Box plots & t-tests

Box Plots
Box plots are a graphical representation of your sample (easy to visualize descriptive statistics); they are also known as box-and-whisker diagrams. Any data that you can present using a bar graph can, in most cases, also be presented using box plots. A box plot provides more information about the data than does a bar graph.

Things to know about box plots
- Your sample is presented as a box.
- The spacings between the different parts of the box help indicate the degree of dispersion (spread) and skewness in the data, and identify outliers.
- A box plot shows a 5-number data summary: minimum, first (lower) quartile, median, third (upper) quartile, maximum.
- The box is divided at the median.
- The length of the box is the **interquartile range** (IQR).
- The 1st quartile is the bottom line.
- The 3rd quartile is the top line.

Example
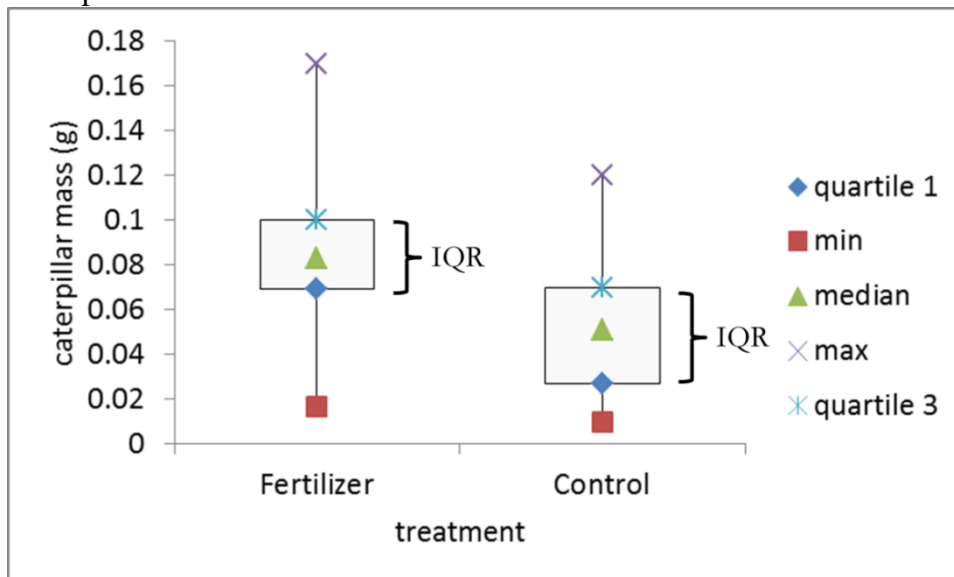
Quartiles divide frequency distributions
- Q_1 :1st or lower quartile: cuts off lowest 25% of the data
- Q_2 :2nd quartile or median: 50% point, cuts data set in half
- Q_3 :3rd quartile or upper quartile: cuts off lowest 75% of the data (or highest 25%)

Q_1 is the median of the first half of your data set.
Q_3 is the median of the second half of your data set.
The difference between the upper and lower quartiles is called the **interquartile range.** The interquartile range spans 50% of a data set, and eliminates the influence of outliers because the highest and lowest quarters are removed.

Example:
A biologist samples 12 red oak trees in a forest plot and counts the number of caterpillars on each tree.
The following is a list of the number of caterpillars on each tree: 34, 47, 1, 15, 57, 24, 20, 11, 19, 50, 28, 37

Calculate the median, 1st and 3rd quartile.
Step 1: Arrange the values in ascending order
1, 11, 15, 19, 20, 24, 28, 34, 37, 47, 50, 57
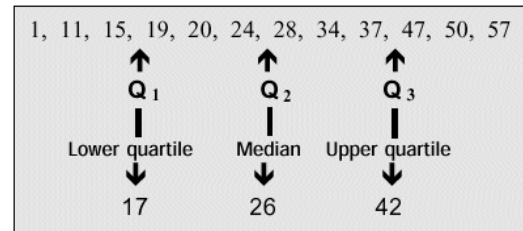
Step 2: Calculate the median
Median = (24 + 28)/ 2 = **26**

