Data Analysis[6]

The major objective of the Plant Biology Labs is to develop your observational skills. But in addition, in some exercises you will be collecting quantitative data. For these lab exercises, you may be required to analyze data. The following is a brief description of the analytical and statistical analyses you may be required to do for your lab write-ups.

Data Distributions. One of the most useful analytical tools used to explore the nature of data is the "stem and leaf" diagram (Tukey, 1977). Suppose we have a set of measurements of O_2 production from the Photosynthesis lab. These are data compiled from all lab groups for a particular light treatment (*please be advised that the data are fictitious and may bear no similarity to your 'real' data*).

In ml g^{-1} h^{-1}: $0.5, 0.6, 0.8, 1.8, 0.4, 0.7, 0.7$

To create a stem and leaf, we usually sort them first, from lowest to highest value:

$$0.4, 0.5, 0.6, 0.7, 0.7, 0.8, 1.8$$

create the **stem**:

```
0 (0.4 to 0.6)
0 (0.7 to 0.9)
1 (1.0 to 1.2)
1 (1.3 to 1.5)
1 (1.6 to 1.8)
```

and add the data to the right (the leaf):

```
0 456
0 778
1
1
1 8
```

Notice that most of the data are clustered between 0.4 and 0.8. One datum is not: 1.8.

Central Values. So, how do we estimate the central value of the data? We can use the **median**. This is the halfway point, where half the data are smaller and half the data are larger. For the sorted data, the median is 0.7 (the 4th ranked of the seven values, sorted from smallest to highest). More generally, once your data are sorted, the median is the nth datum calculated from 0.5(1+N), equal to 4 in our case. If the sample size is an even number, you take the mean of the two middle values to obtain the median.

We can also use the **mean**, which is the sum of all the data, divided by the number of data (sample size, N). The general formula is:

$$\bar{x} = \frac{\sum_{n=1}^{n=N} x_n}{N}$$ where x_n are the data, N is the sample size, and $\bar{x}$ is the mean.

For this data:

$$\frac{(0.4 + 0.5 + 0.6 + 0.7 + 0.7 + 0.8 + 1.8)}{7}$$

 equals a mean of 0.79

---

[6] Developed by Roger Lew, Sapna Sharma and Christopher Luszczek.

Data Spread (Variability). It is often useful to estimate how variable the data are. This can be variability due to experimental error, or variability that exists naturally in the population being analyzed. For our data, the extremes —0.4 and 1.8— already give us an idea, but to measure more precisely how clustered the data are, **hinges** are useful. These are the values halfway between the median and the extremes. For our data, the hinges are 0.55 and 0.75. That is, most of the data are clustered between 0.55 and 0.75. More generally, the hinge values can be determined by first determining the ranked value equal to the median (in our case the 4th value, from 0.5(1+7), then the rank of the hinge values is calculated from 0.5(1 + 4) = 2.5. You can check the ranking procedure with the following example (pre-sorted):

$$0.3, 0.5, 0.5, 0.6, 0.7, 0.8, 1.0, 1.8$$

The median is 0.5(1 + 8) = 4.5th rank: 0.65. The hinges are 0.5(1 + 4.5) = 2.75th rank. Rounded to the 2.5th rank (that is, between the 2nd and 3d rank), the hinges are 0.5 and 0.9.

For this data set, the stem and leaf diagram (with median M and hinges H shown) looks like this:

```
0    3
0  H 55
0  M 67
0  H 8
1    0
1
1
1
1    8
```

A more complex estimate of the spread is the **standard deviation**. This is an estimate of population variance, but be careful, this is only true for so-called "Normal" distributions. The formula for the calculation of the standard deviation (squared, s^2) is:

$$s^2 = \frac{\sum_{n=1}^{n=N}(x_n - \bar{x})^2}{N-1}$$ where x_n are the data, N is the sample size, and $\bar{x}$ is the mean.

