# Quantitative Understanding in Biology Short Course
# Session 1
# Working with Continuous Data; Confidence Intervals

Jason Banfelder
Luce Skrabanek

March 11th, 2025

## 1 Introduction

The linchpin of the scientific method is the reproducibility of our experiments and our measurements. We like to believe that if we describe our experiments in sufficient detail, account for all relevant variables, and are sufficiently careful, then we should get the same results when we repeat our manipulations, and, more importantly, that another competent scientist should be able to do the same, and thus be able to verify our results. Without getting too philosophical about the nature of scientific inquiry and epistemology[1], we can hopefully all agree that this reproducibility and verifiability are essential to the accumulation of knowledge in society.

Our experiences in everyday reality, and in the lab, belie this idealized notion; we know all too well that, in nearly all cases, repeating the same experiment in the lab results in different measurement values. On a good day, our repeated measurements are "close enough" to each other, and we can, with the "right statistical treatments", draw inferences from such measurements. On other days, of course, our measurements are all over the place, and we say that our experiment "doesn't work (yet!)."

The devil is in the details, though; this short course is about what "close enough" is, and what the "right statistical treatments" are. While statistics is necessarily quantitative, and some mathematics are unavoidable, we'll seek to explore the necessary concepts with a minimum of mathematics (this is NOT a math course), and instead appeal to your intuition. This is a bit of a challenge, though, because many statistical realities are not intuitively obvious (and some are counter intuitive). This can be overcome with careful reasoning, but it will require your attention; you can look forward to some mental gymnastics.

[1] Read Karl Popper if you do want to get philosophical; it'll be worth it.

# 2 Characterizing a Distribution

In the course of a biological investigation, we almost always make quantitative measurements. In general, our statistical model is that these measurements are samples taken from some kind of random distribution. You've probably encountered problems involving coin tosses, picking colored marbles from an urn, or socks from a drawer in the dark.

## 2.1 Measures of central tendency

The mean is, of course, the most common way to characterize a distribution of sampled measurements:

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

The concept is easy to understand and should be familiar to everyone.

In addition to the mean (or more precisely the arithmetic mean), you are probably also familiar with other measures of 'central tendency'.

The median is the value above which (and below which) 50% of the data is found. The median is less sensitive to outliers than the arithmetic mean. One case where the median can be convenient is when measuring distributions of times before an event occurs; e.g., how long it takes an animal to learn a task. If you want to report the mean, you need to wait for all the animals in your population to learn the task, which could be a really long time if there are one or two particularly dumb animals in your sample population, and may become undefined if one of your animals dies before it learns the task. You can, however, report the median after just over half of your animals learn the task.

The mode is not often used in biological applications of statistics.

## 2.2 Measures of variation

The simplest measure of variation is the range (lowest, highest). The problem with this is that the range will typically vary systematically with sample size; we say it is a *biased* estimate. Contrast to average: your best guess of the mean of the population is the mean of the sample; thus we say the mean is an unbiased estimate of central tendency.

In addition to the mean, the standard deviation and (to a lesser extent) the variance are also commonly used to describe a distribution of values:

$$\begin{pmatrix} Sample \\ Variance \end{pmatrix} = s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1} \tag{2}$$

© *Copyright 2017, 2025 – J Banfelder, L Skrabanek; The Rockefeller University* 

$$\left( \begin{array}{c} Sample \\ Standard \\ Deviation \end{array} \right) = s = \sqrt{\left( \begin{array}{c} Sample \\ Variance \end{array} \right)} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}} \tag{3}$$

Observe that the variance is an average of the square of the distance from the mean. All terms in the summation are positive because they are squared.

$\overline{x}$ and $SD$ have a particular meaning when the distribution is normal. We'll have more to say about this soon.

When computing the variance or standard deviation ($SD$) of a whole population, the denominator would be $n$ instead of $n-1$. The variance of a sample from a population is always a little bit larger, because the denominator is a little bit smaller. There are theoretical reasons for this having to do with degrees of freedom; we will chalk it up to a "weird statistics thing" (it is actually a correction that turns the $SD$ into an unbiased estimate).

Observe that the standard deviation has the same units of measure as the values in the sample and of the mean. It gives us a measure of how spread out our data are, in units that are natural to reason with.

In the physical sciences (physics, chemistry, etc.), the primary source of variation in collected data is often due to "measurement error": sample preparation, instrumentation, etc. This implies that if you are more careful in performing your experiments and you have better instrumentation, you can drive the variation in your data towards zero. Think about measuring the boiling point of a pure substance as an example. Some argue that if you need complex statistical analysis to interpret the results of such an experiment, you've performed the experiment badly, or you've done the wrong experiment.