Step 3: Determine Q1
Lower quartile = value of middle of first half of data
Q_1 = the median of 1, 11, 15, 19, 20, 24
= (3rd + 4th observations) ÷ 2
= (15 + 19) ÷ 2 = **17**

Step 4: Determine Q3
Upper quartile = value of middle of second half of data Q_3
= the median of 28, 34, 37, 47, 50, 57
= (3rd + 4th observations) ÷ 2
= (37 + 47) ÷ 2 = **42**

Outliers
Observations that are 1.5 x IQR greater than Q3 or less than Q1 are called *outliers* and are distinguished by a different mark, e.g., an asterisk. The actual symbols used don't matter as long as you are consistent and you explain your symbols in the figure legend. In the figure below the arrows are pointing to the outliers.
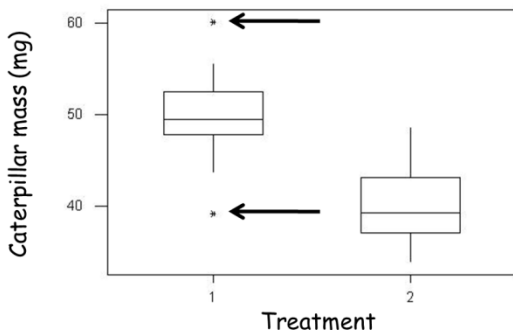
Do not remove outliers from the dataset unless there is good reason to do so. A good reason could be that the outlier is a typo, for example a student records a caterpillar mass of 10 grams instead of 0.10 grams. Or if the equipment used to measure the observation failed, for example if the balance measures a caterpillar as 10 grams. Don't remove outliers just because you want to make your dataset look "prettier". Outliers can point us to interesting patterns or let us know that we may need to increase are sample size.

How do you determine if there are any outliers in your sample?
1. Calculate IQR x 1.5
2. Add this value to Q3. Are there any values greater than Q3 + (IQR x 1.5)? If so, then these values are outliers.
3. Subtract this value from Q1. Are there any values smaller than Q1 – (IQR x 1.5)? If so, then these values are outliers.

Whiskers
The two vertical lines (called *whiskers*) outside the box extend to the smallest and largest observations within 1.5 x IQR (interquartile range) of the quartiles. If there are no outliers, then the whiskers extend to the min and max values.

Comparing populations using *t*-tests

The t-test is a parametric test for comparing two sets of continuous data. It is used for comparing two sample means. For example, we want to know if the average mass of the caterpillars fed leaves from plants receiving the fertilizer treatment is larger than the average mass of caterpillars fed leaves from plants receiving the control treatment. A t-test allows you to determine if there is a statistically significance difference between the two treatments. When you are comparing two samples, then you use a t-test. A t-test doesn't work if you are comparing more than two samples. For example, if you were comparing caterpillars fed leaves from a high, low, and control fertilizer treatments, then you would not be able to use a t-test. The two-sample t-test is a hypothesis test for answering questions about the mean when the data are collected from two random samples of independent observations, each from an underlying normal distribution.

H_1: The means are the two samples (or treatments) are different.
H_0: (null hypothesis): The samples are from the same populations = the means are equal.

A t-test uses 3 pieces of information
1. Sample size = the number of replicates in each sample.

