Statistics in scientific papers

Doug Kieffer

Experiments are set up to test various **treatments** on different **populations**. For example, the effect of increased light (treatment) on stem elongation of poinsettias (population). Obviously, every poinsettia in the world can't be checked so experiments are run on **representative samples**. For a sample to be representative, it must be selected **randomly** from the entire population. There are many things that must be considered when designing an experiment to assure this condition is met.

After the treatments are applied, the response is measured. In the above example, the response would be the length of the stem. Of course, even under perfectly controlled light, irrigation, temperature, … conditions, not every poinsettia plant will have the exact same length stem. This is due to the genetic makeup of each individual plant. In nature, it is assumed that the distribution of a population will follow the **Normal Distribution** curve (see fig. 1). That is, most of the population will be centered around the average length but there is a finite probability that, under normal conditions, members of the population will be noticeably larger or smaller than the mean.

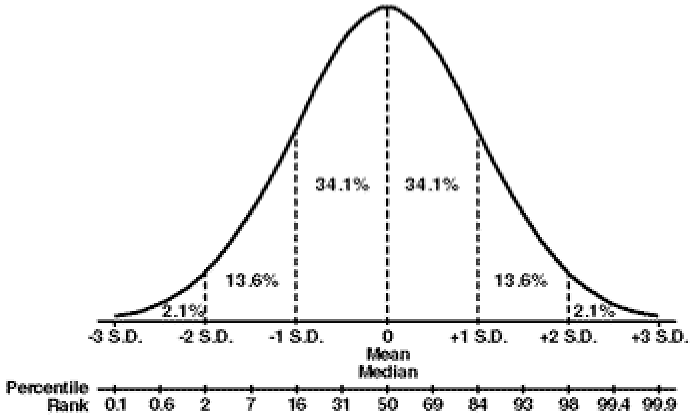

Figure 1. Normal Distribution Curve

The next thing to consider is how dispersed the data is. In other words, does most of the population fall very close to the mean or is there a great deal of variation. If it's the former, the Normal Distribution curve will be squeezed and skinny. If it's the latter, the curve will be spread out. The dispersion is measured by a parameter called the **standard deviation** (s.d.). In a normally distributed population, we would expect 34% of the measurements to be 1 s.d. greater than the mean and an equal number to be 1 s.d. less than the mean. This means that roughly 2/3 of the population is within 1 s.d. of the mean. In our example, assume the mean stem length is 30cm and the s.d. is 1cm. We would expect that 68% of all stems would be between 29 and 31 cm long. Further, from figure 1 we see that 95% of the members should be within 2 s.d's of the mean. That would mean

95% of our stems are between 28 and 32cm. Looked at another way, there is only a 5% chance we would find a stem shorter than 28cm or longer than 32cm. It's possible, but not likely.

Where do these numbers come from? They come from the data. Let's say the poinsettia experiment had 2 treatments.
1. Control - Poinsettias getting light from sun shining through the greenhouse walls
2. Light - Poinsettias also receiving supplemental light from a sodium halide lamp.

These treatment groups represent 2 populations that we assume have been selected randomly. In other words, we didn't pick all the skinny poinsettia seedlings and put them in the Control group and put all the vigorous seedlings into the Light group. This would bias the results.

Mathematical operations on the stem length data for each treatment can be done to determine the shape of the normal distribution curve for each which includes information on the treatment mean and standard deviation. The next question is how different the means have to be from one another before we can say that the Light treatment was **significantly** different than sunlight alone. The term "significance" is an important term in statistical analysis so care should be taken when using it. Just because two values are <u>noticeably</u> different, doesn't mean they are <u>significantly</u> different. And, numbers that may look pretty similar may, in fact, have statistically significant differences. The only way to prove this to another researcher is by doing the math.

The first step is to formulate the **null hypothesis**. The null hypothesis assumes the status quo. In our example, the null hypothesis is that the stem lengths are the same with and without supplemental lighting. If the math shows that the null hypothesis must be rejected (that stem lengths <u>are</u> different), then we conclude the lighting <u>did</u> have an effect. The next question is how certain we want to be.

