CONDUCTING STATISTICAL ANALYSES IN MICROSOFT EXCEL - OBE

For statistical analysis in Excel, you will need a Data Analysis add-on.

a. In order to run statistics in Excel, you will need the ***Data Analysis Tool Pak*.**
    i. In Excel, go to File>Options>Add-ins
    ii. Select "Analysis Tool Pak" from the list. Click "Go"
    iii. A new small window will pop up. Make sure that "Analysis Tool Pak" is checked. Click "OK"
    iv. You should now see a "Data Analysis" option under "Data" or "Tools"
    v. Follow the directions below to use the Data Analysis Tool Pak within Excel.

Descriptive Statistics: With any data set, it is wise to examine the mean and distribution of the data. Means, standard deviation (SD), and standard error (SE) are good ways to get a quick sense for the spread in your data. Descriptive statistics can also tell you about the outliers in your dataset. Why is it important to notice outliers?

a. ***Use Excel to calculate means:***
    i. "=average()". For example, "=average($C$3:$C$12)" would give you the average of the numbers in column C and rows 3 through 12. You can manually enter the data range, or you can select the data by clicking and dragging.

b. ***Use Excel to calculate SE or SD***: both SE and SD are descriptive statistics and tell you about the variability in your data set.
    i. SD characterizes the dispersion of the data points about the mean. Like SD, SE is a measure of variability. However, SD measures variability within your one sample while SE measures variability of an estimate (for example, the mean) among different samples of the same population. SE essentially gives the expected error of an estimate due to sampling. The mean +/- the SD shows you the range most of your data fall within, whereas the mean +/- the SE (error bar) shows you the most likely values for your estimate (for example, the mean).

    ii. The more variable your data are, the larger the SE is going to be.
    iii. To calculate SD, "=STDEV(*cells*)"
    iv. To calculate SE, "=STDEV(*cells*)/SQRT(sample size)"; this equation implies that the larger the sample size, the smaller the SE (more accurate estimates).
    vi. Your sample size is the number of measurements (e.g., measured plants) you have in each sample. You can use" =COUNT()" to calculate sample size.

c. You can use SEs to get a sense for whether groups in your dataset are statistically different from one another; **Reading Error Bars**:

i. If the two means are similar and there is overlap in SE→ the difference you see is likely not statistically significant
ii. If the two means are far apart and there *is* no overlap in SE→ any difference you see is likely statistically significant
iii. Remember: visualizing SE tells you about the degree of overlap between your two groups of data, but you need to conduct a statistical test to determine if this difference is statistically significant.

d. If your dataset contains outliers, it may be more helpful to look at the median in addition to the mean. **How are these two values different from each other)?**
    i. "=MEDIAN()". For example, "=MEDIAN($C$3:$C$12)" would give you the median of the numbers in column C and rows 3 through 12.

*** *Creating*** a scatterplot of your dataset can also help you understand the distribution of the data. It can allow you to visualize any patterns in the distribution and identify outliers; you can insert a plot in a spreadsheet with the following string: **Insert→ charts→ scatter→ select data**

Common statistical tests: *Note: the following directions are written for a PC version of Excel; also, the tests below are parametric tests and as such rely on data being somewhat normally distributed.*

1. ***T-Test*:** This analysis is used to compare **two** populations of data, such as size of prairie dog populations in a grazed *versus* ungrazed field or the size of willow plants growing along Fountain Creek and Monument Creek, for example.

    a. Before beginning any statistical test, define your research question. This crucial step will allow you to choose the correct statistical test and correctly interpret the results.
    b. State the "**null hypothesis (H_o)**" for your research question (e.g., no difference in populations); the alternate hypothesis (H_a.) is that the two 'populations' are biologically unique. The results of your statistical test will provide support for either the H_o or the H_a.
    c. Make sure that the two columns of data you are comparing are side-by-side.
    d. Before you proceed with your test, you should restate your research question. Also, you need to remember that you will be performing a test of a statistical hypothesis, essentially testing whether the two data 'populations' (columns) have means and variances different enough that each can be considered distinctly separate populations.
    e. In Excel, click on '**Tools**" at the top of the screen, select '**Data Analysis**' on the drop-down (or if not an option, click on 'Add Ins' and select 'Analysis Tool Pak'), and select '**t-test: two-sample assuming unequal variances'**. Click 'ok'.
    f. In the small screen that pops up, you will need to add your data in the 'Input'. For the 'Variable 1 range', click on the small box with the red arrow at the right end of the entry area. On the main screen, click and drag your mouse over the entirety of one of your data columns (it does not matter which column)—this essentially enters your range of data. Click the red area in the small box again, which takes you back to the