2. The difference between the two means
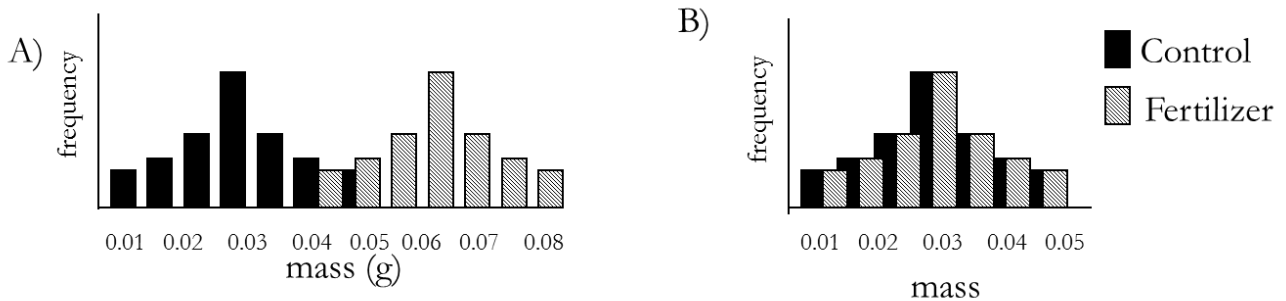
The figures above show frequency histograms of hypothetical caterpillar mass data. The difference between the two means is greater in figure A). These data will produce a more significant t-test.

3. The variance of each sample
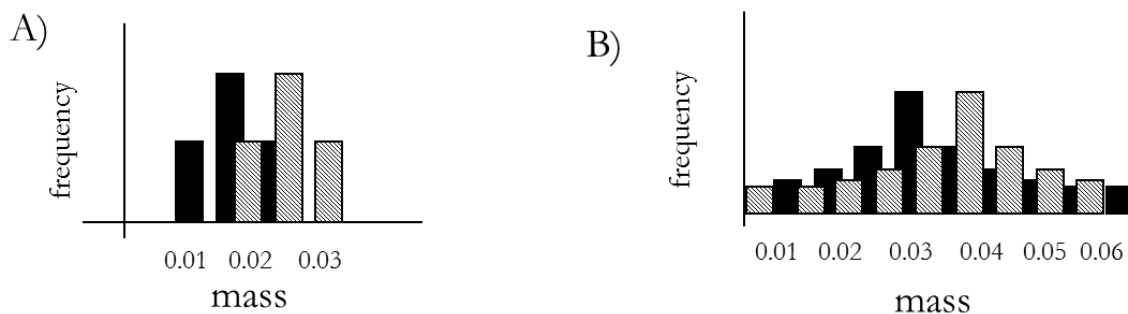
The figures above show frequency histograms of hypothetical caterpillar mass data. The difference between the two means is equal; however, the two samples in A) are more clearly distinguishable because of their smaller variances. The data in sample A) are more likely to produce a significant t-test. Variance is represented by s^2. Variance is a measure of data scattering around the mean. Variance is a way to measure how variable or dispersed the data are. Variance is equal to the standard deviation (standard deviation = s) squared.

Types of t-tests
The type of t-test that you decide to use depends on your hypothesis and methods used to collect your data. This handout provides background information about the t-test and explains how you could calculate it by hand. However, I strongly recommend that you use EXCEL or another statistical software package to do t-tests. The point of doing one by hand is so that you understand what goes into a t-test. If you have to do multiple t-tests, then doing it by hand is tedious.

A **paired t-test** compares two different measures taken from the same individual. For example, I could compare your scores on the ecology pretest with your score on the final to test the hypothesis that you learned something about ecology this semester.

A **t-test for independent samples** compares the averages of two samples. For example, I could compare your scores on the ecology final to the scores from last year's class to test the hypothesis that changes made in the class this year improved student learning. We call this an independent test because the individuals in first sample are not connected to the individuals in the second sample, i.e. the samples are independent of one another.

Additionally, a t-test can also be **one-tailed** or **two-tailed**. This distinction depends on your hypothesis. If your hypothesis is directional, then you would do a one-tailed test. For example "I hypothesize that students in this year's class will score higher on the final than students in last year's class." A one-tailed test completely disregards the possibility of testing for significance in the other direction.

If the difference between the two samples could go in either direction (sample A could be greater than B or sample B could be greater than A) then you would do a two-tailed test. For example "I hypothesize that there will be a difference in exam scores between this year's class and last year's class."

