12. Measures of spread

It is often important to know how much a set of numbers is spread out. That is, do all of the data cluster close to the mean, or are most values distant from the mean. For example, all of the numbers below are quite close to the mean of 5.0 (3 numbers are exactly 5.0).

```
4.9, 5.3, 5.0, 4.7, 5.1, 5.0, 5.0
```

In contrast, all of the numbers that follow are relatively distant from the same mean of 5.0.

```
3.0, 5.6, 7.8, 1.2, 4.3, 8.2, 4.9
```

This chapter focuses on summary statistics that describe the spread of data. The approach in this chapter is similar to Chapter 11, which provided verbal and mathematical explanations of measures of central tendency. We will start with the most intuitive measures of spread, the range and inter-quartile range. Then, we will move on to some more conceptually challenging measures of spread, the variance, standard deviation, coefficient of variation, and standard error. These more challenging measures can be a bit confusing at first, but they are absolutely critical for doing statistics. The best approach to learning them is to see them and practice using them in different contexts, which we will do here, in the Chapter 13 practical, and throughout the semester.

12.1. The range

The range of a set of numbers is probably the most intuitive measure of spread. It is simply the difference between the highest and the lowest value of a dataset (Sokal and Rohlf, 1995). To calculate it, we just need to take the highest value minus the lowest value. If we want to be fancy, then we can write a general equation for the range of a random variable X,

$$Range(X) = \max(X) - \min(X).$$

But really, all that we need to worry about is finding the highest and lowest values, then subtracting. Consider again the two sets of numbers introduced at the beginning of the chapter. In examples, it is often helpful to imagine numbers as representing something concrete that has been measured, so suppose that these numbers are the measured masses (in grams) of leaves from two different plants. Below are the masses of plant A, in which leaf masses are very similar and close to the mean of 5.

```
4.9, 5.3, 5.0, 4.7, 5.1, 5.0, 5.0
```

Plant B masses are below, which are more spread out around the same mean of 5.

```
3.0, 5.6, 7.8, 1.2, 4.3, 8.2, 4.9
```

To get the range of plant A, we just need to find the highest (5.3 g) and lowest (4.7 g) mass, then subtract,

$$Range(Plant\ A) = 5.3 - 4.7 = 0.6$$

Plant A therefore has a range of 0.6 g. We can do the same for plant B, which has a highest value of 8.2 g and lowest value of 1.2 g,

$$Range(Plant\ B) = 8.2 - 1.2 = 7.0$$

Plant B therefore has a much higher range than plant A.

It is important to mention that the range is highly sensitive to outliers (Navarro and Foxcroft, 2022). Just adding a single number to either plant A or plant B could dramatically change the range. For example, imagine if we measured a leaf in plant A to have a mass of 19.7 g (i.e., we found a huge leaf!). The range of plant A would then be $19.7 - 4.7 = 14$ instead of 0.6! Just this one massive leaf would then make the range of plant A double the range of plant B. This lack of robustness can really limit how useful the range is as a statistical measure of spread.

12.2. The inter-quartile range

The inter-quartile range (usually abbreviated as 'IQR') is conceptually the same as the range. The only difference is that we are calculating the range between quartiles rather than the range between the highest and lowest numbers in the dataset. A general formula subtracting the first quartile (Q_1) from the third quartile (Q_3) is,

$$IQR = Q_3 - Q_1.$$

Recall from Chapter 11 how to calculate first and third quartiles. As a reminder, we can sort the leaf masses for plant A below.

```
4.7, 4.9, 5.0, 5.0, 5.0, 5.1, 5.3
```

The first quartile will be the mean between 4.9 and 5.0 (4.95). The second quartile will be the the mean between 5.0 and 5.1 (5.05). The IQR of plant A is therefore,

$$IQR_{plant\ A} = 5.05 - 4.95 = 0.1.$$

We can calculate the IQR for plant B in the same way. Here are the masses of plant B leaves sorted.

```
1.2, 3.0, 4.3, 4.9, 5.6, 7.8, 8.2
```

