Biol 171 Methods in Field Biol. F'2023 Name _____

Lab 11 - Working with data: Excel T-test and introduction to R

In this lab we will expand on previous work in excel. We will start with review to practice and then use a t-test to test for a difference between two groups. Work through the following exercises in excel and R. **Please turn in 1) Excel analysis file with graphs and functions; 2) typed up answers to questions in bold in a Word document (a single file for excel and R graphs/answers is fine), organized by section and subsection; 3) text file with your functioning R code. If you do not have enough time to get this done in lab and turn it in today, work on it outside of class and let me know right away if there are questions. I will set the deadline for your analysis and answers to be turned in to Canvas as Wednesday Oct 25th by 11:59pm. There is an associated peer review due on the interpretations Oct 26th by 11:59pm.**

Whenever you work with data you need to get to know it. What are the central tendencies and variation like? What are the extreme values? Are there any obviously wrong data points? Missing data points? Other errors? Great ways to do this are **1) graphical methods and 2) summary statistics**.

In this exercise we will examine and plot data on body mass (in grams) in a butterfly species from a population in the Bay Area.

Download the dataset from Canvas: "bmassdata171.txt" and open it with Excel. When opening the file make sure you use "Tab" as a column delimiter because this is a tab delimited text file. Remember that text files with columns delimited by tab, space, or comma, are a common format for saving data files because they can be read and imported into many analysis programs.

Examine the data. Sort the data by "sex" using Data→Sort (select the column letter to sort by, or check the "my list has headers" and select the name of the column).

PART I - Excel Basics (review):

Data range

For many of the functions in Excel you will probably want to use, like summing, averaging, finding the maximum, you need to reference a range of cells (=data range). To do this you can click and drag to select the cells you want to include or you can type in the range. For example typing or selecting "G1:G23" means you want cells in Column G, from Row 1 to 23, and typing the function =sum(G1:G23) will return the sum of all the data in that range of cells.

Copying functions across multiple cells

In some cases you may want to repeatedly use a function on a certain column, or row, or even single cell. You can do this by putting "$" in front of the column and row to use data from just a single cell (or in front of the column to freeze the column, or in front of row to freeze the row). For example, maybe you want to use the result of a calculation in cell $C$3 as a value in a second calculation. By using the $ you insure that even if you drag the function, or paste it down 100 cells the cell C3 is still being used every time.

Try this: say we want to add a constant value to our body mass data in column G. We could do this by typing "3" into J1 in the data set and then in J2 type "=G2+J1". You should get 3.302 in J2 if you sorted by sex descending the alphabet. To do the calculation for all the individuals you just Fill → down or grab the bottom right corner and drag down to where you want to end. But it does not work correctly the way we currently have it set up. Having a single cell for the value you want to add is a nice way to set this up if the value in J1 might change – your body mass + J1 is automatically updated. But if you drag or copy/paste the function you just made in J2 down into J3 the calculation is incorrect (3 + 0.274 ≠ 3.576). The program used data from J2 instead of J1! Confirm this by double clicking in J3 to see the cells involved in the function highlighted in color – it used J2 instead of J1 right? The program assumes when

you drag or copy/paste that you want to use the data in the same RELATIVE position as the original function. Now edit the function in J2 (in the formula bar) to be "=G2+$J$1". Next drag or copy/paste and you'll see that J1 is used in the calculation in each cell no matter what row the function is in and J3 = 3.274 as it should be.

Counting items in rows or columns

In many cases you need to count large number of items and will want to use the functions:

=count(data range) ← for numbers

=counta(data range) ← for text

=countif(data range, condition) ← for conditional counting, the condition is a logical test such as a cell value equal to a certain numeric/text value, like in the following example.

Instead of just directly counting the number of females, use the function: =countifs(data range 1, "F", data range 2, ">0"). This function counts all Females ("F") that have a non-zero value (">0") for the variable in range 2. In this function you should select for "data range 1" all individuals in the sex column and for "data range 2" all individuals in the body mass column. Note that for finding text, the condition (Female) is in quotes "". This is not necessary for numbers. You can also use > , < , = in your conditions.

In a new cell alter the code for counting the number of males.

I.1 What is the sample size for females and males with body mass data?