Only use a one-tailed test when you have a very good reason to expect that the difference between the two samples will be in a particular direction. If the difference between the two samples could go in either direction, then do a two-tailed test.

t-test assumptions
- The data in each sample are normally distributed.
- The data in each sample have approximately equal variances. The t-test is fairly robust with regard to this assumption, but if there is a large difference between the variances in each population then you can also do a t-test that assumes unequal variance.

A t-test is fairly robust to deviations from normality; this means that is it usually still OK to do a t-test even when your data do not have a perfect normal distribution. In this class, because of time constraints, we are going to assume that the data are normally distributed without first testing this assumption. However, if you are collecting data for "real" research purposes then you need to test this assumption first.

How to conduct an independent t-test
There are two parts to a t-test: the t-statistic and the p-value. First, you calculate the t statistic and then you determine the p-value that goes along with your t-statistic.

t – statistic:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{(s_1^2 / n_1) + (s_2^2 / n_2)}}$$

In English: The mean of sample 1 minus the mean of sample 2 divided by the square root of the variance of sample 1 divided by the sample size of sample 1 plus the variance of sample 2 divided by the sample size of sample 2.

You also need to calculate the degree of freedom (df) for the t-test:
Degree of freedom = $n_1 + n_2$ - 2

Example
You are researching the impact of urbanization on robin populations. You conduct point-counts in urban and rural settings. From these point-counts you are able to estimate robin density (# of robins per hectare). You conduct 15 point-count surveys in each setting.

The first step is to establish the specific hypotheses we wish to examine.
H_0: Null hypothesis is that the difference between the two groups is 0. **There is no difference in robin density between the urban and rural settings.**
H_1: Alternative or research hypothesis - the difference between the two groups is not 0. **There is a difference in robin density between the urban and rural settings.** This hypothesis would be tested using a 2-tailed t-test. Or you could make a directional prediction – I predict that robin density will be higher in the rural environment. This hypothesis would be tested using a 1-tailed t-test. For this example, we are going to use the 2-tailed hypothesis.

The data

Sample A - Urban	Sample B - Rural
15.6	15.6
19.2	22.8
15.4	23.9
14.9	14.9
25.1	18.4
28.4	19.4
19.5	17.9
29.9	22.1
10.9	20.1
17.9	20.4
17.7	20.7
23.2	20.2
22.6	19.6
28.6	14.6
28.6	21.7

Information you need for a t-test (calculated from the data):
Sample size of a = n_a = 15
Sample size of b = n_b = 15
Mean$_a$ = 21.17
Mean$_b$ = 19.49
Var$_a$ = s_a^2 = 35.64
Var$_b$ = s_b^2 = 7.79

Degree of freedom: $15 + 15 - 2 = 28$

If you plug these numbers into the above t-stat formula, then you will calculate $t = 0.988$. Is this a high or low value?
Now you need to consult a statistical table to determine the significance of the t-statistic. There is a copy of part of a statistical table included at the end of these notes – additionally, if you have taken a statistics class previously and still have your textbook, then there is probably a copy of the table at the end of the book.

1. First, look down the left hand column to find the degrees of freedom that matches the one you calculated. Then, reading across that row, compare your t-value with the number in the second column. **The number in the chart is the critical value of t for p = 0.05**
2. If your calculated t-statistic is smaller than the p = 0.05 t-critical value, then your two samples are NOT significantly different. You must accept the null hypothesis that the two samples came from the same population. There is no difference between the means.
3. If your calculated t-statistic is larger than p = 0.05 t-critical value, then the difference between the means **is statistically significant**. You can reject the null hypothesis with no more than a 5% error rate and accept the alternative hypothesis.
4. The next column in the table presents the t-critical values for p = 0.025. If your value is larger than this number, then you can reject your null with a 2.5% error rate.
5. Then compare the computed p-value to alpha

Step 5: Comparing the computed p-value to alpha

All hypothesis tests are based on the same basic principles and are setup to minimize the probability of drawing an incorrect conclusion.