The first quartile of plant B is 3.65, and the third quartile is 6.70. To get the IQR of plant B,

$$IQR_{plant\ B} = 6.70 - 3.65 = 3.05.$$

An important point about the IQR is that it is more robust than the range (Dytham, 2011). Recall that if we found an outlier leaf of 19.7 g on plant A, it would change the range of plant leaf mass from 0.6 g to 14 g. The IQR is not nearly so sensitive. If we include the outlier, the first quartile for plant A changes from $Q_1 = 4.95$ to $Q_1 = 4.975$. The second quartile changes from $Q_3 = 5.05$ to $Q_3 = 5.150$. The resulting IQR is therefore $5.150 - 4.975 = 0.175$. Hence, the IQR only changes from 0.1 to 0.175, rather than from 0.6 to 14. The one outlier therefore has a huge effect on the range, but only a modest effect on the IQR.

12.3. The variance

The range and inter-quartile range were reasonably intuitive, in the sense that it is not too difficult to think about what a range of 10, e.g., actually means in terms of the data. We now move to measures of spread that are less intuitive. These measures of spread are the variance, standard deviation, coefficient of variation, and standard error. These can be confusing and unintuitive at first, but they are extremely useful. We will start with

the variance; this section is long because we want to break the variance down carefully, step by step.

The sample variance of a dataset is a measure of the expected squared distance of data from the mean. To calculate the variance of a sample, we need to know the sample size (N, i.e., how many measurements in total), and the mean of the sample ($\bar{x}$). We can calculate the variance of a sample (s^2) as follows,

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

This looks like a lot, but we can break down what the equation is doing verbally. First, we can look inside the summation ($\sum$). Here we are taking an individual measurement x_i, subtracting the mean $\bar{x}$, then squaring. We do this for each x_i, summing up all of the values from $i = 1$ to $i = N$. This part of the equation is called the **sum of squares** (SS),

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

That is, we need to subtract the mean from each value x_i, square the result, and add everything up. Once we have this sum, SS, then we just need to multiply by $1/(N-1)$ to get the variance.

An example of how to do the actual calculation should help make it easier to understand what is going on. We can use the same values from plant A earlier.

```
4.9, 5.3, 5.0, 4.7, 5.1, 5.0, 5.0
```

To calculate the variance of plant A leaf masses, we start with the sum of squares. That is, take 4.9, subtract the sample mean of 5.0 ($4.9 - 5.0 = -0.1$), then square the result ($(-0.1)^2 = 0.01$). We do the same for 5.3, $(5.3 - 5.0)^2 = 0.09$, and add it to the 0.01, then continue down the list of numbers finishing with 5.0. This is what the sum of squares calculation looks like all written out,

$$SS = (4.9-5)^2 + (5.3-5)^2 + (5-5)^2 + (4.7-5)^2 + (5.1-5)^2 + (5-5)^2 + (5-5)^2.$$

Remember that the calculations in parentheses need to be done first, so the next step for calculating the sum of squares would be the following,

$$SS = (-0.1)^2 + (0.3)^2 + (0)^2 + (-0.3)^2 + (0.1)^2 + (0)^2 + (0)^2.$$

Next, we need to square all of the values,

$$SS = 0.01 + 0.09 + 0 + 0.09 + 0.01 + 0 + 0.$$

If we sum the above, we get $SS = 0.2$. We now just need to multiply this by $1/(N-1)$, where $N = 7$ because this is the total number of measurements in the plant A dataset,

$$s^2 = \frac{1}{7-1}\,(0.2)\,.$$

From the above, we get a variance of approximately $s^2 = 0.0333$.

