

## Senior Science Seminar

### The 'Good Data' Handout

*How to get the most from your experimentation.*

**Data is “collected before you collect it.”**

In other words, how the sampling is done will determine the quality of the data. How many times will you sample at a given time? How many samples do you need to collect? In what manner does the experimentation need to be repeated?

**Planning your experiment: Draw an imaginary graph with pretend data in it.**

This helps visualize the axis of the graph, and really helps with the experimental design. Nope, this isn't your actual data. Your actual data will be placed inside a similar, “blank” graph as you collect your data.

All experiments need proof that they were actually done—that is photographic images of the data so that we know you didn't make it up. Some of these images need to be in your presentation.

**Do other people like your imaginary graph?**

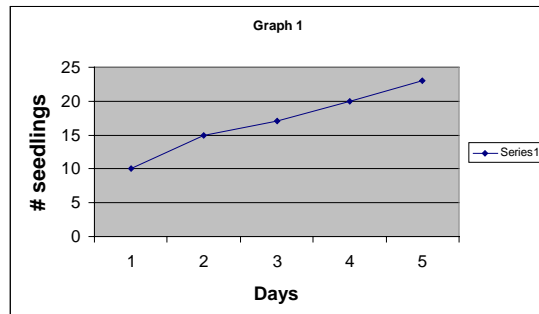
Do they agree with you that this is good data to collect?

**Statistics.**

OK, now that we have an imaginary data set, let's consider how many times we need to sample (collect the data) and if we need to repeat our experiments.

**The Beauty of Data over time.**

Are you watching the leaves turn color or counting the number of green jelly beans in a jar? Are you collecting the same data a few days later, then a few days later, and so on for an extended duration? This type of data is internally controlled providing it shows a trend.



Let us say that you have planted some seeds, and graph 1, above, indicates how many of the seeds have germinated. the X axis on Graph 1, above, is the # of seedlings viewed in a pot. The Y axis is “days.” It appears that in this pot, there are more seedlings sprouting from the soil, as “time goes on.”

Is this a good data set? Yes and No. It is good data, and clearly shows a trend. To tell you the truth, the data is enough to tell me that I believe your seeds are germinating. However, in science, we need to be more confident about the data. Therefore, we have the option of repeating the experiments and showing both data sets, or, to save time, doing parallel experiments.

## Parallel Experimentation.

Frequently, parallel experimentation gets us out of repeating experiments. I have read many papers that have been accepted to very prestigious medical journals that have not repeated their study. It is just too hard to do, if not impossible since the original experimental conditions of the trial were unique. However, these studies must be set up in very believable ways. In many respects, it is important to repeat even these prestigious and accepted data sets, and frequently the data is repeated by doing other trials that will be reported by scientists as a separate body of work.

Regarding our seedling experiment, we therefore plant a total of 3 pots with seeds and see if we get the same number or a similar number of seedlings on the same days. It is very important that we treat all of our pots the same way and that we do not water a pot more than the others or put one in the window-sill and the others in garage under a grow-lamp. The Pots should be alongside one another in the window-sill, receiving the same amount of light. They should also receive exactly the same amount of plant food and water. There are now 3 separate sub-experiments that are being doing as a part of the larger data set.

Our data now looks like this:

| Day | Pot # 1 | Pot # 2 | Pot # 3 |
|-----|---------|---------|---------|
| 1   | 10      | 11      | 13      |
| 2   | 15      | 17      | 15      |
| 3   | 17      | 18      | 18      |
| 4   | 20      | 21      | 21      |
| 5   | 23      | 24      | 24      |

How do we analyze this data? By making a similar graph as shown above. We can take the average of the # of seedlings in the three pots at a given day, which gives us the following values:

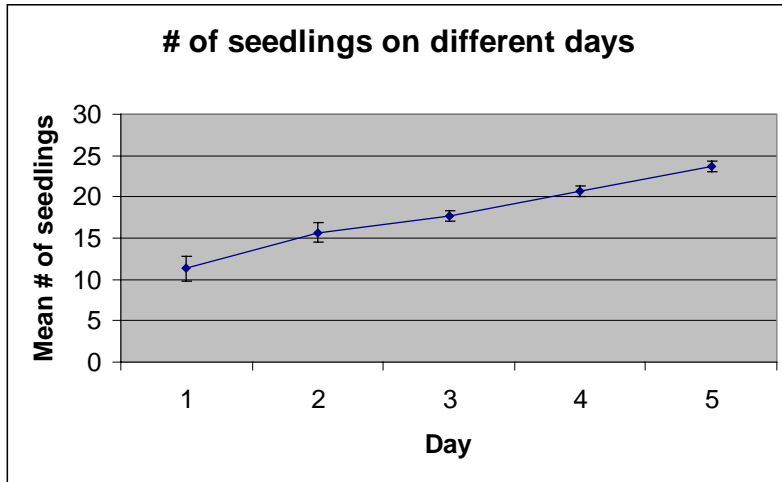
| Day | Mean:    |
|-----|----------|
| 1   | 11.33333 |
| 2   | 15.66667 |
| 3   | 17.66667 |
| 4   | 20.66667 |
| 5   | 23.66667 |

And, we would still see the trend as shown in the first graph. However, the data isn't really good enough for presentation, because it does not take into account the differences between the 3 pots. This information can be visualized, by calculating the standard deviation of the number of seedlings at each of the days. Our data will then be:

| Day | Mean:    | SD:      |
|-----|----------|----------|
| 1   | 11.33333 | 1.527525 |
| 2   | 15.66667 | 1.154701 |
| 3   | 17.66667 | 0.57735  |
| 4   | 20.66667 | 0.57735  |
| 5   | 23.66667 | 0.57735  |

The easiest way to calculate Standard Deviation is using excel. Enter =STDEV(B41:D41) , where B41,C41 and D41 are the "cells" that have the number of seedlings in the pot on a given day. I'll show you how to do this in the computer lab. The formula for calculating the average of these cells is =AVERAGE(B41:D41) in excel. I used the Excel program to generate these graphs.

Our graph will now look like this:



Can you build this graph using Excel? This will be one of our objectives during the computer lab visit, is to construct this graph.

**Here is some helpful information regarding Means and Standard Deviation, that you should understand:**