Two possibilities (reality):
Null hypothesis is true.
Null hypothesis is false.

Two outcomes of a hypothesis test:
We reject the null hypothesis.
We fail to reject the null hypothesis.

Reality → Decision	Null hypothesis is true. Populations are the same	Null hypothesis is false. Populations are different
Null hypothesis is rejected	Incorrect decision	
Null hypothesis is accepted.		Incorrect decision

Two possible types of mistakes:
- Type 1 error: The null hypothesis is true but is rejected. The probability of a type 1 error is denoted by alpha (α).
- Type 2 error: The null hypothesis is false, but is it not rejected. The probability of a type 2 error is denoted by beta (β).

A p-value represents the probability, if H_0 is actually true, that random chance could produce your observed results. If you calculate p = 0.01, then this means that there is a 1% chance of the null hypothesis being true (no difference between the means) given your observed results. If you calculate p = 0.57, then this means that there is a 57% chance of the null hypothesis being true given your observed results.

To determine if you should accept your hypothesis (i.e. reject your null hypothesis), you calculate a t-statistic and p-value.
You accept your hypothesis if $p < \alpha$ (greek letter alpha)

For most scientific studies, the accepted value of α is 0.05

The data support the research hypothesis if $p < 0.05$.
There is a statistically significant difference between the two populations.
The data fail to support the null hypothesis.

The data fail to support the research hypothesis if $p > 0.05$.
There is no statistically significant difference between the two populations.
The data support your null hypothesis.

For a t-test, as the difference between your calculated t-statistic and t-critical **increases**, the p-value **decreases**.

Back to the example:
$t = 0.988$ Is this a high or low value?

By looking at the table, you can see that this value is less than the t-critical value (2.05) and therefore it is **not** significant at p = 0.05. Therefore, we reject our research hypothesis and support our null hypothesis. There is no significant difference in robin density between urban and rural settings.

In this example, $p > 0.05$

But what if I'm using a computer program to analyze my data?
If you are using a computer program to calculate your statistics (which is what we normally do), then you don't need to figure out the t-critical value using a table. Most computer statistical programs, including the statistics you will do in EXCEL, compute the p-value for you. In this example, p = 0.331. This value is much larger than the accepted p-value of 0.05. This large p-value indicates that we cannot reject our null hypothesis (another way of saying this is that we fail to support our research hypothesis).

Reporting statistics in your results section
Report the t-statistic, degrees of freedom and p-value using the following format: Sentence stating your results (t_{df} = t-stat; p-value).

Continuing with the robin example:
- If you used a table to compare your t-statistics and t-critical:
  Robin density did not differ between the urban and rural settings ($t_{28} = 0.988$; $p>0.05$).
- If you used a computer program to analyze your statistics:
  Robin density did not differ between the urban and rural settings ($t_{28} = 0.988$; $p= 0.331$).

Report your statistics, **don't discuss p-values and t-statistics.**
This is wrong: My p-value was 0.046, this is less than the accepted value of 0.05, so I can reject my null hypothesis and accept my alternative hypothesis.

In the results section you just state the results (including the statistics). In the discussion section, you interpret and explain your results.

Don't ignore your statistical results.
In this example, average robin density was higher in the urban setting (21.19 robins per count) than in the rural setting (19.49 robins per count). However, these means are NOT significantly different. So you cannot make any statements about differences in robin density between the two environments. It is **incorrect** to say that robin density was higher in the urban setting.