Now use the =min() and =max() function to find the high and low values for each sex. Type into the cell =min(data range for males) to find the minimum value of body mass for males.

I.2 What is the range of values for females (min and max)?
For males?

I.3 Does this range seem reasonable after visually scanning the data? Does anything seem odd?

It is very important to always visually inspect your data and a great way to do this is with graphical methods such as a box plot or scatterplot. In this case we'll make a quick and dirty scatter plot for each sex on the same graph.

I.4 Make a scatter plot of the data for both males and females on the same plot. Sex should be on the x-axis as a grouping variable with body mass plotted on the y-axis. You'll have to add a column of "dummy" x-data for the groups, for example a column for "sex" that =1 for males and =2 for females. The "NA's" in our data set will come out as zeros and so it is probably best to delete them from the dataset now. Make a scatterplot by inserting a chart and then in the Chart menu select the source Data option and add a series for males and a series for females where you select the x and y data for each sex. You can remove the numbers on the x axis and make a horizontal axis title that says "Male Female".

I.5 Do any of the measurements seem like outliers or errors? If so which ones?

This commonly happens and is why it is a good idea to enter your numerical data twice and calculate the difference to identify errors. In this case if you were to check the raw data in the notebook you would find there was a transcription error and that individual SCC-1 accidently had a zero after the decimal making it 0.0223 instead of 0.223. Remove this body mass datum from the dataset before continuing your calculations. Body mass was never recorded for SCC-1, it is bogus data added for this activity.

Now use the =average(data range) to calculate means for males and females. Make sure to delete the body mass datum for individual SCC-1.

I.6 What is the average value for males?
What is the average value for females?

Report with correct significant digits – see "Assignment writing guidelines" if unsure.

I.7 Do the male and female distributions overlap?
Do they look like different distributions?

This can be answered with statistics, specifically a t-test. We will do this next.

PART II - Excel T-test

We will test the null hypothesis that body mass does not differ between the sexes in this population. Putting this in terms of distributions of data, we are testing the hypothesis that males and females are from the same distribution. If we can reject this null hypothesis then we accept that the male and female distributions differ, and conclude there is sexual dimorphism for this trait.

We will use a t-test to test our hypothesis. A t-test assumes several things about the distributions being compared. Perhaps most importantly it assumes that each data point is independent of the others. One way this could be violated is if we measure individuals more than once and include them in the dataset as independent measurements. Another assumption is that we have normally distributed data. What this means is that the distributions for both populations (male and female) are approximately bell shaped with a central tendency and equal tails on both sides. There are ways to quantitatively describe this that are beyond our scope here but we can do it by eye.

Frequency histogram

These data are perfect for making frequency histograms to inspect the distribution of our data, one for each sex.

First, you have to put the data into appropriate sized bins - here is a trick to do it:
label column M "Female body mass intervals (g)"; and column N "number Female individuals"; column O "Male body mass intervals (g)" and column P "number Male individuals"
Create 0.025 g wide intervals starting with a bin that includes the lowest value in the entire dataset: a males has FW=0.148 g, so make the first bin 0.125 to 0.1499 so 0.148 falls within.
Use the formula "=COUNTIF($G$2:$G$17,">="&M2)-COUNTIF($G$2:$G$17,">="&M3)" to count the number of individuals in the first bin (0.125 to 0.1499). This formula counts all individuals greater than or equal to a certain value (e.g. 0.125) and subtracts out all that are in the next higher bin. Make sure the equation make sense to you. Copy and paste the formula into the next cell down for all the other body size bins. Remember the "double $" allows you to paste the formula and always select the same exact range of cells (i.e. it nullifies using the relative position of cells in the equation).
Once you paste down for the females, set up the equation to work with the male data.
***Note: 1) you will have to change the data range to do this for the males, and 2) your selected cells may be different depending on how you set up your spread sheet. Carefully check that you are selecting the correct cells for females and males.

Use the graphing functions in the Insert Chart menu (select first vertical column icon); select Source data from the Chart menu. Then for Series Values select the frequency in each bin (this is the y-axis). Then we need to select the body mass 'bins' for x-axis. This is straightforward on a Mac just enter the x-axis category bin values. Do this by clicking the series name (male/female) and selecting EDIT under horizontal axis labels and you then select the range of cells containing your bins.