Fortunately, it will almost never be necessary to calculate a variance manually in this way. Any statistical software will do all of these steps and calculate the variance for us (Chapter 13 explains how in Jamovi). The only reason that we present the step-by-step calculation here is to help explain the equation for s^2. The details can be helpful for understanding how the variance works as a measure of spread. For example, note that what we are really doing here is getting the distance of each value from the mean, $x_i - \bar{x}$. If these distances tend to be large, then it means that most data points (x_i) are far away from the mean ($\bar{x}$), and the variance (s^2) will therefore increase. The differences $x_i - \bar{x}$ are squared because we need all of the values to be positive, so that variance increases regardless of whether a value x_i is higher or lower than the mean. It does not matter if x_i is 0.1 lower than $\bar{x}$ (i.e., $x_i - \bar{x} = -0.1$), or 0.1 higher (i.e., $x_i - \bar{x} = 0.1$). In both cases, the deviation from the mean is the same. Moreover, if we did not square the values, then the sum of $x_i - \bar{x}$ values would always be 0 (you can try this yourself)[1]. Lastly, it turns out that the variance is actually a special case of a more general concept called the *covariance*, which we will look at later in Week 9 and makes the squaring of differences make a bit more sense.

We sum up all of the squared deviations to get the SS, then divide by the sample size minus 1, to get the mean squared deviation from the mean. That is, the whole process gives us the *average* squared deviation from the mean. But wait, why is it the sample size minus 1, $N - 1$? Why would we subtract 1 here? The short answer is that in calculating a *sample* variance, s^2, we are almost always trying to estimate the corresponding *population* variance (σ^2). And if we were to just use N instead of $N - 1$, then our s^2 would be a biased estimate of σ^2 (see Chapter 4 for a reminder on the difference between samples and populations). By subtracting 1, we are correcting for this bias to get a more accurate estimate of the population variance. It is not necessary to do this ourselves; statistical software like Jamovi and R will do it automatically. This

[1] If you are wondering why we square the difference $x_i - \bar{x}$ instead of just taking its absolute value, this is an excellent question! You have just invented something called the mean absolute deviation. There are some reasons why the mean absolute deviation is not as good of a measure of spread as the variance. Navarro and Foxcroft (2022) explain the mean absolute deviation, and how it relates to the variance, very well in section 4.2.3 of their textbook. We will not get into these points here, but it would be good to check out Navarro and Foxcroft (2022) for more explanation.

is really all that it is necessary to know for now, but see this footnote[2] for a bit more detailed explanation to try to make this intuitive (it is actually quite cool!). Later, we will explore the broader concept of *degrees of freedom*, which explains why we need to take into account the number of parameters in a statistic that are free to vary when calculating a statistic[3].

This was a lot of information. The variance is not an intuitive concept. In addition to being a challenge to calculate, the calculation of a variance leaves us with a value in units squared. That is, for the example of plant leaf mass in grams, the variance is measured in grams squared, g^2, which is not particularly easy to interpret. For more on this, Navarro and Foxcroft (2022) have a really good section on the variance. Despite its challenges as a descriptive statistic, the variance has some mathematical properties that are very useful (Navarro and Foxcroft, 2022), especially in the biological and environmental sciences.

For example, variances are additive, meaning that if we are measuring two separate characteristics of a sample, A and B, then the variance of A+B equals the variance of A plus the variance of B; i.e., $Var(A + B) = Var(A) + Var(B)$ [4]. This is relevant to genetics when measuring heritability. Here, the total variance in the phenotype of a population (e.g., body mass of animals) can be partitioned into variance attributable to genetics plus variance attributable to the environment,

$$Var(Phenotype) = Var(Genotype) + Var(Environment).$$

This is also sometimes written as $V_P = V_G + V_E$. Applying this equation to calculate

