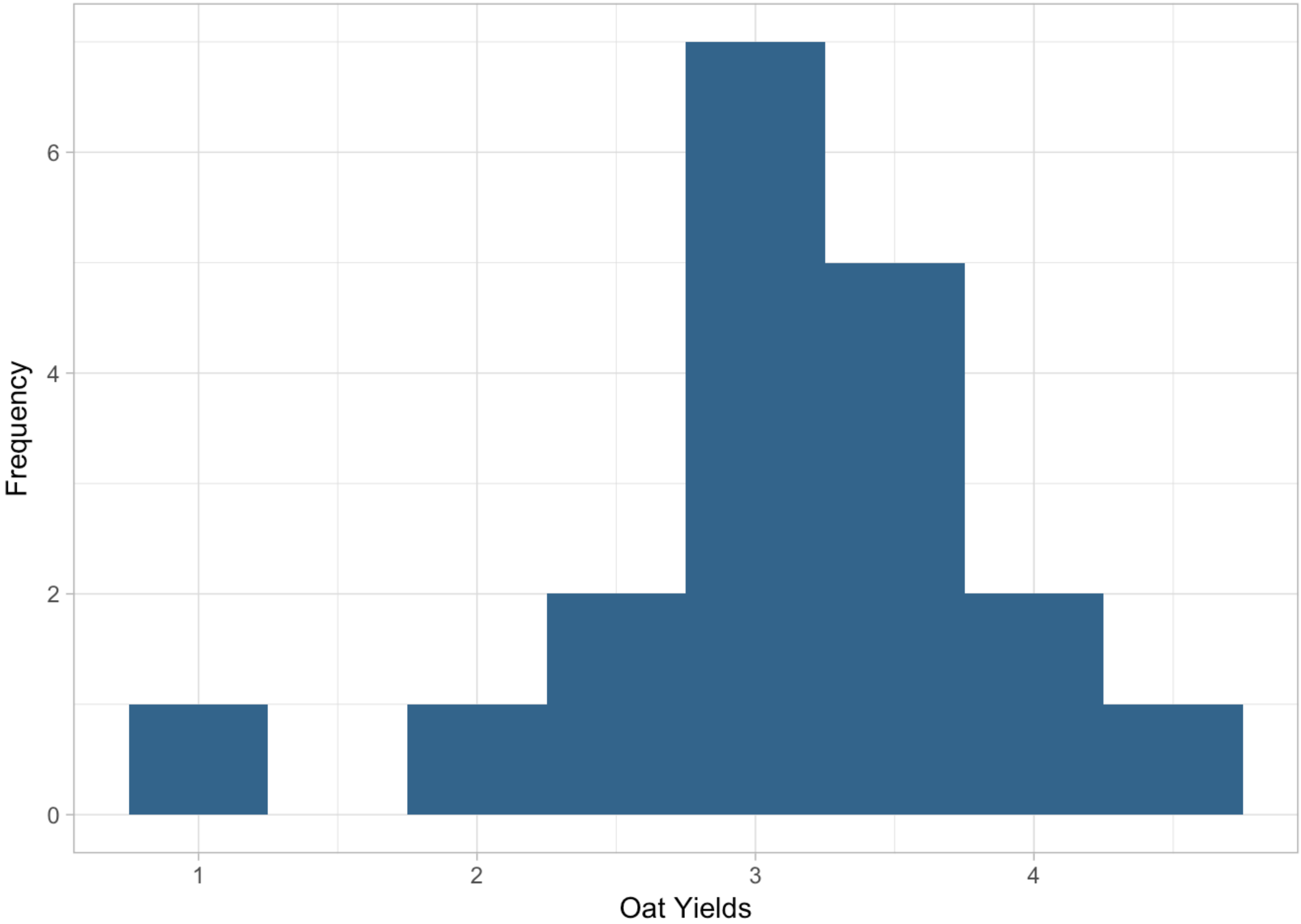
You will note the presence of `NA` in some columns. This just means the data is missing - either it wasn’t recorded that year, the records were lost, the farms weren’t planted because everyone was dying of the Black Death or in some awful civil war or something like that.