Delete the horizontal lines to make a nice clean plot. Keep legend info to show male vs female. Other legend information describing the data goes in a figure legend below the figure.
Add axis labels and a figure legend to your plot. Delete the plot title.

Label your axes and provide a short figure legend below your graph.

II.1 Do our distributions look approximately normally distributed?

A final assumption we will mention here is that our data have **equal variances** (**homoscedastic**; **heteroscedastic** would be **unequal variances**). Again there are ways to quantitatively assess homoscedasticity, namely by calculating the variance.

II.2 Calculate and report the variance (the square of the standard deviation) for our samples using =var.s(data range of female data); do this for both sexes.

II.3 Based on the variance you calculated, do our data appear to have equal variances?

II.4 Which sex has higher variance?

II.5 Does this make sense given the distributions you plotted above?

Calculate the *p*-value for the null hypothesis that the males and females come from the same distribution.

- In excel use the function "=T.TEST(array 1, array 2, tails, type)" where you select the range of male data for array 1, and select the range for female data for array 2.

- Do a one-tailed test by entering 1 for 'tails'. A one-tailed test is used when you want to test for a difference in the distributions in a specific direction instead of either direction (which would be a two-tailed test). For example here, we expect females are larger (have a larger mean) than males so a one tailed test is appropriate.

- "type" is the type of test. You can do 'paired t-tests' if you measure the same individual before and after a drug treatment. We did not do this so we choose one of the two-sample options. Based on our calculations of variance above we should enter 3 for type because our variances do not seem equal.

- The resulting p-value indicates the probability of getting a difference as large as, or larger than, what we see in our dataset if the null hypothesis is correct. Even if the males and females are from the same distribution and the null hypothesis is true, there will be some small percentage of studies that find a difference as a result of sampling a subset of the population. We use the standard scientific cutoff of $p<0.05$ for our test.
    - In this test if $p<0.05$ then we reject the null hypothesis that male and female samples are from the same population (reject our null hypothesis) because there is a less than 5% chance that the two samples came from the same distribution, and a significantly low probability that the samples are different just by chance alone. In other words, we feel confident that we are seeing a 'real' or 'significant' difference and not just a chance event

and we accept the alternate hypothesis: male and female samples are from different populations.

- Finally, calculate the value of the *t* statistic in excel by using "=T.INV(probability, degrees of freedom)" and select your *p*-value (select the cell in which you calculated the p-value above). For this test we will enter 25 degrees of freedom, two less than the total number of independent samples (which here is 27).

II.6 Write out a statement summarizing the result of the test that includes: a null hypothesis, a report of the p-value and t-statistic in complete sentences.

Do we reject the null hypothesis?

What do you conclude about sexual dimorphism of body mass in the population?

PART III – Introduction to R

For this section remember to 1) save and turn in electronically a TextWrangler/TextEdit/WordPad document with the R-code that you get to work (that way you can adapt it and use it for other analyses later), and 2) add your answers to the numbered questions below to your Word file, along with your graphs, organized with the Letters and Numbers. Turn them in on Canvas by Monday Oct 26th by 11:59pm.

***Each time we use R I want you to create a text document (textedit, bbedit, textwrangler, or other, NOT Word b/c of problems with quotes) to write your good working code into as you work through this activity. Save the code that works for each lab!!! This way you can use that code to do the same or similar kind of analyses later.**

What is R you ask?
R is a statistical programming language that is a free open source version of S (a common stats programming language) with HUGE number of people out there constantly tweaking, modifying, writing new statistical methods, and answering questions about it. Go to http://www.r-project.org to begin to get a sense of it. There are wonderful mailing lists with searchable archives, and people always willing to give you pointers.

Not only does R have all of the basic statistical tools at your easy disposal, but it also contains library upon library of obscure and powerful statistical methods that you might need at a moments notice. That, and the code, once you wrap your head around the basic syntax, is very simple. You can do an ANOVA on your dataset in three lines.

R is largely object oriented. This means that, instead of saying "run my ANOVA!" and getting results printed up, you create an ANOVA object with your data. The object contains every bit of information about the ANOVA you would ever want. To see this information, or work with it later, you need to call a function on the ANOVA object itself, for example:

```
my.anova<-anova(response ~ treatment)
summary(anova)
```

For future reference, text like that is R code.