[2] To get the true population variance σ^2, we would also need to know the true mean μ. But we can only estimate μ from the sample, $\bar{x}$. That is, what we would really want to calculate is $x_i - \mu$, but the best we can do is $x_i - \bar{x}$. The consequence of this is that there will be some error that underestimates the true distance of x_i values from the population mean, μ. Here is the really cool part; to determine the extent to which our estimate of the variance is biased by using $\bar{x}$ instead of μ, we just need to know the expected squared difference between the two values, $(\bar{x} - \mu)^2$. It turns out that this difference (i.e., the bias of our estimate s^2) is just σ^2/N; that is, the true variance of the population divided by the sample size. If we subtract this value from σ^2, so $\sigma^2 - \sigma^2/N$, then we can get the expected difference between the true variance and the estimate from the sample size. We can rearrange $\sigma^2 - \sigma^2/N$ to get $\sigma^2 \times (N-1)/N$, which means that we need to correct our sample variance by $N/(N-1)$ to get an unbiased estimate of σ^2. If all of this is confusing, that is okay! This is really only relevant for those interested in statistical theory, which is not the focus of this module.

[3] Briefly, in the case of sample variance, note that we needed to use all the values x_i in the dataset and the sample mean $\bar{x}$. But if we know what all of the x_i values are, then we also know $\bar{x}$. And if we know all but one value of x_i and $\bar{x}$, then we could figure out the last x_i. Hence, while we are using N values in the calculation of s^2, the use of $\bar{x}$ reduces the degree to which these values are free to vary. We have lost 1 degree of freedom in the calculation of $\bar{x}$, so we need to account for this in our calculation of s^2 by dividing by $N - 1$. This is another way to think about the $N - 1$ correction factor (Sokal and Rohlf, 1995) explained in the previous footnote.

[4] This has one caveat, which is not important for now. Values of A and B must be uncorrelated. That is, A and B cannot covary. If A and B covary, i.e., $Cov(A, B) \neq 0$, then $Var(A + B) = Var(A) + Var(B) + Cov(A, B)$. That is, we need to account for the covariance when calculating $Var(A + B)$.

heritability ($H^2 = V_G/V_P$) can be used to predict how a population will respond to natural selection. This is just one place where variance reveals itself to be a highly useful statistic in practice. Nevertheless, as a descriptive statistic to communicate the spread of a variable, it usually makes more sense to calculate the standard deviation of the mean.

12.4. The standard deviation

The standard deviation of the mean (s) is just the square root of the variance,

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2}.$$

This is a simple step, mathematically, but it also is easier to understand conceptually as a measure of spread (Navarro and Foxcroft, 2022). By taking the square root of the variance, our units are no longer squared, so we can interpret the standard deviation in the same terms as our original data. For example, the leaf masses of plant A and plant B in the example above were measured in grams. While the variance of these masses were in g^2, the standard deviation is in g, just like the original measurements. For plant A, we calculated a leaf mass variance of $s^2 = 0.0333\,g^2$, which means that the standard deviation of leaf masses is $s = \sqrt{0.0333\,g^2} = 0.1825\,g$. Because we are reporting s in the original units, it is a very useful measure of spread to report, and it is an important one to be able to interpret. To help with the interpretation, here is an interactive tool showing how the heights of trees in a forest change across different standard deviation values[5].

> Click here for an interactive application to illustrate the standard deviation.

Here is another interactive tool showing how the shape of a histogram changes when the standard deviation of a distribution is changed[6].

> Click here for an interactive application to visualise how a histogram changes given a changing standard deviation.

The practical in Chapter 13 explains how to calculate the standard deviation in Jamovi.

