**Bio183 Lab Summer 2023 Name \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**15 points – Due at the beginning of lab during the week of June 6.**

**Short introduction to statistical analysis (Student’s t-test)**

The purpose of this document is to start familiarizing you with the importance of statistics in data analysis.

Since the beginning of this semester, you have been designing experiments in a scientific manner. As you already know, when scientists gather data, they are always confronted with the task of comparing experimental groups between them, or with a control group.

The following table represents data that were collected by 20 groups of students who compared the respiration rate of mung beans with that of bean beetles. For this experiment, students placed equal weights of bean beetles and mung beans in 2 different respiration chambers similar to the ones you used in lab, and allowed the organisms to respire for 10 minutes. At the end of the 10 minutes, the CO2 concentration in the chamber was recorded (in parts per million) and entered in the table below (i.e. the table contains 20 replicates of the same experiment).

|  |  |  |
| --- | --- | --- |
|  | **Mung Beans** | **Bean Beetles** |
| **1** | 10.40817972 | 11.45594902 |
| **2** | 9.034635194 | 13.73553831 |
| **3** | 8.81916301 | 12.3095564 |
| **4** | 8.442095534 | 11.59162107 |
| **5** | 8.776707016 | 12.0402397 |
| **6** | 8.345589197 | 12.47250484 |
| **7** | 9.858450408 | 11.6921308 |
| **8** | 10.29463473 | 12.93743748 |
| **9** | 10.79704177 | 13.22190712 |
| **10** | 8.904041254 | 11.72468125 |
| **11** | 10.0829407 | 12.01128412 |
| **12** | 10.39230658 | 13.08776349 |
| **13** | 8.473160125 | 11.70709261 |
| **14** | 10.05532488 | 11.60232903 |
| **15** | 9.573302875 | 11.81669814 |
| **16** | 9.713062038 | 11.47132447 |
| **17** | 12.09911265 | 11.47220908 |
| **18** | 9.640241403 | 12.53026095 |
| **19** | 10.52857026 | 11.51297888 |
| **20** | 10.82086222 | 11.56236617 |

This table is way too messy to allow you to draw any kind of conclusion by quickly looking at it.

A comparison between the mean values of these two columns would help us simplify the data.



**1. What can be determined from this? Can we safely say that these 2 groups are different? Why/why not?**

If instead of representing that data with a bar graph, we represented it with a scatter plot, would we be able to determine any additional important information?

Let’s give it a try:



1. **What is the advantage of plotting data this way? What can we determine from this data that we could not determine from bar graph above?**

Additionally, if we consider the following 2 sets of data, we can see that the column on the right contains data that mostly matches the column on the left, with the exception of 1 data point. This isolated point is most likely a hiccup (although not necessarily), but it is sufficient to affect the whole mean value. By plotting our data in a bar graph, we wouldn’t have been able to acknowledge the presence of this isolated data point. After careful assessment of our data collection method, and if we have evidence that this particular data point is highly unlikely to represent the population as a whole (e.g. value that is physically impossible or mis-recorded, experimental subject mishandled), we may want to exclude this unusual data point from our analysis, as its presence appears highly unusual, and may affect our analysis negatively. This unusual data point is known as an outlier. We cannot overemphasize the importance of not rejecting what we consider an outlier without thorough investigation why this data point looks different than the rest.



In many scientific journals, you will also see the following type of graph. This type of bar graph represents the mean value as the “tower,” and the “antenna” at the top represents the variability of the data. Although not necessarily the best way of displaying data, this type of graph is probably the most common type of graph used in biological journals.



**Quick summary:** We have now established that when trying to compare the differences between 2 sets of data, one must pay attention to 2 things: difference between the means, and variability within each group.

Statisticians realized this a long time ago, and luckily for us, they came up with an easy and convenient way of quickly assessing the likelihood that 2 sets of data are different or not from one another.

One of the simplest statistical methods used to assess the differences between 2 groups is known as a Student’s t-test. “Student” was the pseudonym of the statistician who introduced this test in the early 1900s – it has nothing to do with an actual student!