Analysing data by looking at shapes

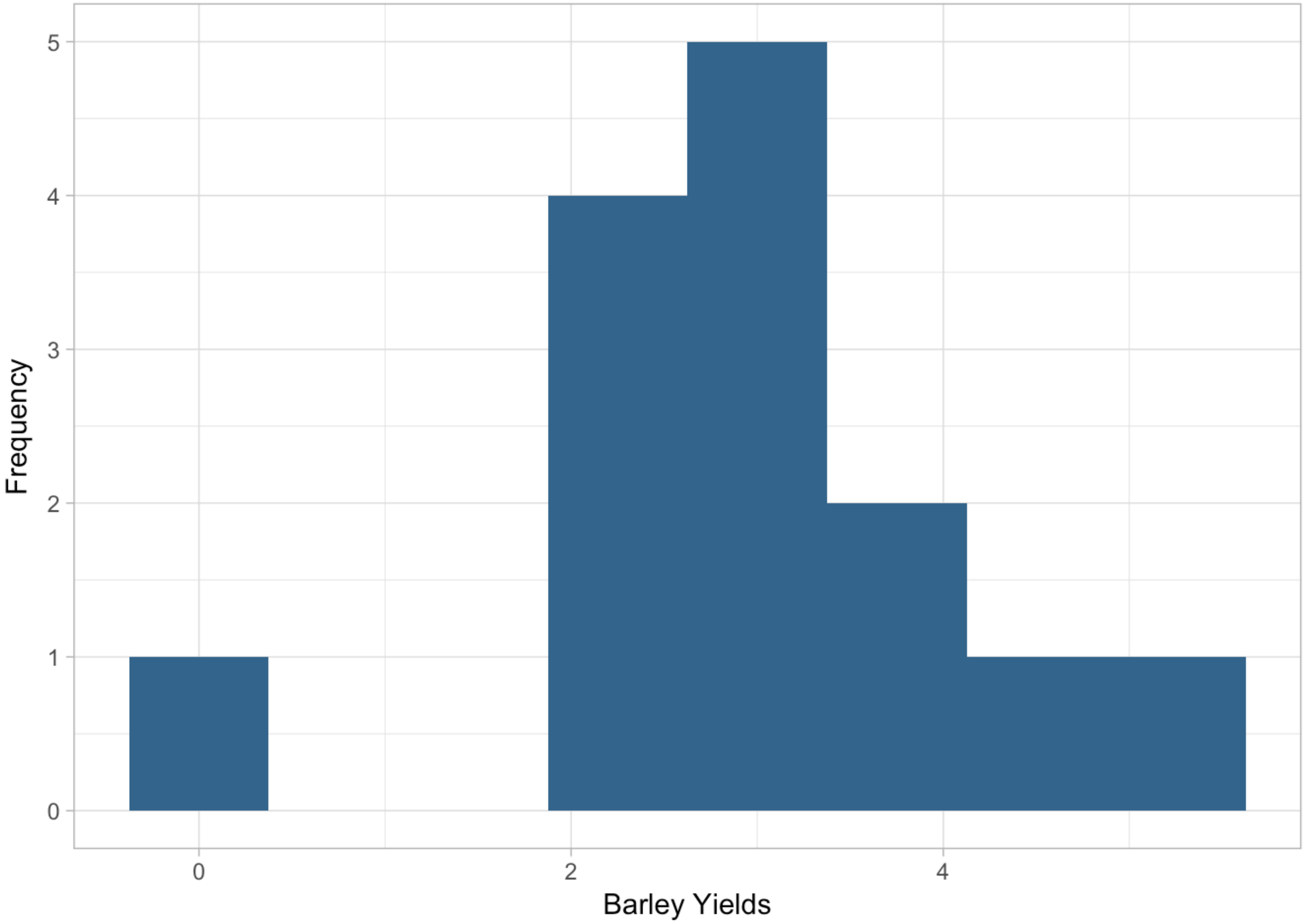
One of the most effective ways of analysing data is by observing its shape. This generaly means simple charts to begin with! Let’s begin with a histogram of each of the crops that were planted.