[5] Here is the full URL: https://bradduthie.shinyapps.io/forest/
[6] Here is the full URL: https://bradduthie.shinyapps.io/normal_pos_neg/

12.5. The coefficient of variation

The coefficient of variation (CV) is just the standard deviation divided by the mean,

$$CV = \frac{s}{\bar{x}}.$$

Dividing by the mean seems a bit arbitrary at first, but this can often be useful for comparing variables with different means or different units. The reason for this is that the units cancel out when dividing the standard deviation by the mean. For example, for the leaf masses of plant A, we calculated a standard deviation of 0.1825 g and a mean of 5 g. We can see the units cancel below,

$$CV = \frac{0.1825\ g}{5\ g} = 0.0365.$$

The resulting CV of 0.0365 has no units; it is *dimensionless* (Lande, 1977). Because it has no units, it often used to compare measurements with much different means or with different measurement units. For example, Sokal and Rohlf (1995) suggest that biologists might want to compare tail length variation between animals with much different body sizes, such as elephants and mice. The standard deviation of tail lengths between these two species will likely be much different just because of their difference in size, so by standardising by mean tail length, it can be easier to compare relative standard deviation. This is a common application of the CV in biology, but it needs to be interpreted carefully (Pélabon et al., 2020).

Often, we will want to express the coefficient of variation as a percentage of the mean. To do this, we just need to multiply the CV above by 100%. For example, to express the CV as a percentage, we would multiply the 0.0365 above by 100%, which would give us a final answer of $CV = 3.65\%$.

12.6. The standard error

The standard error of the mean is the last measurement that we will introduce here. It is slightly different than the previous estimates in that it is a measure of the variation in the *mean* of a sample rather than the sample itself. That is, the standard error tells us how far our sample mean $\bar{x}$ is expected to deviate from the true mean μ. Technically, the standard error of the mean is the standard deviation *of sample means* rather than the standard deviation *of samples*. What does that even mean? It is easier to explain with a concrete example.

Imagine that we want to measure nitrogen levels in the water of Airthrey Loch (the loch at the centre of campus at the University of Stirling). We collect 12 water samples and

record the nitrate levels in milligrams per litre (mg/l). The measurements are reported below.

```
0.63, 0.60, 0.53, 0.72, 0.61, 0.48, 0.67, 0.59, 0.67, 0.54, 0.47, 0.87
```

We can calculate the mean of the above sample to be $\bar{x} = 0.615$, and we can calculate the standard deviation of the sample to be $s = 0.111$. We do not know what the *true* mean μ is, but our best guess is the sample mean $\bar{x}$. Suppose, however, that we then went back to the loch to collect another 12 measurements (assume that the nitrogen level of the lake has not changed in the meantime). We would expect to get values similar to our first 12 measurements, but certainly not the *exact* same measurements, right? The sample mean of these new measurements would also be a bit different. Maybe we actually go out and do this and get the following new sample.

```
0.47, 0.56, 0.72, 0.61, 0.54, 0.64, 0.68, 0.54, 0.48, 0.59, 0.62, 0.78
```

The mean of our new sample is 0.603, which is a bit different from our first. In other words, the sample means vary. We can therefore ask what is the variance and standard deviation *of the sample means*. In other words, suppose that we kept going back out to the loch, collecting 12 new samples, and recording the sample mean each time? The standard deviation of those sample means would be the standard error. **The standard error is the standard deviation of $\bar{x}$ values around the true mean μ.** But we do not actually need to go through the repetitive resampling process to estimate the standard error. We can estimate it with just the standard deviation and the sample size. To do this, we just need to take the standard deviation of the sample (s) and divide by the square root of the sample size ($\sqrt{N}$),

$$SE = \frac{s}{\sqrt{N}}.$$

In the case of the first 12 samples from the loch in the example above,

$$SE = \frac{0.111}{\sqrt{12}} = 0.032.$$

The standard error is important because it can be used to evaluate the uncertainty of the sample mean in comparison with the true mean. We can use the standard error to place confidence intervals around our sample mean to express this uncertainty. We will calculate confidence intervals in Week 5, so it is important to understand what the standard error is measuring.