This can be simplified somewhat:

$$s^2 = \frac{\sum_{n=1}^{n=N}x_n^2 - \frac{\left[\sum_{n=1}^{n=N}x_n\right]^2}{N}}{N-1}$$

Although many wouldn't consider this much of a simplification, it does make the calculation easier! Now, most calculators can calculate the standard deviation (*s*) automatically. But, should your batteries run out, the calculation for our data is:

$$\sum x^2 = (0.4)^2 + (0.5)^2 + (0.6)^2 + (0.7)^2 + (0.7)^2 + (0.8)^2 + (1.8)^2$$

equal to 5.63. and

$$\left[\sum x\right]^2 = [0.4 + 0.5 + 0.6 + 0.7 + 0.7 + 0.8 + 1.8]^2$$

equal to 30.25, divided by 7 (our sample size) is equal to 4.32. Now we can take these values (5.63 and 4.32) and calculate $s^2 = [(5.63-4.32) / 6] = 0.22$. Taking the square root: the standard deviation, $s = 0.47$.
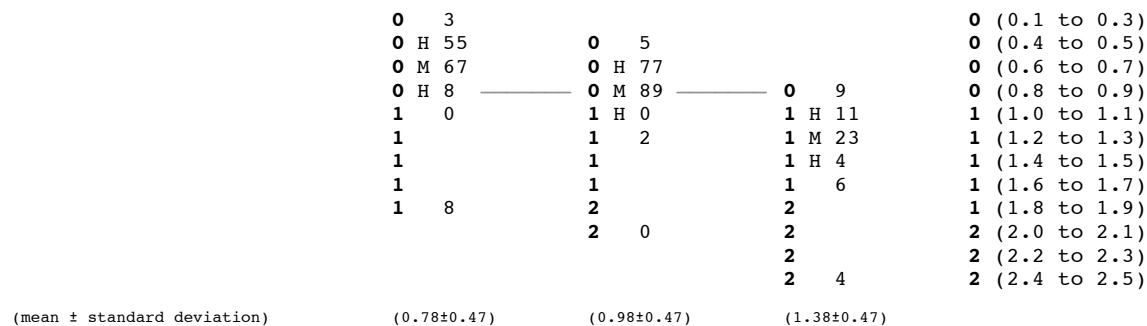
So, the mean ± standard deviation is: 0.79 ± 0.47 (n = 7). Note that it is very helpful to indicate the sample size.

For the purposes of Plant Biology lab exercises, the mean ± standard deviation would be considered useful summary statistics. Stem and leafs are more useful in exploratory analysis. For example, in our data, that 1.8 really falls outside the clustered distribution. Why? Is it experimental error? A new discovery?

Evaluating Differences — graphical comparators. Much of statistical analysis is devoted to comparisons of two (or more) datasets to determine if the datasets are different. Evaluating statistical differences is complex: In scientific work, t-tests and ANOVA (<u>An</u>alysis <u>of</u> <u>V</u>ariance) are the traditional techniques. These techniques can fail when the sample number is small, when the populations are not "Normal" distributions, or when the data spreads of the two (or more) datasets are different.

The hinges of the stem and leaf provide us with a natural measure of the spread of the data. If two datasets are different, then the hinges are unlikely to overlap. But be warned, unlike the traditional techniques (t-tests and ANOVA), the 'non-overlap' does *not* translate into a statistical metric, it is *only* a graphical comparator.

For example, the first two datasets are *not* different, the first and third *are*:

```
0    3                                              0  (0.1 to 0.3)
0 H 55         0    5                               0  (0.4 to 0.5)
0 M 67         0 H 77                               0  (0.6 to 0.7)
0 H 8   ─────  0 M 89  ─────  0    9                0  (0.8 to 0.9)
1    0         1 H 0         1 H 11                 1  (1.0 to 1.1)
1             1    2         1 M 23                 1  (1.2 to 1.3)
1             1             1 H 4                   1  (1.4 to 1.5)
1             1             1    6                   1  (1.6 to 1.7)
1    8         2             2                       1  (1.8 to 1.9)
               2    0         2                       2  (2.0 to 2.1)
                             2                       2  (2.2 to 2.3)
                             2    4                   2  (2.4 to 2.5)
```

(mean ± standard deviation)   (0.78±0.47)   (0.98±0.47)   (1.38±0.47)

Similarly, for the summary statistics of mean ± standard deviation (shown above), if the ranges of the two datasets (from the mean *minus* the standard deviation to the mean *plus* standard deviation) do not overlap, than it is likely that the results are different. Using this criterion, the first and third datasets are *not* different. This may be because the standard deviation is affected by the outlying value, while the hinge values are not.

