Chapter 1.5 - Measures of Position

> "To be sure of hitting the target, shoot first and call whatever you hit the target." - Ashleigh Brilliant

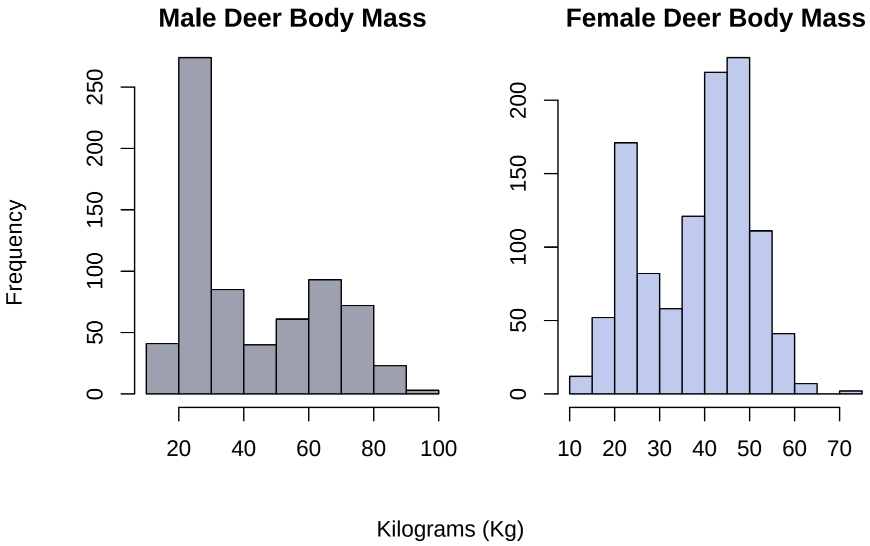
Suppose we observe a male white-tailed deer who's 60 kg and a female who's 45 kg. We have the tools to say how different they are from each other, but how can we describe their differences within their specific group? Our intuition might lead us towards stratification— if we *subset* the data based off of the groups we're interested in we could compare any individual we come across with their appropriate demographic.

This is, obviously, clunky and inconvenient. Since we're statisticians, and thus exceedingly lazy, we need to develop a better method for working through this problem.

---

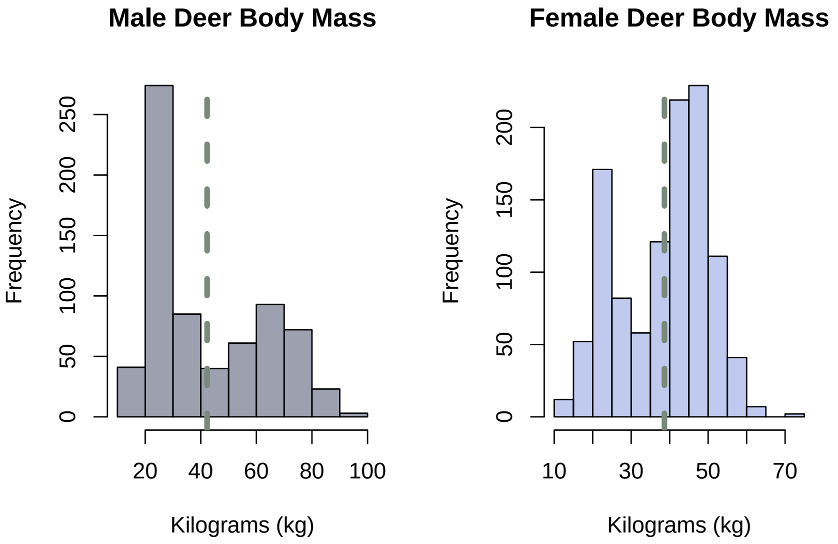
z-scores

It's a fact of nature that male and female mammals have differing body compositions. It should logically track that they're going to have different *average weights*.

