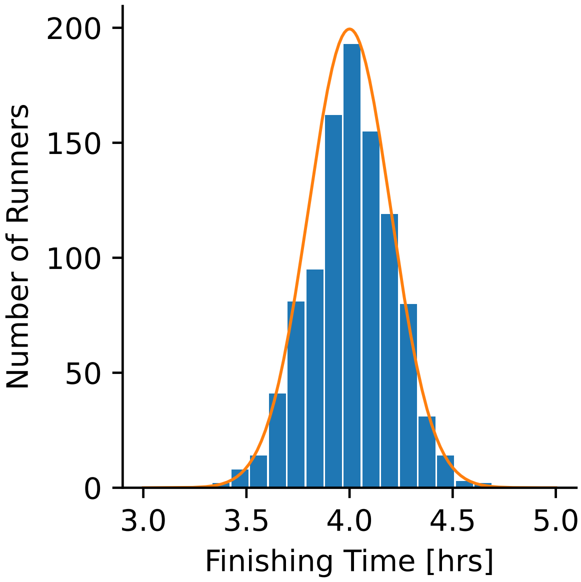
Population vs Sample Statistics

James Pickering, School of Chemistry, University of Leicester

To further elaborate on the *'what's the difference between population and sample statistics'* question, it's helpful to consider what we are trying to achieve with statistics, and this is best illustrated with an example. Imagine we have 1000 people running a marathon[1], and we look at the statistics of their finishing times: I simulated this by generating 1000 times to be normally distributed with a mean of 4 hours, and a standard deviation of 0.2 hours (12 mins). As we can see, their finishing times are (roughly) normally distributed around a mean time of 4 hours (the blue bars show a histogram of the finishing times, and the orange line shows the 'ideal' normal distribution). If I wanted to know the mean finishing time of all runners (and it's standard deviation), then I can calculate it using the formulae we know (treating the data we have as an **entire population**):



$$\text{Population Mean}: \ \mu = \frac{1}{n}\sum_i x_\mathrm{i} \tag{1}$$

$$\text{Population Standard Deviation}: \ \sigma = \sqrt{\frac{1}{n}\sum_i (x_\mathrm{i} - \mu)^2} \tag{2}$$

If I do this calculation, then I'll find that for all 1000 runners, the mean $\mu = 4.009$ hours, and the standard deviation $\sigma = 0.199$ hours[2]. This matches pretty well with what we know is true from the simulation parameters I used (mean of 4 hours, standard deviation of 0.2 hours).

[1] This example shamelessly stolen from Michael Burt, University of Oxford.

[2] The data is attached in a file if you don't believe me and want to calculate yourself.

But, what if in our marathon, we only have limited timing equipment, and so we can only actually measure the finishing time of ten runners? Can we use the data from these ten runners to try and estimate the mean and standard deviation of the finishing times of all 1000 runners? **Yes we can**, and this is where we use the formulae for **sample** mean and standard deviation:

$$\text{Sample Mean : } \bar{x} = \frac{1}{n}\sum_i x_{\mathrm{i}} \tag{3}$$

$$\text{Sample Standard Deviation : } s = \sqrt{\frac{1}{n-1}\sum_i (x_{\mathrm{i}} - \bar{x})^2} \tag{4}$$

This is the key idea of sample statistics, and the key take-home:

> The point of **sample statistics** is that we are trying to get the **best estimate** of the **population statistics** from limited data (a **sample** of the **population**).

Let's randomly pick ten times from our list of 1000:

$$4.0359, 4.2089, 3.9416, 4.0910, 3.7579, 3.6973, 4.3213, 4.1614, 3.9589, 4.0044$$

We can now calculate the sample mean and standard deviation using equations (3) and (4) above, and we find that the sample mean, $\bar{x} = 4.018$ hours, and the sample standard deviation, $s = 0.193$ hours. So, picking just ten runners didn't do a bad job in estimating our population statistics!

But, what if we had instead used the formulae for population mean and standard deviation? Clearly the mean would be the same, but if we used equation (2) rather than equation (4) for our standard deviation (having the denominator as n not $n-1$)? In this case, we would find that our standard deviation was 0.183 hours, quite a bit lower than the true standard deviation. This is generally the case, and we can say that:

> Using n rather than $n-1$ in the formula for sample standard deviation, will **always** result in you underestimating the population standard deviation.

Remember, our goal is to use $\bar{x}$ and s to **estimate** μ and σ. In an ideal world where we have infinitely many data points and an infinitely large sample, $\mu = \bar{x}$ and $\sigma = s$.

An important subtlety here is that we are free to define for ourselves what the 'population' is. We might take our data from 1000 runners and treat it as a population (like we did here). Alternatively, we might treat that as a *sample* of 1000 runners out of all the people running marathons globally (probably several million runners). In this case, we could estimate the mean and standard deviation of the **global** population using the sample formulae, taking our 1000 runners as a small sample. Note however, that as n gets big, $n-1 \approx n$, so the difference matters less and less. The bottom line is that it's important to think about what you're doing, and whether or not you are looking at a whole population, or just a sample of a larger population. My rule of thumb is: **if in doubt, use the sample formulae**, but no doubt some statisticians would disagree with me[3].

On a more advanced note, we can unpick the difference between equations (2) and (4) a bit further:

$$\text{Sample Standard Deviation : } s = \sqrt{\frac{1}{n-1}\sum_i (x_\mathrm{i} - \bar{x})^2}$$

$$\text{Population Standard Deviation : } \sigma = \sqrt{\frac{1}{n}\sum_i (x_\mathrm{i} - \mu)^2}$$

The difference is that factor of $n-1$ in the denominator, as we know. This factor is known as **Bessel's Correction** and essentially aims to correct for the issue we just saw in the running example - if we don't use it, we end up underestimating the population standard deviation. But, why is the factor $n-1$ and not anything else? The reason for this is quite subtle but there are two things that may make it a bit clearer:

1. (More simple): Imagine taking a sample of 1 data point. Clearly, we can't define the standard deviation of one data point, and the factor of $n-1$ reflects this (as then our denominator would be $1-1=0$, and our standard deviation undefined - division by zero).
2. (More tricky): If we have n points in our sample data, then we have to use these n points to calculate our sample mean, $\bar{x}$. This means that all of our points are going to be *biased* and be nearer $\bar{x}$ than μ (μ being the 'true' mean of the population that we want to estimate). This bias means that the $(x_\mathrm{i}-\bar{x})^2$ part of the formula (the **sum of residuals**) for s is going to be smaller than the $(x_\mathrm{i}-\mu)^2$ part of the formula for σ. Bessel's correction helps to remove this bias, by dividing the sum of residuals residual by a smaller number, compensating for the fact that it is probably an underestimate due to the definition of $\bar{x}$.

This is more than enough and just being aware of all these things is important. Statistics quickly becomes very complex once you dig deeper, but being aware of the possible problems and subtleties is as much as we generally need as chemists!

[3] In researching a bit, I have found several places advocating for either sample or population as the default choice - so I'm not sure there's a universal answer!