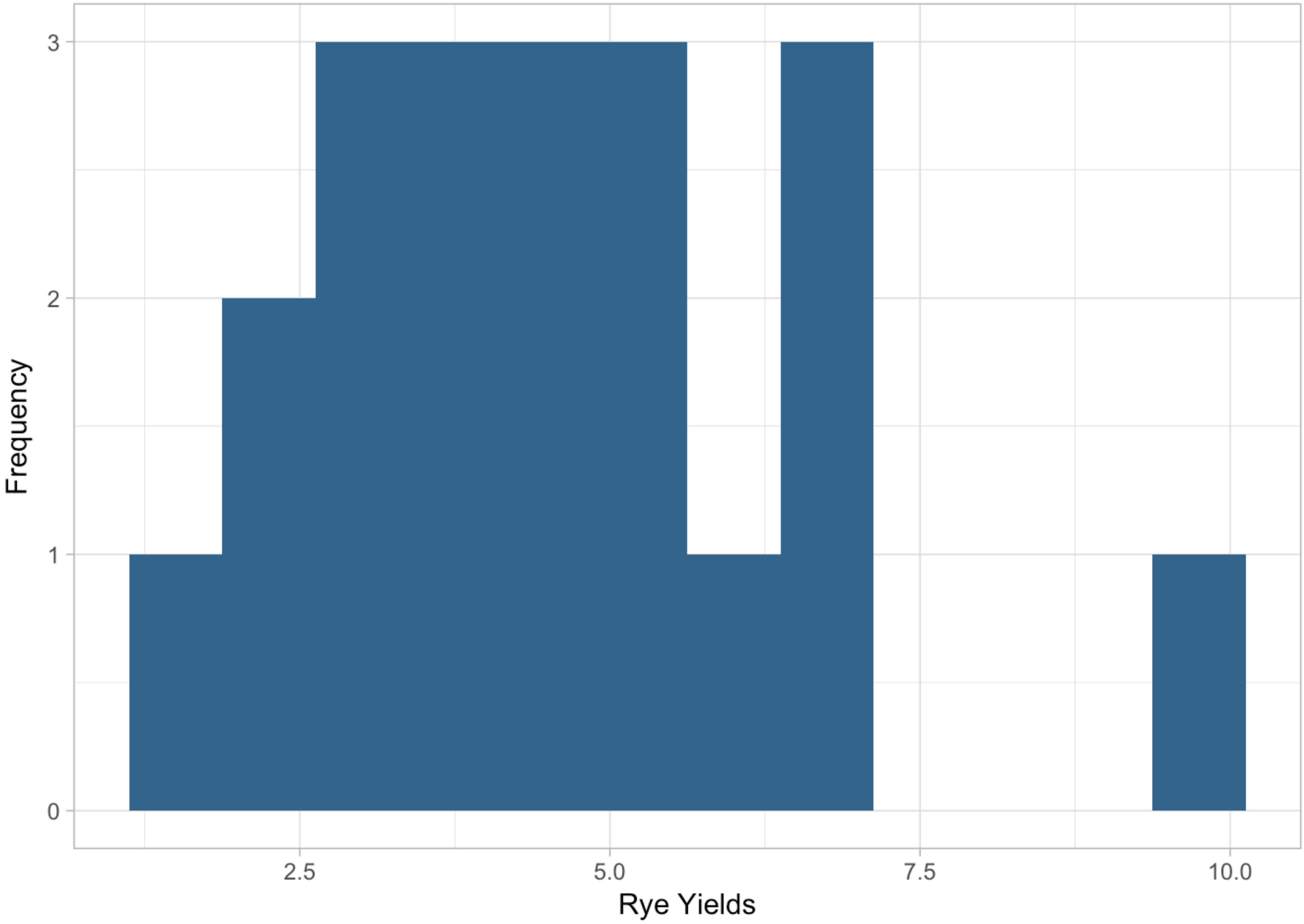
Wheat has most of the data clustered around the upper portion of the yield ratio. There are very few observations in the lower portion of the yield ratios, around 0 - 1. This tail to the left is called a *negative skew*.



Oat yeilds are more symmetric, the data is more evenly spread around a central point of the distribution, although there is some slight negative skewness here too.