For either approach, the comparisons are neither quantitative nor definitive. Even so, the simple techniques of comparing data spreads to evaluate differences are easy to perform, and, offer a robust approximation that will help you as you consider how different your data are.

Evaluating Differences — statistical tests. Biologists often rely upon a more formal comparative technique — the t–test— to evaluate differences quantitatively. This is especially useful for ecologists, who often work with large datasets with inherent variability. But it is also important for bench scientists who want to test well-defined hypotheses. For example, do two genetic strains of an organism grow at the same rate (or not)? Does a phytochemical cure cancer? In these examples, the *null* hypotheses would be that the strains grow at the same rate, and cancer rates are the same with or without treatment with the phytochemical. The t-test provides a

way to test whether differences are significant. As implied above when explaining graphical comparators, the t-test statistic is complex: it assumes a normal distribution, and readily fails when samples sizes are small. But it is a common statistical test that is traditionally used by biologists, so students should understand and be able to use it.

In a formal way, the t-test compares two means, say $\overline{X}_A$ and $\overline{X}_a$. The simplest way to do this is to subtract one from the other $\overline{X}_A - \overline{X}_a$. But, this doesn't account for the variability (spread) in the data. To account for variability, we can divide by the standard deviation: $\dfrac{\overline{X}_A - \overline{X}_a}{s}$ to normalize to the variability. This tells us how many standard deviations there are between $\overline{X}_A$ and $\overline{X}_a$. But the equation has to be cast in a more complete form that accounts for the variability of *both* means, and, the sample size: $t = \dfrac{\overline{X}_A - \overline{X}_a}{\sqrt{\frac{1}{2}(s_{X_A}^2 + s_{X_a}^2)} \cdot \sqrt{\frac{2}{n}}}$. This is simpler than it looks! To compare two data sets, we compare the means ($\overline{X}_A$ and $\overline{X}_a$) but must also account for the variability of both samples ($\sqrt{\frac{1}{2}(s_{X_A}^2 + s_{X_a}^2)}$ accounts for both standard deviations) and how much sampling we have done ($\sqrt{\frac{2}{n}}$, n is the sample size). Now, the larger t is, the greater the difference between the two samples, and the more likely the samples are different. The equation states that the greater the difference in $\overline{X}_A$ and $\overline{X}_a$, the larger t will be, that small standard deviations will make t larger (since they are in the denominator), and that the larger the sample size, the larger t will be. A graphical explanation should help (figure 1).
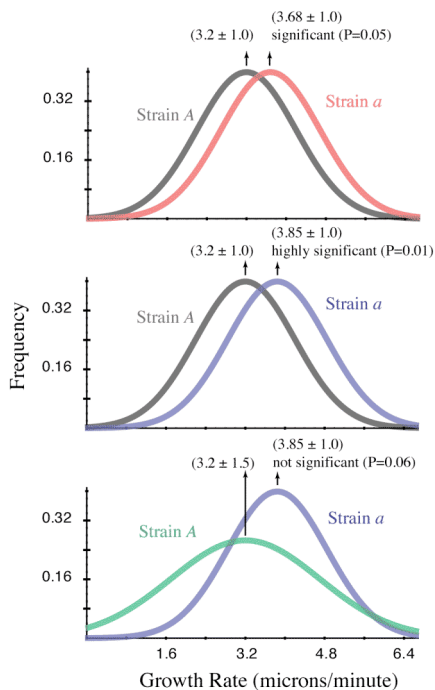

Fig. 1. Graphical Explanation of the t-test. The curves are normal distributions for the growth rates of two strains: A (3.2 microns/minute) and a (3.68 and 3.85 microns/minute). Are the growth rates of the two strains different? The probabilities were calculated from t assuming a sample size of 20. 3.68 is significantly faster (**top panel**), and 3.85 even more so (**middle panel**). But if we increase the variability of the growth rate of strain A from 1 to 1.5, neither of the a strains grows significantly faster (**bottom panel**). In other words, significant differences depend on how much faster, and how much variability there is in your data.