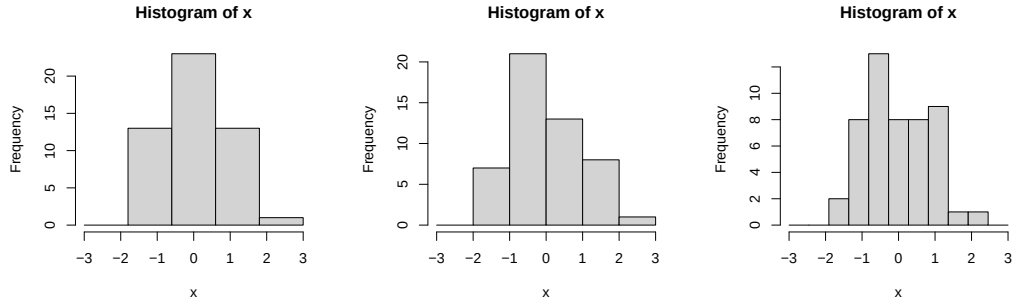
Although one might imagine that an experimenter would always use the best possible measurement technology available (or most affordable), this is not always the case. When developing protocols for CT scans, one must consider that the measurement process can have deleterious effects on the patient due to the radiation dose required to carry out the scan. While more precise imaging, and thus measurements (say of a tumor size), can often be achieved by increasing the radiation dose, scans are selected to provide just enough resolution to make the medical diagnosis in question. In this case, better statistics means less radiation, and improved patient care.

In biological systems, the primary source of variation is often "biological diversity". Cells and patients are rarely identical and will generally not be in identical states, so you expect a non-trivial variation, even under perfect experimental conditions. In biology, we must learn to cope with (and ultimately embrace) this naturally occurring variation.

*© Copyright 2017, 2025 – J Banfelder, L Skrabanek; The Rockefeller University* 

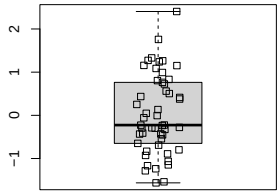
# 3 Graphing a Distribution

Histograms convey information about a distribution graphically. They are easy to understand, but can be problematic because binning is arbitrary. There are essentially two arbitrary parameters that you select when you prepare a histogram: the width of the bins, and the alignment, or starting location, of the bins. For non-large $n$, the perceptions suggested by a histogram can be misleading. Consider the three histograms below; they are all prepared from the same data, but give different impressions of the data.




Although it doesn't solve all such issues, a useful rule of thumb is to use $\sqrt{n}$ bins when preparing histograms.

While histograms can work well when $n$ is large, a better approach when $n$ is small is to simply show all the data.



Here we've also overlaid a box-and-whisker plot. Boxplots usually show quartiles. The heavy bar in the middle is typically the median, not the mean. The box above the median is the third quartile; 25% of the data falls in it. Similarly, the box below the median holds the second quartile.

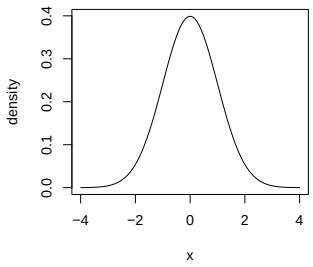
The whiskers in such a plot can be chosen in many ways; when using the "Tukey" method, if the underlying distribution is normal, roughly 1 in 100 data points will fall outside of

© *Copyright 2017, 2025 – J Banfelder, L Skrabanek; The Rockefeller University* 

their range, which usually gives good results for any $n$. These are putative outliers that you may want to inspect further.

# 4 The Normal Distribution

For the rest of today, we'll be considering data that can be treated as being sampled from some normal distribution; i.e., from a bell curve. As this is not a math class, we won't even write down the equation for the normal distribution, but we will sketch the canonical one here, where the mean is zero and the standard deviation is one.

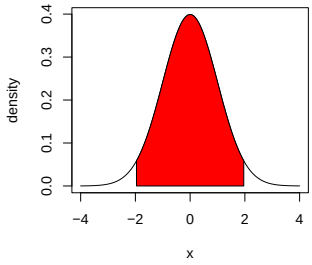



Of course, most of our laboratory measurements won't have a mean of zero and an $SD$ of one; for the normal distribution this is just a matter of shifting the origin and rescaling.