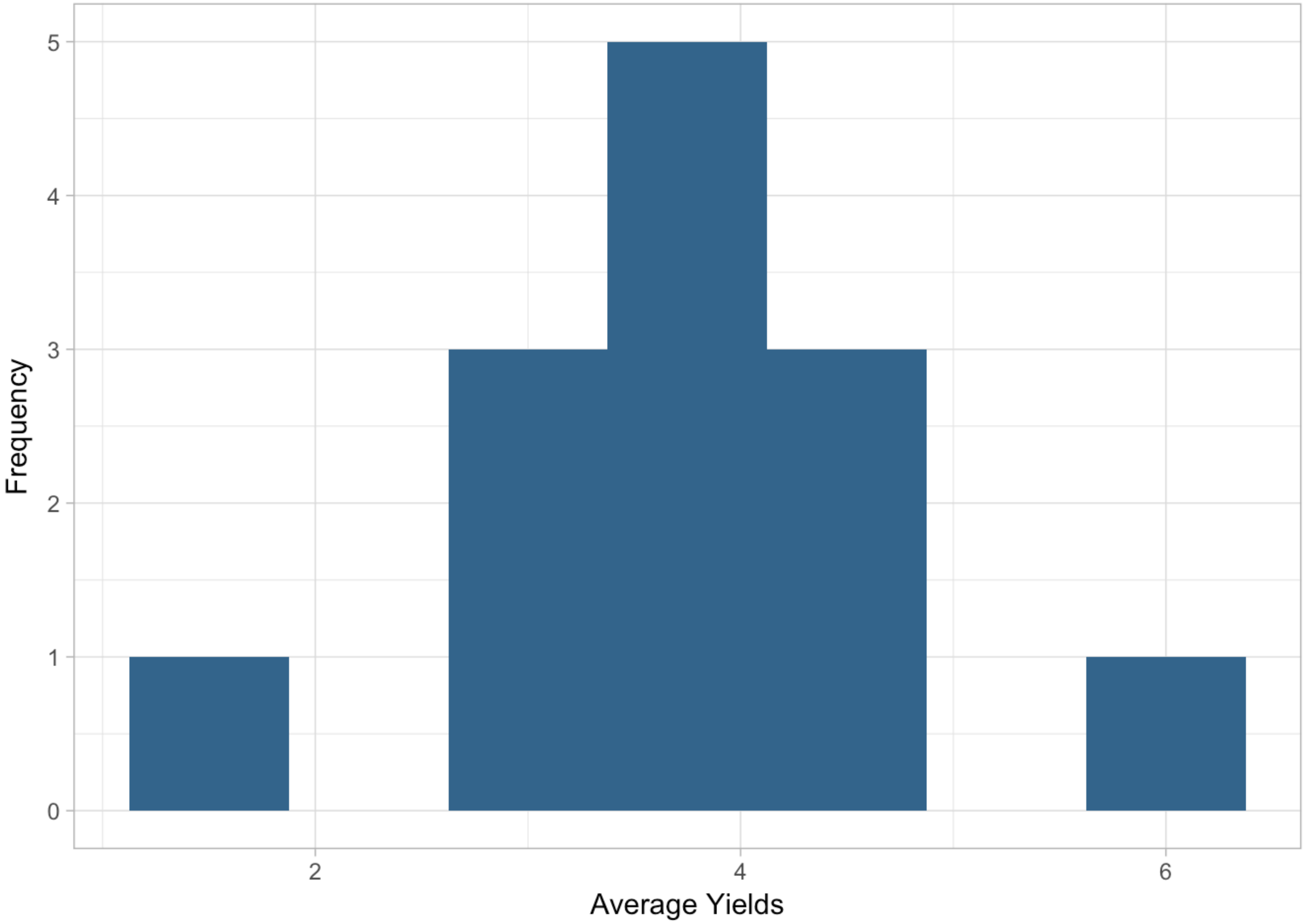


Barley yields on the other hand are skewed in the other direction - positively. With the exception of a single zero-yield observation, most of the data is clustered at the lower end of the yield distribution, around 2. The rest is spread out in a ‘tail’ towards the higher parts of the yield distribution. This is called *positive skewness*.



On the other hand, Rye yields are clustered fairly centrally, with a single observation far out to the right with a huge yield of 10. This is what we call an *outlier*. It’s a part of the data a long way from the rest of the distribution.