Let's try some…

Getting R:
Go to http://www.r-project.org and click on the top left CRAN (Comprehensive R Archive Network) link. Then scroll down to the link for Berkeley and then select your operating system. On a Mac. click on the most recent .pkg link that is compatible with your computer; on a PC click on the "install for first time" link and then find a link to download the package that best fits your system.

Once it is installed, open the program R and set the working directory.
If you want to do RStudio you can download that here (https://rstudio.com/products/rstudio/download/), but you will also need to have R installed.

Setting the working directory: **setwd()**
You need to tell R what directory your data files are in so you can work with them.
Let's start by setting up a new "DataAnalysis" folder located within your "Documents" folder.

Do this now.
In the future you can have folders for different datasets or different labs.

You can use **getwd()** to find out what the current work directory is and use this to help format your code in setwd(). Try it.
When I first open R and use the getwd() command it returns ("/Users/rhill4").
To set a new work directory I then type:

```
setwd("/Users/rhill4/Documents/DataAnalysis")
```

One trick at finding the exact path of a file or folder is to drag a file or folder into the workspace in R. It should show you the path leading to that file/folder and you can use this information to set the working directory. You can also probably see the path below an open window if you have selected that option.

Another useful command is **dir()**. Typing dir() will return a list of items in the active directory.

Save the code that works in a text file! The R workspace can be saved, but it will be a good idea to keep a text file (you need to turn this in after lab) with the code that works for different analyses so you can come back to it easily. Any text program will work including the simpler text editors that come on your PC (avoid using Word for this!). For Macs I recommend BBEdit https://www.barebones.com/products/bbedit/ or Textwrangler (http://www.barebones.com/products/textwrangler/download.html).
R also has it own editor and you can start a new file of code by clicking on the blank page to the left of the printer at the top of the R workspace.

Getting Help Inside of R:
You can get help on any specific function while running R by typing help(function), ?function, or by asking for an example with example(function).
If you have questions beyond this, google and the R website will most likely have your answers. The manual is at http://cran.r-project.org/manuals.html and you can find great other tutorials at http://cran.r-project.org/other-docs.html. There are also a few other great resources out there:
Searchable archives of the R mailing list http://tolstoy.newcastle.edu.au/~rking/R/
RTips http://www.ku.edu/~pauljohn/R/Rtips.html
R Graph Gallery http://addictedtor.free.fr/graphiques/index.php

Bringing data into R:
Make a duplicate of the dataset you downloaded called "bmassdata171.txt" and move the duplicate to your "DataAnalysis" folder or wherever you think is convenient. Make sure you remove the "bogus" SCC-1 body mass data and also remove the "copy" from the filename. Once you do this and it is in its new directory open the file by using code like this:

```
massdata=read.delim("bmassdata171.txt", header=T, sep="\t", as.is=T)

#NOTE that this is a tab delimited file and so the sep is "\t" instead of "" for "space"
#NOTE – remember that in R, the # is a comment delimiter used to leave a note in the code to remind yourself what the function did or why you did something a certain way.
```

What I have done here is to tell R that "massdata" is and object containing the data in the "bmassdata171.txt" file. Note you can use the = or <- to name objects.

I usually store my data as a text file because it uses less memory than an Rdata file, is safe and because it's easier to then open it in Excel or other programs if needed. Use **read.table() or read.delim()** to read text files into R**. read.delim()** is a version of **read.table()** that is designed for tab-delimited data that has a header row at the top. If your data is different you can import just about anything by changing arguments in **read.table()**. The arguments regularly changed are:

file="file location and name"

sep="\t" indicates how your data is delimited. ("" means space, "\t" means tab, "," is for comma-delimited, etc.)

as.is=T (the default for this argument is F, and I have found that it causes me a lot of headaches. When this is set to true it reads in character data as characters, integer data as integers, etc. Otherwise, it may import some fields as *factor* data when you may not want that.

header=T if your data has a header row at the top. If you do have that, but don't tell R, it will find that first row of data, then think all of your columns are character data, and numeric or other data in the table will show up as NA.

na.strings=c("dead","-99") tells R to import certain characters or codes as NA.