The value of t tells you how different the two samples are. But this is usually cast in the form of a probability (which lies between 0 to 1). To convert to a probability, you have two options. You can use a software program —SPSS is one example, even excel can do it— or a table of t values (Table I) to calculate the probability.

Evaluating Differences —constructing a hypothesis. When using statistical tests of differences, it is absolutely crucial to construct a formal hypothesis. This is especially true when working with natural populations —including clinical research on humans— because of the inherent population variability that exists. A carefully constructed hypothesis minimizes the risk of research that can't be replicated. An overview of hypothesis testing is shown in figure 2.

Fig. 1. Graphical Explanation Hypothesis Testing. The left diagram explains the process. Your TAs will provide you with a more formal explanation. The right diagram explains the importance of replicating your experimental results, courtesy of Randall Munroe at xkcd (http://xkcd.com/242/).
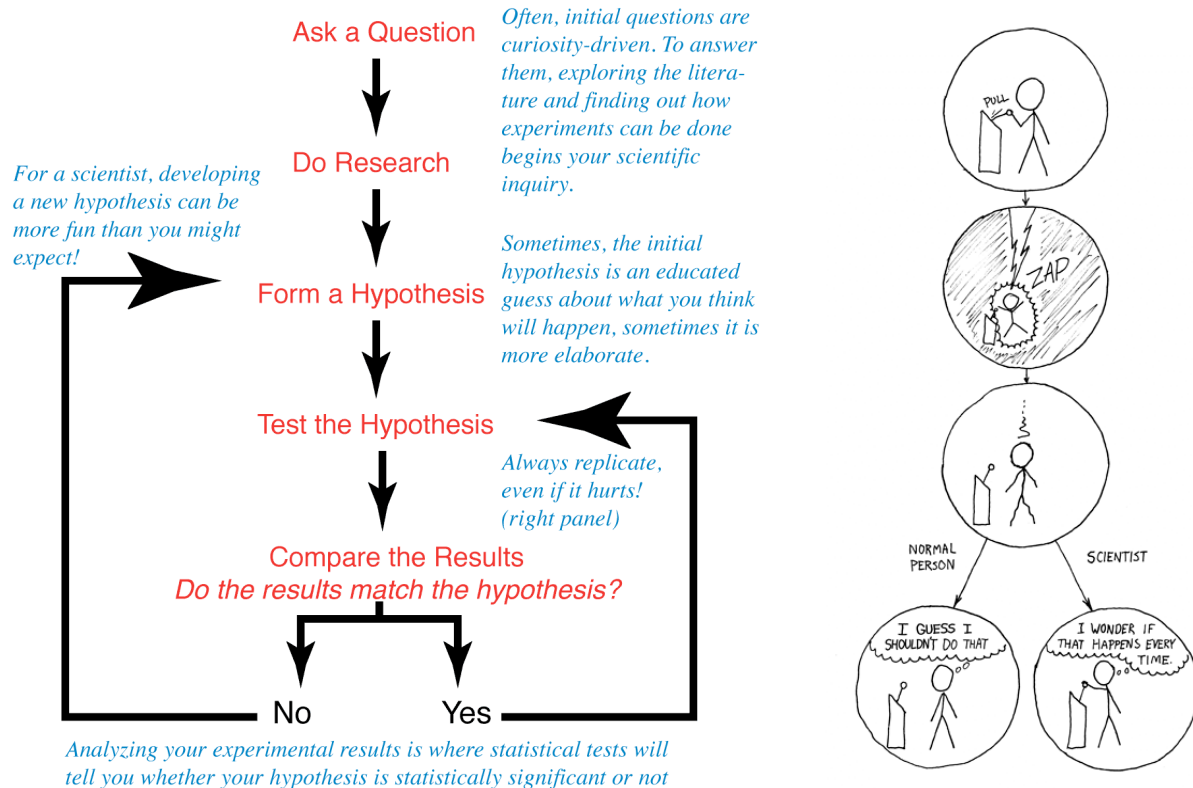
t–Test Recipe. There are at least 7 steps that one can follow to perform a statistical test. We will review each of the steps using a hypothetical example and you will perform the same procedure on your data.

 1. What is the appropriate statistical test to use? There are numerous statistical tests that one can perform. However, you need to be aware of the type of data that you have and what your null hypothesis is to help you identify which statistical test is appropriate for your data.