We’re budding data scientists here, we don’t have to just deal with the data we’re given: we can use our data to make more *interesting* data. Let’s look at the histogram of average yield ratios. That is, the average of all the yield ratios in any given year.



The histogram is very symmetric, the data is evenly spread around the central point of 3.5. This suggests that low-average-yield-years happened about as often as high-average-yield-years during the data collection period.

Quartiles and percentiles

Each histogram is a frequency distribution, it describes how common each part of the distribution is. We can divide up our frequency distributions into *percentiles*. For example, the 50th percentile is the observation where 50% of the data is smaller than that number. The 35th percentile is the observation where 35% of the data is smaller than that number.

The most commonly used percentiles are *quartiles*. These are the 25th, 50th and 75th percentiles, they divide the data into quarters. The 25th percentile for wheat yields is 2.95 while the 75th percentile is 4.98.

Analysing data by looking at measures of location

When we examined the histograms of the crop yields, we talked alot about *location*. Data was located symmetrically, or to the upper end of the yield ratios, or the lower end. Statistics define these measures of location more concretely and they can be grouped according to which kinds of location they're describing.

There are:

- Measures of central tendency. These describe what's happening to the middle portion of the data.

These include the mean, median and mode.

- Measures of spread. These describe how widely the data is dispersed. Is it very broadly dispersed, or is it tightly grouped around one point?
- Measures of skewness: there's only one measure of skewness we use commonly and that's, well, *skewness*. It does exactly what it says: it tells us if the data is symmetric or skewed, and if so - in what direction.

Measures of central tendency

Let's talk a little about what the measures of central tendency actually are:

Mean this is the classic statistic you probably learnt in school. Add all the numbers up and divide by the number of observations and you're done! In algebra, it looks like this:

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

(If the notation with Σ is unfamiliar to you, don't worry too much. Flip to the Appendix titled *An Introduction to Sigma Notation*.)

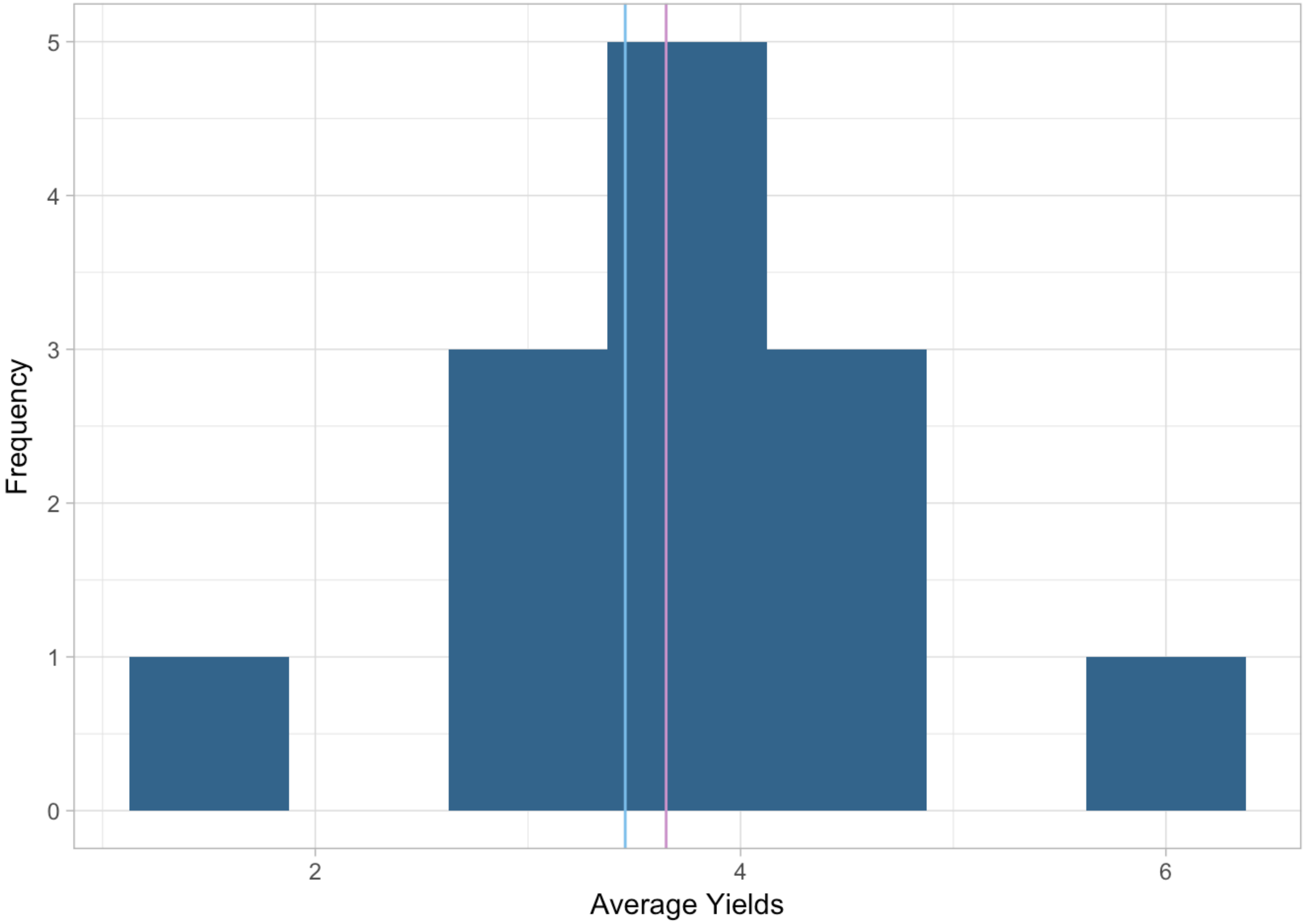
For example, the mean of wheat yields in this data is 3.87, while the mean of rye yields is 4.52. Overall, the mean average yield (that could be worded better) is 3.65.