If the concept of standard error is still a but unclear, we can work through one more hypothetical example. Suppose again that we want to measure the nitrogen concentration of a loch. This time, however, assume that we somehow *know* that the true mean

N concentration is $\mu = 0.7$, and that the standard deviation of water sample N concentration is $\sigma = 0.1$. Of course, we can never actually know the *true* parameter values, but we can use a computer to simulate sampling from a population in which the true parameter values are known. In Table 12.1, we simulate the process of going out and collecting 10 water samples from Airthrey Loch. This collecting of 10 water samples is repeated 20 different times. Each row is a different sampling effort, and columns report the 10 samples from each effort.

Table 12.1.: Simulated samples of nitrogen content from water samples of Airthrey Loch. Values are sampled from a normal distribution with a mean of 0.7 and a standard deviation 0.1.

Sample_1	0.84	0.63	0.71	0.76	0.74	0.60	0.59	0.75	0.70	0.61
Sample_2	0.64	0.54	0.48	0.78	0.68	0.56	0.62	0.66	0.61	0.66
Sample_3	0.90	0.74	0.76	0.81	0.75	0.68	0.68	0.85	0.53	0.58
Sample_4	0.64	0.54	0.66	0.74	0.75	0.61	0.62	0.57	0.61	0.72
Sample_5	0.57	0.71	0.79	0.79	0.69	0.69	0.66	0.81	0.75	0.44
Sample_6	0.77	0.64	0.68	0.71	0.64	0.64	0.61	0.68	0.63	0.72
Sample_7	0.72	0.53	0.67	0.62	0.55	0.75	0.62	0.58	0.82	0.83
Sample_8	0.79	0.57	0.72	0.52	0.69	0.71	0.81	0.72	0.80	0.75
Sample_9	0.72	0.67	0.59	0.67	0.72	0.79	0.72	0.63	0.65	0.84
Sample_10	0.58	0.82	0.81	0.75	0.78	0.80	0.72	0.59	0.64	0.69
Sample_11	0.65	0.73	0.72	0.59	0.79	0.71	0.82	0.72	0.65	0.58
Sample_12	0.86	0.80	0.69	0.56	0.69	0.68	0.65	0.76	0.61	0.65
Sample_13	0.73	0.59	0.74	0.68	0.67	0.57	0.58	0.60	0.69	0.52
Sample_14	0.55	0.50	0.59	0.70	0.81	0.79	0.86	0.56	0.80	0.86
Sample_15	0.66	0.65	0.56	0.75	0.57	0.61	0.68	0.83	0.76	0.76
Sample_16	0.56	0.84	0.50	0.66	0.79	0.84	0.61	0.62	0.64	0.72
Sample_17	0.79	0.78	0.63	0.73	0.72	0.90	0.59	0.66	0.80	0.63
Sample_18	0.68	0.75	0.72	0.70	0.69	0.55	0.72	0.67	0.86	0.59
Sample_19	0.61	0.74	0.83	0.47	0.83	0.62	0.58	0.63	0.87	0.73
Sample_20	0.74	0.51	0.57	0.87	0.74	0.66	0.62	0.60	0.74	0.63

We can calculate the mean of each sample by calculating the mean of each row. These 20 means are reported below.

```
##       [,1]  [,2]  [,3]  [,4]  [,5]  [,6]  [,7]  [,8]  [,9] [,10]
## [1,] 0.693 0.623 0.728 0.646 0.690 0.672 0.669 0.708 0.700 0.718
## [2,] 0.696 0.695 0.637 0.702 0.683 0.678 0.723 0.693 0.691 0.668
```

The standard deviation of the 20 sample means reported above is 0.0273982. Now suppose that we only had Sample 1 (i.e., the top row of data). The standard deviation

of Sample 1 is $s = 0.083006$. We can calculate the standard error from these sample values below,

$$s = \frac{0.083006}{\sqrt{10}} = 0.0262488.$$

The estimate of the standard error from calculating the standard deviation of the sample means is therefore 0.0273982, and the estimate from just using the standard error formula and data from only Sample 1 is 0.0262488. These are reasonably close, and would be even closer if we had either a larger sample size in each sample (i.e., higher N) or a larger number of samples.