You will use a t-test when you have continuous data ($-\infty$ to ∞) and you would like to compare the means of two populations. For example, you could be interested in knowing the difference in heights, weights, or growth rates of different genetic strains.

Example. The example that we will follow for this exercise is comparing growth rates of different mating types of the *Neurospora crassa* fungus. The data are as follows (and are hypothetical):

Growth rates of mating type A (cm h^{-1})	6	8	7
Growth rates of mating type a (cm h^{-1})	2	1	3

2. What is the statistical significance (P) at which I will evaluate my data? The P value at which you test statistical significance is a pre-determined value. In biology, the convention to follow is to use an P value of 0.05 which suggests that there is only one chance in twenty that a situation was a coincidence.

3. What are your statistical hypotheses?

A. <u>Null hypothesis</u>: The null hypothesis usually states that there is **NO** difference between the two groups, such that the mean of group 1 is equal to the mean of group 2. If you are doing an experiment and you are looking at the effect of a treatment on the group of individuals, the null hypothesis would state that there is no effect of the treatment on the means of the individuals in the study before and after the treatment. So, for the fungal growth rates, the null hypothesis would be that there are no differences in the growth rates of A and a mating type strains.

> *What would be the null hypothesis for the following research questions:*
> • Will tomato plants grown with or without sunlight be the same height?
> • Does attending class have an influence on student performance on exams?

B. <u>Alternate hypothesis</u>: The alternate hypothesis usually states that there is **A** difference between the two groups, such that the mean of group 1 is **not** equal to the mean of group 2. If you are doing an experiment and you are looking at the effect of a treatment on the group of individuals, the alternate hypothesis would state that there is an effect of the treatment on the means of the individuals in the study before and after the treatment. So, for the fungal growth rates, the null hypothesis would be that there are differences in the growth rates of A and a mating type strains.

> *What would be the alternate hypothesis for the following research questions:*
> • Will tomato plants grown with or without sunlight be the same height?
> • Does attending class have an influence on student performance on exams?

4. What is the calculated t-statistic? To calculate the t-statistic, you will need to know the mean ($\overline{X}$), variance (S^2), and sample sizes (n) of the two groups:

$$ t = \frac{\overline{X}_A - \overline{X}_a}{\sqrt{\frac{1}{2}(s^2_{X_A} + s^2_{X_a}) \cdot \sqrt{\frac{2}{n}}}} = \frac{\overline{X}_A - \overline{X}_a}{\sqrt{(\frac{s^2_A}{n_A} + \frac{s^2_a}{n_a})}} $$

Remember that the formula for the mean is: $\overline{x} = \dfrac{\sum\limits_{n=1}^{n=N} x_n}{N}$ and for the variance is $s^2 = \dfrac{\sum\limits_{n=1}^{n=N}(x_n - \overline{x})^2}{N-1}$

Example: Let's try this out with our hypothetical dataset.

First, let's calculate the values we will need for mating type A:
  1. Mean of mating type A = (6 + 8 + 7)/3 = 21/3 = 7
  2. Variance of mating type A = ((6-7)^2 + (8-7)^2 + (7-7)^2)/3-1 = (1 + 1 + 0)/2 = 2/2 = 1
  3. n = 3

We do the same for mating type a.
  1. Mean of mating type a = (2 + 1 + 3)/3 = 2
  2. Variance of mating type a = ((2-2)^2 + (1-2)^2 + (3-2)^2)/3-1 = (0 + 1 + 1)/2 = 2/2 =1
  3. n = 3

Now, let's put it all together into the formula to calculate the t-statistic:

$t_{calculated}$ = (7-2)/sqrt (1/3+ 1/3)
= 5/0.81
=6.17

5. What are the degrees of freedom? It is the number of values in the final calculation that are free to vary; a measure of independence between points. The formula to calculate degrees of freedom = n–1; where n is the number of observations in the groups. In our example, we have 2 groups with three individuals each.

Degrees of freedom = 3-1 = 2

6. Compare the calculated t-statistic against the critical t-statistic (Table I).

- If $t_{calculated}$ > $t_{critical}$ —Reject the Null hypothesis (that is, no difference in growth rates).
- If $t_{calculated}$ < $t_{critical}$ —Accept the Null hypothesis.