When referring to a sample mean, you will often see it notated as $\bar{x}$.

Median The median is what we get if we line up all the numbers in order from smallest to biggest and chose the middle one. If there are an even set of numbers, we can choose halfway between the middle two numbers. The median of oat yields is 2.99, while the median of barley yields is 2.89. Not a huge difference between them.

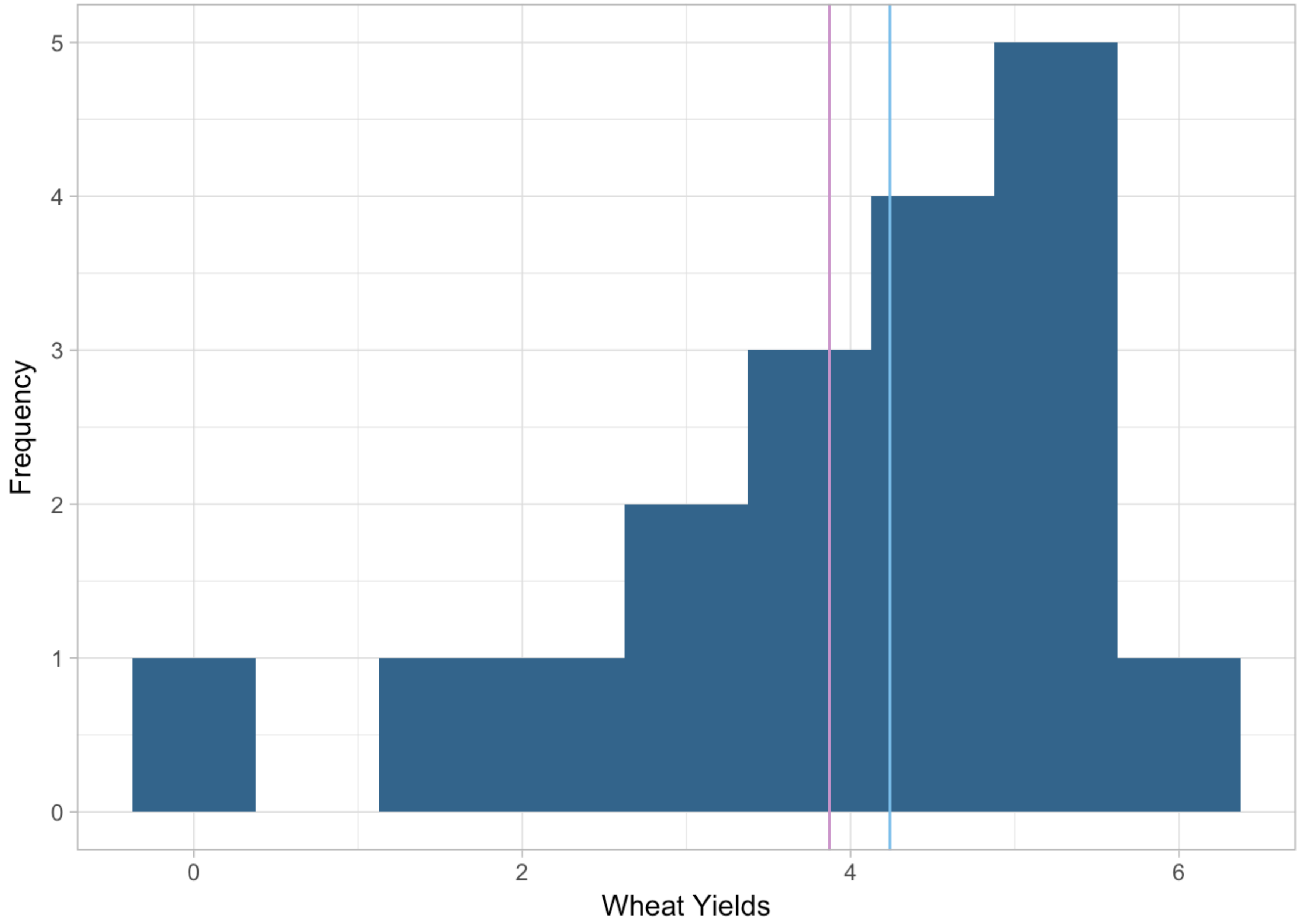
Mode The mode is simply the most *common* observation. In continuous data, it's very probable that a modal observation won't exist - there may be no two numbers the same. And in any case, if you had a dataset of 100 numbers and two were in common, would that necessarily say anything relevant about your data? We can, however, have a *modal class*. This is when we divide the data up into groups and the *modal class* is the most common group. In our histogram of average yields above, the modal class was 3.5-4. This was the most common group.