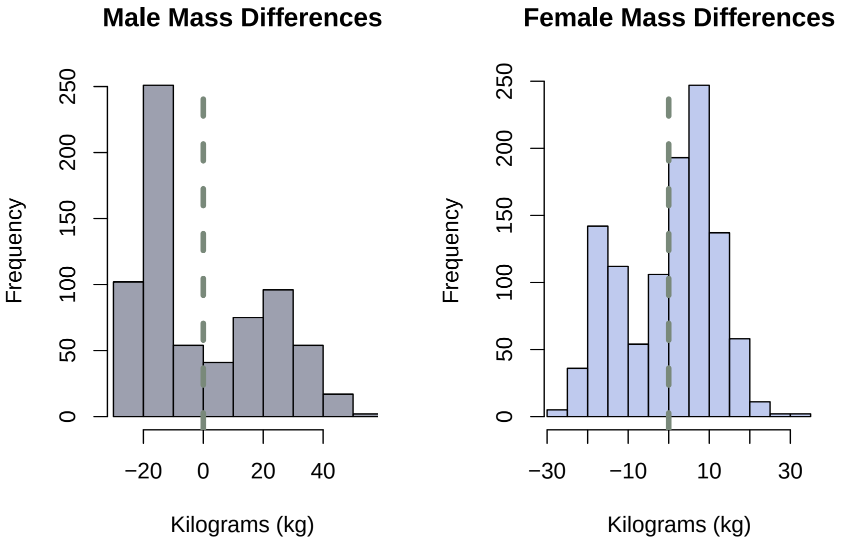


Comparing any individual deer within its group is simple enough and there's an argument to be made that the "easier" or "lazier" method here would be to just stratify and make comparisons within those strata. This works well until we have to make inferences *across* groups, i.e., How should a 60 kg male compare to a 45 kg female?

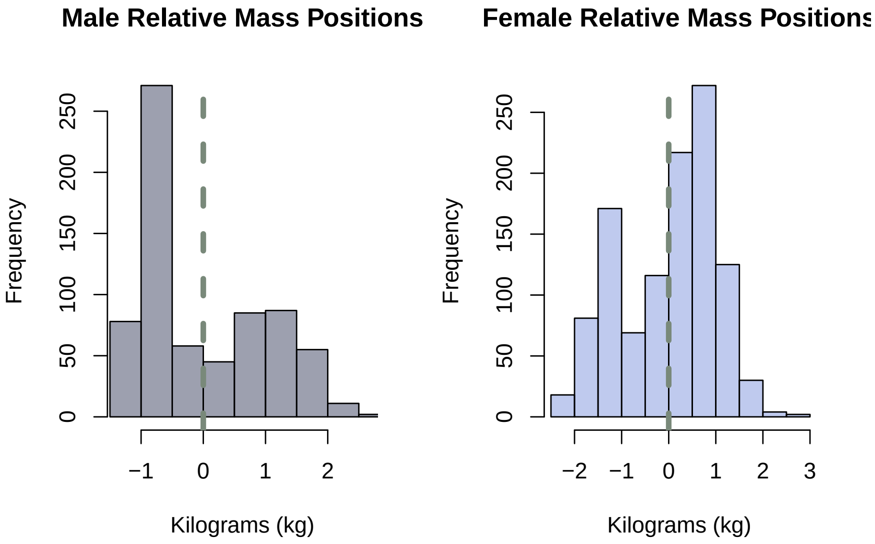
It's been a good idea so far to start at the mean, so let's do that.



We're going to go back to our previous trick of re-defining the mean as 0 so that we're looking at *differences* instead of direct values:



What we have currently is an adjusted data set of *signed* (as in positive and negative) differences between each data point in our two strata and their respective means, which we've calculated before! This time, however, we're going to leverage the **standard deviation** by using it as a *reference point.* If we divide every value in these new data sets by the standard deviation of the **original data sets** something very interesting will happen:



It might be hard to spot on the histograms alone, but when we look at the units in the underlying mathematics:

$$\frac{\text{Mass (Kg)} - \text{Mean Mass (Kg)}}{\text{Standard Deviation of Mass (Kg)}}$$

We placed each of our strata into a position relative to their mean then *removed the units.* What's left behind is a data set of *unitless* values that represent *each deer's* **number of standard deviations** they are from the mean. With this we could easily make comparisons within groups *and across them.*

This is the process of calculating a **z-score** (also known as standardizing data) and it's one of the most important techniques statistics has to offer.

A z-score for any value x is the **number of standard deviations** x is from the **mean** of the data set.

- $z < 0 \Rightarrow$ the value of x is **less than the mean**
- $z = 0 \Rightarrow$ the value of x is **equal to the mean**
- $z > 0 \Rightarrow$ the value of x is **greater than the mean**

Let x be a value from a **population** with mean μ. The z-score is:

$$z = \frac{x - \mu}{\sigma}$$

For a sample:

$$z = \frac{x - \bar{x}}{s}$$

The Empirical Rule

Recall that, given approximately bell-shaped data:

- $\approx 68\%$ of the data will be between $\mu - \sigma$ and $\mu + \sigma$
- $\approx 95\%$ of the data will be between $\mu - 2\sigma$ and $\mu + 2\sigma$
- $\approx 99.7 - 100\%$ of the data will be between $\mu - 3\sigma$ and $\mu + 3\sigma$