The t-test allows us to calculate what is known as a p-value, which is a measure of the confidence with which we can assert that a group of data points is different from another group. When performing a t-test, we have to start by making the assumption that the 2 groups of data are not different from each other (a concept known as the null hypothesis). By convention, biologists usually agree that a p-value of less than 0.05 indicates that we are 95% confident that we can reject the null hypothesis (i.e. we are 95% confident that there is a statistical different between the two groups analyzed). It should be noted that we are somewhat oversimplifying some statistical concepts, here.

Here are suggestions on how to express the results of a t-test:

**p-value < 0.05:** Based on the data we collected, we can reject the null hypothesis. Therefore, we have statistical evidence that on average there is a significant difference between Groups A and B at the 95% confidence level.

**p-value >= 0.05:** Based on the data collected, we cannot reject the null hypothesis. This means, we did not collect sufficient evidence to conclude that there is a significant difference between the group means. However, this does not mean that we can conclude the means are equivalent. There is another type of test (called equivalence test) that would be required if you are trying to prove that two groups means are the same (or equivalent).

**Note:** It is important to mention that before performing a Student’s t-test, one must insure that

1. the data are normally distributed (i.e. are distributed in a pattern that fits under a bell-shaped curve)
2. The 2 groups being compared show equal variance (i.e. same distribution pattern – see scatter plot above)
3. The data used to carry out the tests are sampled independently from the two populations being compared (i.e. we are not measuring the same subject multiple times)

 For the purpose of this assignment, we will perform our analysis in Microsoft Excel. However, it is worth mentioning that Microsoft Excel is far from being the software of choice for statistical analysis! We selected Excel for its ubiquity and ease of use, but programs such a Jump, SAS, R, Graphpad Prism (just to name a few common ones) are much more capable and offer a much wider selection of statistical analyses than Excel does.

We would like you to perform a statistical Student’s t-tests to assess whether the rate of respiration of pinto beans is different from that of black eyed peas (some of you may have chosen to do this comparison this week).

Last year, one lab compared the rate of respiration per gram of pinto beans and black eyed peas, and each group of students (8 total) recorded their values in the table below:

|  |  |  |
| --- | --- | --- |
|  | **Pinto Beans** | **Black Eyed Peas** |
| **Group 1** | 0.107 | 0.310 |
| **Group 2** | 0.177 | 0.189 |
| **Group 3** | 0.157 | 0.220 |
| **Group 4** | 0.043 | 0.233 |
| **Group 5** | 0.059 | -0.010 |
| **Group 6** | 0.118 | 0.216 |
| **Group 7** | 0.181 | 0.168 |
| **Group 8** | 0.186 | 0.233 |

Let’s assume that this data is normally distributed, that the 2 groups show equal variance, and that students were not measuring the same beans/peas multiple times.

For your assignment, start by entering the two columns of data above in Microsoft excel

* Select an empty cell in which you want to display your p-value
* In this cell, write the following: **=ttest(*column A*, *column B*, 2, 2)**

 To enter column 1, just highlight your first column with the mouse, and repeat the same thing for column 2.

The two “2” values entered after “column B” are variables whose importance you will learn later in your statistics courses.

**3. What p-value did you obtain? What does it mean?**

1. **Now, look at your raw data. Is there anything in there that looks strange or abnormal? What is it, and how do you think that this data entry ended in this table?**
2. **Now, let’s re-do the t-test calculation while omitting the outlier. What p-value did you obtain now? What does it mean?**

Once again, we cannot overemphasize the importance of carefully analyzing whether to keep or reject a certain data point in your analysis. Data selection for the sole purpose of supporting a pre-determined outcome would be considered scientific fraud! However, data points that clearly fall outside of a “normal” range values (such as a human height recorded as 16 feet 7 inches, or an outside temperature of 236 degrees Fahrenheit) can be considered outliers and omitted in your analysis. In some cases such as this one, removing an outlier can make a significant difference in the outcome of your analysis.

**\*\*\*\* This is good practice for you as you prepare your bean beetle report. You will be required to do a simple Student’s t-test on your data for inclusion in your oral and written reports.**