Let's look at these in context. Here we have our average yields histogram again, but with the mean in purple, the median in blue - the mode does not exist for this variable, there is no common value. However, we can also see that the modal *class* in this histogram is between 3.5 and 4.

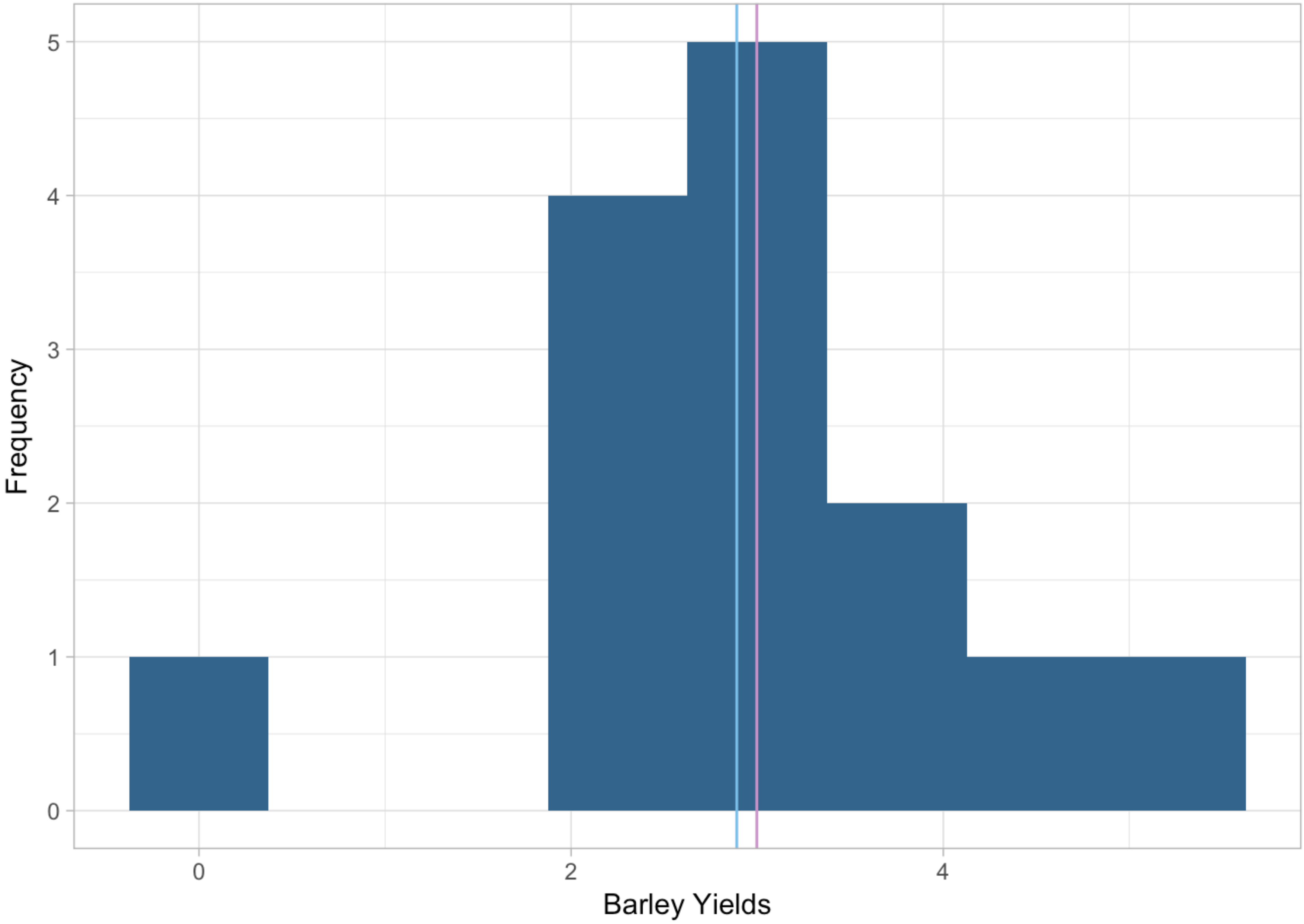


Note how close the median and mean are here. In a perfectly symmetric distribution, they are equal.

Let’s take a look at wheat yields now. Again, the mode doesn’t exist and the mean is in purple, while the median is in blue. Notice how the two are a little further apart with the median to the right of the mean: this is a feature of *negatively skewed* data.



Finally, barley yields again.This is a more symmetric distribution and the mean and median are closer together. However, note that the median is now to the left of the mean: this is a feature of *positively skewed* data.



Measures of spread

It's helpful to know where most of the data is located, but knowing how widely it's dispersed from that point is also important.

There are several measures of dispersion or spread. These include:

The range. What's the biggest number? What's the smallest number? Take the biggest from the smallest and you've got your range. This is a very simple statistic, but one that's really helpful. The range of rye yields is 8.38.

The inter quartile range. This is the range between the 25th and 75th percentiles. In the case of the rye yields, the interquartile range is 2.6.