This directly translates to z-scores:

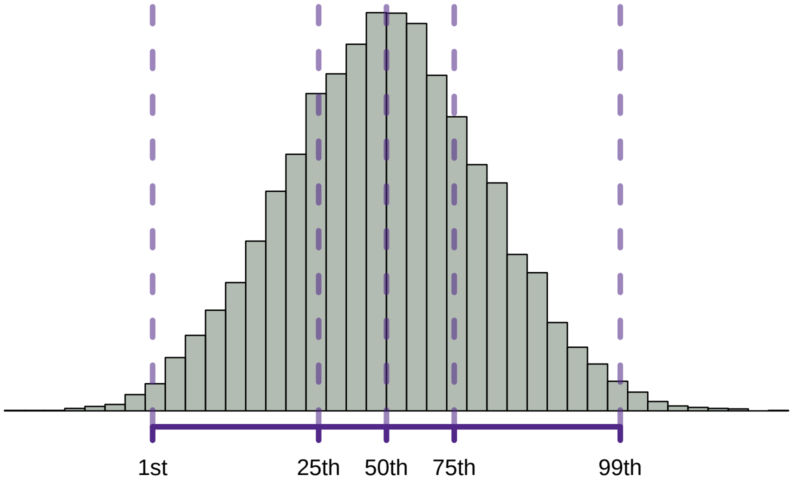
- $\approx 68\%$ of the data will be between $z = -1$ and $z = 1$
- $\approx 95\%$ of the data will be between $z = -2$ and $z = 2$
- $\approx 100\%$ of the data will be between $z = -3$ and $z = 3$

Percentiles and Quartiles

We'd already discussed a measure of position prior to this chapter— the median.

The median describes the central position in the data, so while it's a measure of center it's achieving this by *measuring a position*. We can expand on this idea quite easily, what if we were concerned about the center point between the first individual in the data and the median?

Percentiles are a more *precise* version of the median. If we define a number, p, between 1 and 99, the p^{th} **percentile** separates the lowest $p\%$ of the data from the highest $(100 - p)\%$. What we now have is a data set that's been sliced into 100 separate parts. We avoid describing the 0^{th} and 100^{th} percentiles because those are just the minimum and maximum respectively.



A special case of percentiles are the 25^{th}, 50^{th}, and 75^{th} percentiles, also known as **quartiles**.

Every data set has three quartiles:

- The 1^{st} quartile, denoted $\mathbf{Q_1}$ separates the **lowest** 25% of the data from the **highest** 75%
- The 2^{nd} quartile, denoted $\mathbf{Q_2}$ separates the **lowest** 50% of the data from the **highest** 50% ($\mathbf{Q_2}$ = Median)
- The 3^{rd} quartile, denoted $\mathbf{Q_3}$ separates the **lowest** 75% of the data from the **highest** 25%

The same method we use for computing percentiles can be used to compute quartiles— and it looks very similar to the method of derivation for locating the median:

1. Arrange the data in increasing order
2. Let n be the number of values in the data set. For the p^{th} percentile, compute the value:

$$L = \frac{p}{100} * n$$

3. If L is a whole number, the p^{th} percentile is the average of the number in position L and the number in position $L + 1$
   - If L is **not** a whole number, round it up the the next higher whole number. The p^{th} percentile is the number in the position corresponding to the rounded-up value

It's easy to show that regardless of the *magnitude* of n, this formula will *always* provide us with the median when computed at $p = 50$:

$$\mathrm{Q}_2 = \frac{50}{100} \times n = \frac{1}{2} \times n = \frac{n}{2}$$

The two remaining quartiles cover the halfway points between the median and the minimum/maximum. As such, we can consider Q_1 to be the median for the lower half of the data and Q_3 to be the median for the upper half of the data. While we can continue to use the formula for percentiles to locate these two quartiles:

$$\mathrm{Q}_1 = \frac{25}{100} \times n = \frac{1}{4} \times n = \frac{n}{4}$$

$$\mathrm{Q}_3 = \frac{75}{100} \times n = \frac{3}{4} \times n = \frac{3n}{4}$$

We can also separate the data at the halfway point (the median) and locate the median for each half.

Five-Number Summary

The **five-number summary** is a set of five measures of position computed from a data set. The idea behind this summary is to **thoroughly** describe the shape of a data set without visualizing it. The summary consists of:

Min	Q_1	Median	Q_3	Max

Below is a table of total of infected counts from a series of Foot-and-mouth disease (FMD) outbreaks in cattle.