Now you should have a feel for how R treat things like objects. Type "massdata" and R will return a view of the whole dataset. Compare what you imported, shown now in R, with what is in the text file as shown in a text editor. Notice how R inserted periods between some text because there were problems with the column delimiters like "tab", "period", "space", "comma". This is common in working with data and something you have to examine carefully each time you start a new project. This is why I encourage you to put underscores ("_") to represent spaces in data columns.

III.A. Histograms

These data are obviously the same as the ones we worked with in Excel so the point here is to learn another way to analyze data. We did a histogram in Excel and it took some careful coding, but look how easy it is in R.

First we make objects for just the male and female data to plot.

```
malebmass<-subset(massdata, sex=="M")
femalebmass<-subset(massdata, sex=="F")
```

Now look at those objects to confirm you have objects for each sex separately.

Then use the hist() function, setting the x axis limits (xlim=c()) to make the graphs comparable.

```
hist(malebmass$b.mass.g, xlim=c(0.125,0.35))
    #using $ selects that column in the object indicated
quartz() or windows() #this makes a second blank plot for the female graph, quartz() on Mac, windows() on PC.
hist(femalebmass$b.mass.g, xlim=c(0.125,0.35))
```

#Note you can make nice axis labels using xlab= (or ylab=) as subcommands in the hist() function. Here for example: hist(femalebmass$b.mass.g, xlim=c(0.125,0.35), xlab= "Female body mass (g)").

Try this for females and males.

If you'd like to eliminate the gaps between axes, try adding the argument

, xaxt = "n" to make the plot without an x-axis, then add the x-axis on a new line with axis(1, pos = 0)

Paste these new graphs into your word document (select all =command-A and then paste) **under a section labeled "III.A"**.

Once pasted into Word examine the overlap between males and females and compare the plots with what you made in Excel.

III.1. Do males and females look like they overlap? Are they significantly different? Record your answers on the exercise sheet to turn in. Does your R histogram agree with the one you made in Excel?

III.B. Boxplots

Box and whiskers plots are very commonly used to visualize variation between groups. The middle line in the box indicates the median of the data and the box edges indicate the 25th and 75th percentile of the data (25% of the data are below the 25th percentile and 25% are above the 75th percentile. The end lines represent the extremes. Try thinking about this like you are looking down on a histogram or Gaussian distribution. Both approaches describe the shape of the distribution.

Paste your boxplot into your word document under a section labeled "III.B"

III.2. Explain how the shape of the box plots relates to the shape of the histograms.

```
boxplot(massdata$b.mass.g ~ massdata$sex, xlab="Sex", ylab="Body mass (g)")
#Again, note the use of xlab= and ylab=. Paste the graph into your word document.
```

III.C. Measures of Variation: Standard Deviation, Variance

Variance is a measure of variability in the data. It is the average of the squared deviations from the mean. Variance is calculated by getting the sum of squared deviations from the mean and dividing by the degrees of freedom N-1. Variance is the square of standard deviation (s^2). Commonly used in statistical formulae, it is not as practical for working with the data b/c it is in units squared.

Calculate the variance for male body mass:

```
var(malebmass$b.mass.g, na.rm=T) #here the na.rm=T indicates that the NA's should be removed – the calculation won't work if the NA's are included.
        #var() and sd() both use denominator n-1 for samples
```

Standard Deviation – the standard deviation (s) is a measure of variability derived from the variance. Standard deviation is expressed in the same units as the data and so is commonly used to describe datasets.

Calculate the standard deviation for male body mass:

```
sd(malebmass$b.mass.g, na.rm=T)
```

III.3. Report the standard deviations for each sex. Explain how the shape of the plots fits with the values for standard deviation in each of the sexes.

III.D. T-test of differences between TWO groups

To test whether males and females are different we will use a t-test.

A t-test tests the null hypothesis that two groups come from the same distribution.

```
t.test(massdata$b.mass.g ~massdata$sex, paired=FALSE, alternative=c("two.sided"))
        #here we have indicated a statistical model with the tilde symbol (~) that says do a t-test on body mass by sex. It is essentially the dependent continuous variable by the independent grouping variable.
```

III.4. What is our p-value and t-value for the test of the null hypothesis that male and female come from the same distribution? Write out the results in complete sentences.

III.5. Write out what you conclude about the null hypothesis based on this test.