The Variance. Watch out, algebra alert here. This statistic is the sum of the squared deviations from the mean. That looks like this in algebra:

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

There are a few variations on this statistic, one for the population with $\frac{1}{n}$ instead of $\frac{1}{n-1}$ for a sample. We'll stick with the sample version fo the moment.

While this looks complicated, all it's saying is that we should add up the squared differences between each observation x and its sample mean and divide by $n - 1$.

The variance of wheat yields is NA.

The standard deviation. This is simply the square root of the variance. The advantage of the standard deviation is that it is measured in the same units as the variable of interest itself. This makes it helpful for comparisons. The standard deviation of wheat yields is NA.

The coefficient of variation. This one isn't used very often, but it's a useful way of comparing two distributions with different means and standard deviations. It's simply the standard deviation divided by the mean. The coefficient of variation of wheat yields is 0.382 while that for oats is 0.252. This suggests that oat yields are less dispersed from their mean than wheat yields are.

Summary

This has been a brief introduction to some common statistics used in data analysis, but is hasn't been an introduction to *how* we might perform a data analysis in practice: stay tuned!

Where did this data come from?

Bruce M. S. Campbell (2007), Three centuries of English crops yields, 1211-1491 [WWW document]. URL http://www.cropyields.ac.uk (http://www.cropyields.ac.uk) 11/11/2017]