Practice problems with answers
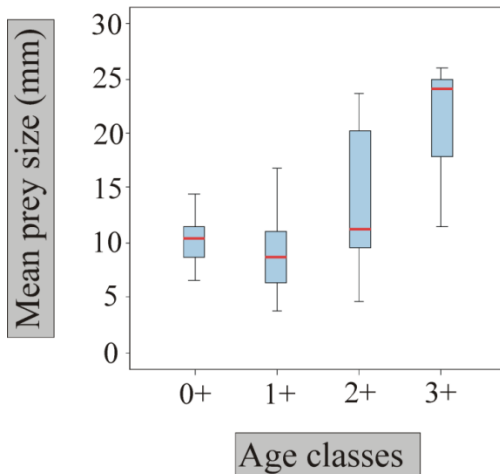Problem 1 – Brown trout box plots

Figure 1: Brown trout eat a variety of freshwater invertebrates and the size of the food that they eat can vary with trout age. This figures shows box plots of the age-related variation in prey size of *Salmo trutta*, brown trout, in the Furelos River (NW Spain) during summer. From: Ontogenetic Dietary Shifts in a Predatory Freshwater Fish Species: The Brown Trout as an Example of a Dynamic Fish Species By Javier Sánchez-Hernández, María J. Servia, Rufino Vieira-Lanero and Fernando Cobo

a) What is the independent variable? What is the dependent variable?

b) Describe how prey size changes as trout age class increases.

c) Visually estimate the minimum, maximum, median, Q1, Q3 and the IQR for teach treatment.

Age class	Min.	Max.	Median	Q1	Q3	IQR
0+						
1+						
2+						
3+						

d) Which of the box plots would most likely have normal? Explain your answer.

2. Creating and interpreting box plots
   a) Use the data below to construct box plots of the two samples. Round to whole numbers.

Sample A	Sample B
11	14
38	16
15	17
35	20
11	11
8	13
32	17
10	16
16	22
16	15
27	13
18	13
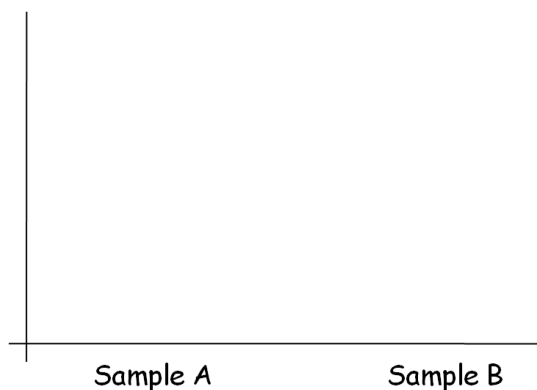
b) Do you think that samples A and B were collected from the same population? Why or why not?

c) Which sample would you predict would have a greater standard deviation? Explain your answer.

3. A researcher hypothesized that phosphorus-fertilization of host plants will impact caterpillar growth. 8 caterpillars were reared on plants fertilized with phosphorus and 8 caterpillars are reared on plants that receive a water only (no phosphorus) treatment. The data are presented below.

Pupal Mass - Phosphorus	Pupal Mass - No phosphorus
12	8
7	7
3	4
11	14
8	6
5	7
14	12
7	5

Sample Mean		
Sample Size		
Variance	13.70	11.84

a) What would be the null hypothesis in this study?

b) What would be the alternate (research) hypothesis?

c) What type of t-test would you use?

d) How many degrees of freedom for your t-test?

e) What is your t-critical?

f) What is your computed t-statistic?

g) Is there a significant difference between the two groups?

h) Write these results up as you would in the results section of a lab report (include a statement of the results and the statistics).

i) Interpret these results as you would in the discussion section of your lab report (i.e. what do these results mean).

Answers
1. Brown trout

a) What is the independent variable? What is the dependent variable?
Independent variable: Brown trout age class
Dependent variable: Prey size
b) Describe how prey size changes as trout age class increases.
   Older fish eat larger prey.
c) Visually estimate the minimum, maximum, median, Q1, Q3 and the IQR for teach treatment. (Since these are estimates, it's OK if your values are slightly different).