One important point about treating your measurements as coming from a normal distribution, though, is that, in principle, you must admit the possibility of any real value, including negative values and arbitrarily large values. In practice, we don't always hew to this rule; for example, if we were collecting data on the heights of people, we could reasonably model the values as coming from a normal distribution, even though negative heights are not physically realistic. This is safe because the mean is sufficiently far from zero, and the data reasonably clustered around the mean, that negative values are vanishingly unimportant for any reasonable dataset. In contrast, if you've recorded the number of deaths per 100,000 population due to flooding in different counties each year, the mean would be close to zero, and the normal curve for this data would have non-negligible values below zero; thus, the model of a normal distribution would **not** be a good representation of that data.

We mentioned above that $SD$ is a characterization of how spread out your data are. If the underlying distribution is normal and $n$ is large, then 95% of the samples are expected to fall within the range: $\overline{x} \pm 1.96 \cdot SD$, shown in red below. This is one of the few numerical facts you should memorize!

© *Copyright 2017, 2025 – J Banfelder, L Skrabanek; The Rockefeller University* 




## 5 Standard Deviation vs. Standard Error of the Mean

While appreciating the underlying variance in our data is important, we are usually more interested in how well we can estimate the true mean of the distribution we are measuring (sampling) from. This is a very different question that statistics can also help us with. Two factors influence this: how spread out the data are ($SD$), and how much data we have ($n$). A new quantity, the Standard Error of the Mean, is introduced:

$$SEM = \frac{SD}{\sqrt{n}} \tag{4}$$

For large $n$, we can be 95% sure that the true mean of the underlying population is in the range...

$$\overline{x} \pm 1.96 \cdot SEM \tag{5}$$

...where $\overline{x}$ is the sample mean.

Understanding the difference between the $SD$ and the $SEM$ is critical. To reiterate, the $SD$ gives us an indication of how spread out the data in the underlying population are. The $SEM$ is an indication of how confident we are in our estimate of the true mean of the underlying population. If you increase $n$, you expect the $SEM$ to decrease; by collecting more data, you've measured the mean more precisely. However, you have no expectation about how the observed $SD$ will change; the underlying, true $SD$ of the distribution you are sampling from remains unchanged, and the value you happen to compute for your samples from this distribution will not vary systematically with $n$ (because $SD$ is an unbiased estimator).

Many plots in publications show error bars. There is no standard as to what these represent; it could be $\pm SD$, $\pm SEM$, $\pm 1.96 \cdot SD$, $\pm 1.96 \cdot SEM$, or something else. If the publication does not explicitly state what the error bars represent, they are of no use to you (and you might begin to question the underlying analysis). The article by Cumming in your homework addresses this more fully.

© *Copyright 2017, 2025 – J Banfelder, L Skrabanek; The Rockefeller University* 

## 6 Confidence Interval of a Mean

We just said that for large $n$, we are 95% sure that the true mean is in the range $\overline{x} \pm 1.96 \cdot SEM$. It is worth thinking carefully about what this means, as it introduced the essential notion of a confidence interval of a mean. What we are saying here is that if we were to, hypothetically, repeat our experiment many times, collecting new data from the same underlying distribution, then 95% of the time the range $\overline{x} \pm 1.96 \cdot SEM$ that we compute would contain the true mean. This can be challenging to get your head around, in large part because it is difficult to accept that we'll never know the true mean when doing experiments in the lab. But this is what we mean when we refer to a 95% CI of a mean.

When $n$ is not large, things are a bit more uncertain. You can still compute a 95% CI of a mean, but the underlying math is more complex. Fortunately, statistical programs will tend to the details for us. For a single set of measurements assumed to be taken from a normal distribution, we can get our hands on a 95% CI of the mean by asking for a "One sample t-test," or by computing "Descriptive Statistics" for that column.

In fact, you can ask for intervals of any confidence level. While 95% is common, it is completely arbitrary, and there there is often good reason to compute other CIs; 99% and 80% CIs have their uses.

It is important to appreciate how CIs will vary with $n$, the degree of confidence, and the $SD$ of the underlying data. Your homework for this session will explore this interplay.

## 7 Confidence Interval of a Difference Between Two Means

While quantifying how well we have measured a single mean is nice, in "real science", we are usually seeking to measure the effect of some treatment, risk factor, or other variable. This question is usually cast as the effect of the treatment or factor on a mean. So you'll often find yourself computing not just the mean of a set of values, but the difference between the mean of a control group and a test group.

To compute the CI of a difference between two means, first compute the difference between the means:

$$\Delta = \overline{x_A} - \overline{x_B} \tag{6}$$

Even if the treatment or risk factor has absolutely no effect on your measurements, random variation dictates that the measured $\Delta$ in your experiment will probably not be exactly zero. But we can put a CI around the $\Delta$, and if that CI doesn't include zero, then we can conclude that the treatment or risk factor did have some effect. And, the CI puts bounds

© *Copyright 2017, 2025 – J Banfelder, L Skrabanek; The Rockefeller University* 

on the plausible size of that effect. Another way of thinking about this is that the CI gives the range of plausible effects, and, if it includes zero, then no effect at all is also plausible. The strength of the CI (95%, 99%, or 80%, say) governs what "plausible" means.