Let's say we took the data and drew out the normal distributions for both the Control and Light treatment. We may get something like figure 2. Let's, again, assume the Control mean is 30cm and the standard deviation is 1cm. What we need to do now is compare the mean from the Light treatment to this. Let's say the mean length of poinsettia stems subjected to supplemental lighting was found to be 32.5cm. If this mean is only within one standard deviation of the Control mean, we're only 68% sure that this difference isn't just due to random genetic differences in the plant. That's not that great. If it's within two s.d's, we're 95% sure (5% chance of an error). Not bad. And if it were within 3 s.d.s, we're more than 99% sure. Great! The term used here is "significance level". This is usually identified as a **p** value. The p-value is the chance of making an error and is usually expressed as a decimal rather than a percent. So, a p-value of 0.05 means we have a 5% chance of being wrong. A p-value of 0.01, means a 1% chance of being wrong. This is a stricter definition. In our example, we reject the null hypothesis (meaning the stem lengths are <u>not</u> the same) with a significance level of $p = 0.05$, but fail to reject the hypothesis if $p = 0.01$. The last part of that sentence is important. Failing to reject the hypothesis is not the same as accepting the alternative. In other words, just because we can't say the two treatments are different does not, statistically speaking, mean they are the same. It just means the data doesn't support the conclusion.

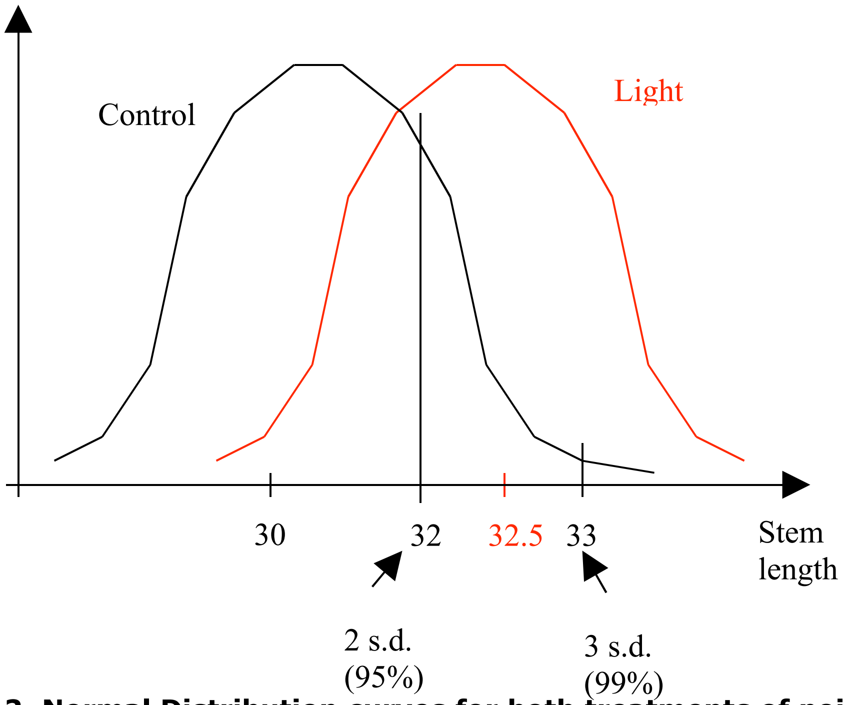

Figure 2. Normal Distribution curves for both treatments of poinsettia example

It is often the case that several treatments are tested at one time. In our example, this could be different levels of increased lighting. The question then becomes whether each of these treatments are different from one another and/or different than the Control treatment. The most common tool for testing for significant differences (sometimes called "separation of means") of multiple treatments is by ANalysis Of VAriance (**ANOVA**). The ANOVA table produces a parameter (the **F**-value) that, coupled with the p-value, allows us to say if one treatment is significantly different from another. The data is commonly presented in tabular or bar-graph form with the treatment means accompanied by lower-case letters. All means accompanied by the same letter are considered to be not statistically different (see table 1).

Treatment (hours of suppl.light supplied)	Mean Stem Length (cm)
0 (Control)	a 30
1	a 30.2
2	b 32.5
3	c 35.7
4	c 36.1
5	d 39

Table 1. Sample table showing which values are significantly different.