Age class	Min.	Max.	Median	Q1	Q3	IQR
0+	*6*	*15*	*11*	*9*	*12*	*3*
1+	*4*	*17*	*9*	*6*	*11*	*5*
2+	*5*	*24*	*12*	*10*	*20*	*10*
3+	*12*	*26*	*24*	*18*	*25*	*8*

d) Which of the box plots would most likely have normal distributions (you may want to review the pine needle handout)? Explain your answer.
   The box plots for age classes 0+ and 1+ most likely were produced from data that are normally distributed. When data are normally distributed (bell curve) then the mean = median and there are approximately an equal number of replicates above and below the median (the median looks like it is in the middle of the distribution of data).

2. Use the data below to construct box plots of the two samples. Round to whole numbers.

Sample A	Sample B
11	14
38	16
15	17
35	20
11	11
8	13
32	17
10	16
16	22
16	15
27	13
18	13
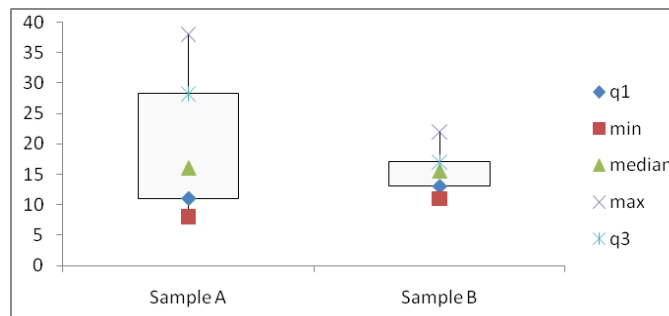
a) Do you think that samples A and B were collected from the same population? Why or why not?
   Answers may vary – a correct answer depends on a correct explanation. The two samples have the same median but the range of values is greater in sample A. There is much more variation within the sample A population and the mean is greater.

b) Which sample would you predict would have a greater standard deviation?
   Sample A (greater range of data). If you actually calculated the values:

	Sample A	Sample B
average	20	16
stdev	10.5	3.1

Problem 2: A researcher hypothesized that phosphorus-fertilization of host plants will impact caterpillar growth. 8 caterpillars were reared on plants fertilized with phosphorus and 8 caterpillars are reared on plants that receive a water only (no phosphorus) treatment. The data are presented below.

Pupal Mass - Phosphorus	Pupal Mass - No phosphorus
12	8
7	7
3	4
11	14
8	6
5	7
14	12
7	5

Sample Mean	8.375	7.875
Sample Size	8	8
Variance	13.70	11.84

a) What would be the null hypothesis in this study?
There will not be a difference in mass between the caterpillars fed phosphorus-treated leaves and the caterpillars fed control leaves.

b) What would be the alternate (aka research) hypothesis?
There will be a difference in mass between the caterpillars fed phosphorus-treated leaves and the caterpillars fed control leaves.

c) What type of t-test would you use?
A 2-tailed t-test. In the description above, the researcher hypothesizes that phosphorus will impact caterpillar mass, so there could be a positive or negative effect of phosphorus on caterpillar mass.

d) How many degrees of freedom for your t-test? *14*

e) What is your t-critical? $t_{critical} = 2.145$

f) What is your computed t-statistic? $t_{stat} = 0.279$

g) Is there a significant difference between the two groups? *No*

h) Write these results up as you would in the results section of a lab report (include a statement of the results and the statistics).
Pupal mass of caterpillars fed phosphorus treated leaves was not significantly different than pupal mass of caterpillars fed control leaves ($t_{14} = 0.279$, $p > 0.05$).

i) Interpret these results as you would in the discussion section of your lab report (i.e. what does these results mean).
Phosphorus had no effect on caterpillar growth. Perhaps phosphorus is not a limiting nutrient for caterpillars, so adding more phosphorus to the plants did not increase caterpillar mass. Other explanations are possible. However, it is important that you don't discuss you t-stat or p-value in the discussion.

Critical values for the t-distribution

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