CIs of differences of means are computed using an "Unpaired t-test".

In the special case where groups are of equal size ($n_A = n_B$):

$$\left( \begin{array}{c} \text{SE of} \\ \text{Difference} \end{array} \right) = \sqrt{\text{SEM}_A^2 + \text{SEM}_B^2} \tag{7}$$

Note that this is simply a triangle rule; it implies that the uncertainty in a sum or difference is more than any one individual uncertainty, but less than the sum of the two uncertainties.

## 8 Paired Studies

The above analysis is applicable when you have two unrelated sets of samples for two different populations. A much more statistically powerful technique can be used when you've performed a paired study. In a paired study, each value in set A has a corresponding value in set B. Often, paired studies are before-and-after studies, where measurements are taken on the same subject before and after a treatment. It offers much more statistical power because you are able to factor out much of the biological diversity in the population.

When working with data from paired studies, you should compute a $\Delta$ for each pair of subjects, then compute $\overline{\Delta}$ and its CI using the techniques for a single distribution. This bookkeeping is often done for you under the title of a "Paired t-test".

© *Copyright 2017, 2025 – J Banfelder, L Skrabanek; The Rockefeller University* 

# 9 Homework

- Read Cumming, Fidler, and Vaux. Error bars in experimental biology. *J Cell Biol.* 2007 Apr 9; 177(1): 7-11.
- Install GraphPad's Prism software. You'll be given license information.
- Explore the relationship between $n$, confidence level, and the width of a CI for both univariate and unpaired continuous data.

1. Assume that the true mean of the heights of people in a population is 68 in, and the SD is 2.6 in (n.b., in "real life" you won't ever know the true mean; but by simulating data in which you do can help get your head around statistical concepts.)

   (a) Generate a random sample of 12 datapoints simulating measuring heights of people sampled from this population.

   – Hint 1: In Prism, generating random data is considered an Analysis; so choose Analyze and scroll down to "Simulate Data," and then choose "Simulate column data".

   – Hint 2: Prism doesn't store analysis results; it recomputes them from your data whenever anything changes. In the case of randomly generated data, this can be annoying, as your data may change when you least expect it. So freeze your "results" as soon you generate them (use the snowflake on the "Sheet" section of the menu bar).

   (b) Analyze this data with both a descriptive statistics and a one-sample t-test analysis. Report the 95% CI of the data. In your case, is the true mean within the reported range? Hint: pay attention to the Hypothetical value part of the Experimental Design of the t-test.

   (c) There are 31 students in our class this year, plus two instructors, who've done this exercise. Assume everyone worked independently and generated different random datasets. Very roughly (use your intuition), how many times do you expect the true mean to be outside of the computed 95% CI?

   (d) Redo your analysis on the **same data**, but this time compute the 80% CI. Is the interval wider or narrower than the 95% CI? Explain why this is so?

   (e) Similarly, compute the 99% CI? How wide is this interval?

2. In a different population, the true mean for the heights of men is 65.9 inches, and the true mean of the heights of women is 64.8 inches; use the same SD as before.

© *Copyright 2017, 2025 – J Banfelder, L Skrabanek; The Rockefeller University* 

(a) Generate random data that simulates measuring 12 heights from each group (24 samples in all); plot the data showing all data points and nothing else (no error bars, means, etc).

(b) Looking at just the data and using your intuition, is there sufficient evidence to conclude that the mean height of men is greater than the mean height of women in your population?

(c) Perform an unpaired t-test to obtain the 95% CI of the difference between the heights of men and women. Does this interval include zero? How do you interpret this finding?

3. Try the same comparison, but with 10,000 samples from each population (20,000 measurements in all).

(a) What is the 95% CI of the difference between the means? Does this interval include zero? How do you interpret this finding?

(b) Based on your reading of Cumming, et al., prepare a plot that uses error bars that you think most appropriately tells the story of the last analysis. Explain the rationale for your choice.

Submit your work as a Prism file and an accompanying write-up as a PDF; no Microsoft Word documents, please! Work in groups of two or three, and hand in the homework as a group by emailing your two files to both Luce and Jason. Be sure that the names of all group members are included in the write-up. And do make sure everyone in the group learns how to drive Prism to get these results.

© *Copyright 2017, 2025 – J Banfelder, L Skrabanek; The Rockefeller University* 