192	152	90	124	178	180	127	182	196	118

The five number summary for this data is:

Min	Q_1	Median	Q_3	Max
90	124	165	182	196

The best use case of the five number summary is to make comparisons with individual data points. Consider that a new outbreak occurs with 190 cattle confirmed as infected.

$$Q_3 < 190 < \text{Max}$$

This is greater than 75% of the outbreaks but is not the largest.

How about an outbreak of 128?

$$Q_1 < 128 < \text{Median}$$

This is greater than 25% of the outbreaks, but less than of the others.

Interquartile Range (IQR)

To wrap up our discussion on measures of position, we'll discuss a measure of spread. The **IQR** is a measure of spread that is often used to detect outliers. We've held off discussing it until now because of it's nature, we won't find much use for the IQR *outside of* detecting and describing outliers. As an actual measure of spread it works similarly to the **range** (given that it's calculated the same general way) but it's not providing information on something we (statisticians, analysts, scientists) typically care about.

To find the IQR we take the difference between Q_1 and Q_3:

$$\text{IQR} = \text{Q}_3 - \text{Q}_1$$

Notice that the IQR contains the **middle** 50% of the data. This is where most of the **density** in our data is found, so we now have a way to reference what we could consider to be the **standard values** our data should take on. To find outliers with the IQR we use the creatively named **IQR Method**.

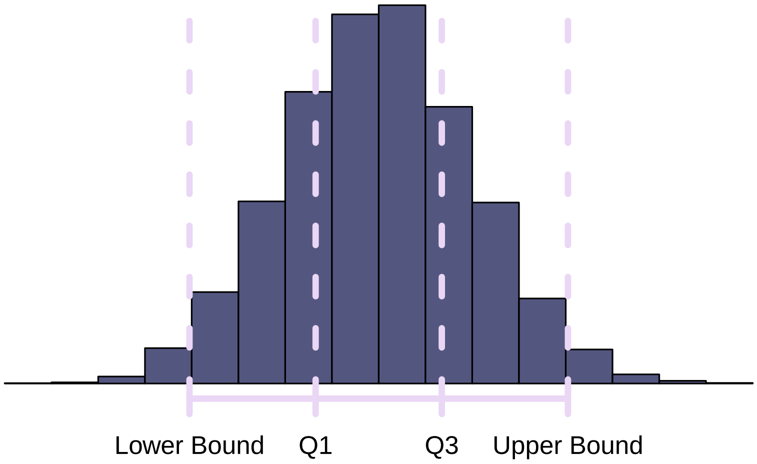
1. Define **outlier boundaries:**

$$\text{Lower Outlier Boundary} = \text{Q}_1 - 1.5 \times \text{IQR}$$

$$\text{Upper Outlier Boundary} = \text{Q}_3 + 1.5 \times \text{IQR}$$

2. Check to see if any data is outside of these boundaries:

$$\text{Upper Boundary} < x < \text{Lower Boundary}$$



Recall the five-number summary for the FMD data:

Min	Q_1	Median	Q_3	Max
90	124	165	182	196

The IQR for this dataset is:

$$\text{IQR} = \text{Q}_3 - \text{Q}_1 = 182 - 124 = 58$$

Defining the **outlier boundaries**:

$$\mathrm{Q}_3 + 1.5 * \mathrm{IQR} = 182 + (1.5 * 58) = 269$$

$$\mathrm{Q}_1 - 1.5 * \mathrm{IQR} = 124 - (1.5 * 58) = 37$$

Any value in the dataset < 37 or > 269 is an **outlier**. The IQR method isn't a perfect method and defining outliers becomes more nuanced and subjective as our analyses become more complex. However we can make the broad generalization that anything identified as an outlier *by the IQR method* is most definitely an outlier, while not every value missed by it can inherently be considered a *non*-outlier measurement.

Outliers

An **outlier** is a value that is considerably large or smaller than most of the values in a data set. This is only one of many discussions on outliers that will occur throughout this book, so don't worry if some questions remain unanswered.

Outliers can be rather uncomfortable for scientists and statisticians alike. We sometimes reconsider our entire experiment when an outlier is observed— which is a very natural, albeit extreme and misguided, reaction. It's important to recognize that outliers rarely break scientific understanding and are often mis-measurements. That said, we should only remove outliers when we've **confirmed beyond any reasonable doubt** that they're mis-measurements. Otherwise we should retain outliers because they can reveal interesting realities of the data generating process we may not have considered prior.

Remember that we defined a difference between resistant and non-resistant statistics. This wasn't for purely entertainment purposes. Scientists, analysts, and statisticians alike need to *work with* outliers rather than *fighting against* them.