Example: Our t-calculated value is: 6.17
Based on 2 degrees of freedom and a significance level of 0.05, the t-critical value is 4.30 (Table I).

Since t-calculated is greater than t-critical, we reject our null hypothesis.

Degrees of freedom	P=0.05	P=0.025	P=0.01	P=0.005
1	12.71	25.45	63.66	127.32
2	4.30	6.20	9.92	14.09
3	3.18	4.17	5.84	7.45
4	2.78	3.50	4.60	5.60
5	2.57	3.16	4.03	4.77
6	2.45	2.97	3.71	4.32
7	2.36	2.84	3.50	4.03
8	2.31	2.75	3.36	3.83
9	2.26	2.68	3.25	3.69
10	2.23	2.63	3.17	3.58
11	2.20	2.59	3.11	3.50
12	2.18	2.56	3.05	3.43
13	2.16	2.53	3.01	3.37
14	2.14	2.51	2.98	3.33
15	2.13	2.49	2.95	3.29
16	2.12	2.47	2.92	3.25
17	2.11	2.46	2.90	3.22
18	2.10	2.44	2.88	3.20
19	2.09	2.43	2.86	3.17
20	2.09	2.42	2.84	3.15
21	2.08	2.41	2.83	3.14
22	2.07	2.41	2.82	3.12
23	2.07	2.40	2.81	3.10
24	2.06	2.39	2.80	3.09
25	2.06	2.38	2.79	3.08
26	2.06	2.38	2.78	3.07
27	2.05	2.37	2.77	3.06
28	2.05	2.37	2.76	3.05
29	2.04	2.36	2.76	3.04
30	2.04	2.36	2.75	3.03
40	2.02	2.33	2.70	2.97
60	2.00	2.30	2.66	2.92
120	1.98	2.27	2.62	2.86
infinity	1.96	2.24	2.58	2.81

Table I. Critical t-statistic values. The left column lists the degrees of freedom. The columns to the right list the critical t-statistics for the probabilities at 0,05, 0.025, 0.01 and 0.001. Biologists tend to use a probability of 0.05, describing it as statistically significant.

For the *example* of mating type strains A and a, the calculated t-statistic was 6.17 and the critical t-statistic is 4.30 (the P=0.05 column for 2 degrees of freedom). Since the calculated t-statistic is greater than the critical t-statistic, the growth rates are significantly different at the 0.05 level.

7. What is the biological interpretation of your statistical analysis? Here, we relate the rejection or acceptance of the null hypothesis back to our study. If we accept the null hypothesis, we can say that there is no statistical difference between the means of the two groups. If we reject the null hypothesis, we would state that there is a statistical difference between the two groups.

Example: We rejected our null hypothesis. Therefore, there is a statistical difference in the growth rates between A and a mating types of the fungus. The growth rates of mating type A are faster than the growth rates of mating type a.

Exercise 1. Follow the 7 steps to determine if there's a statistical difference in the following hypothetical dataset. The research question is: is there a statistical difference between the heights of two trees, X and Y?

Height of Tree X (m)	Height of Tree Y (m)
1	5
2	6
3	7
4	8

1. What is the appropriate statistical test to use?
2. What is the statistical significance (α) at which I will evaluate my data?
3. What are your statistical hypotheses?
   A) Null hypothesis
   B) Alternate hypothesis
4. What is the calculated t-statistic?
5. What are the degrees of freedom?
6. Compare the calculated t-statistic against the critical t-statistic.
7. What is the biological interpretation of your statistical analysis?

Exercise 2: Follow the 7 steps to determine if there's a statistical difference in your dataset.

Summary. There are many ways to analyze data. Some, like the stem and leaf, provide you with a straightforward way to look at how your data are distributed, and even graphical comparisons of two (or more) datasets. This technique is part of an analytical technique called exploratory data analysis. Often, scientists need to use more quantitative techniques. One of these —the t-test— is very common. The analytical techniques that you may use in your lab exercises will vary from year to year. Your TAs will provide you with the additional information you need to analyze data you obtain in your experiments.

Reference: Tukey, J.W. (1977) Exploratory Data Analysis. Addison-Wesley Publishing Company. pp. 27–32.