't-test" screen and do the same thing for the 'Variable 2 range' with your 2nd data column. Then, under 'Output Options', click on 'Output range', and then to the right of this phrase click on the small box with the arrow. On the main screen, decide where you want the results of your statistical test to be displayed (you need an area of approximately 3 cells wide by 12 cells deep), and click on any empty cell. Click 'ok'. In your t-test results, focus on three parts: 't **Stat**', '**P(T<=t) two- tail**', and '**t Critical two-tail**'. If the 't Stat' value is larger than the 't Critical', then you know your H_a is supported, and if the reverse is true, then you know your H_o is supported.

-The **'P' value** tells you the statistical probability of obtaining your results by chance alone. Biological researchers typically use a cutoff value (α) of 0.05 for determining whether the data support the null hypothesis or the alternate hypothesis. If the P-value ≤ 0.05, then there is only a 5% probability that the pattern in your data are attributable to chance, and a 95% probability that the pattern has a biological explanation.

2. ***Linear Regression***: This analysis is used to compare two continuous variables to determine how they are related to each other. A simple linear regression shows you the relationship between the independent and dependent variable, e.g., how much of the variability of X predicts the variability of Y. You can run a regression through 'Tools', then 'Data Analysis' and selecting 'Regression'.
    a. For the 'Input Y range', click on the small box with the red arrow at the right end of the entry area. On the main screen, click and drag your mouse over the entirety of your **dependent variable** (the one whose values presumably depend on, or are influenced by, the corresponding values of the **independent variable**)—this essentially enters your range of data. Click the red area in the small box again, which takes you back to the 'Regression" screen and do the same thing for the 'Input X range' with your 2nd data column.
    b. Among the output for this analysis, you will see "**r-squared**" which describes the degree to which your regression line or linear model explains the variation in your data, i.e., the goodness of fit. Also, the **slope of the line or "m"** is the degree to which your x variable predicts the y variable – note: the analysis will also sometimes assign a **p-value to the slope term, indicating whether your regression line slope is significantly different from zero**.

3. ***ANOVA***: In contrast to a T-test which compares the dependent variable between two independent groups or populations (i.e., control vs. treatment, or pop. 1 vs. pop. 2), an Analysis of Variance (ANOVA) compares the **dependent variable among three (or more) independent groups or populations**. For example, imagine you want to know if the diameter at breast height (DBH) of ponderosa pine trees differs significantly among three sites.
    a. Each site occupies its own column, so you will have three columns. Each cell is a DBH measurement for one tree.

b. Data → data analysis → ANOVA: single factor
c. Select the input range, grouped by columns. Keep alpha at 0.05. **What is alpha and how does it relate to the p-value?**
d. The summary output will give you basic descriptive statistics for each column (site). To determine if the three groups differ significantly from each other, report the F, degrees of freedom (df = n-1), and the p-value. Report the values for between groups (indicating whether the populations are different from each other), not within groups (this is an estimate of the experimental error).

Data Presentation Tips: *Note: the following directions are written for a PC version of Excel.*

1. ***Tables***: used to present a substantial portion of data to the reader—too much to show in a particular graph or chart. Tables should present data that have been transformed to more easily show patterns or data summaries (such as means and standard errors).

2. ***Graphs***: used to present important patterns in the data. Include a very descriptive caption at the **bottom** of the page. How to create graphs in MS Excel:

a. Line (in MS Excel)
i. Select desired data columns by dragging mouse over all desired cells of data.
ii. In Excel, click on 'Insert' at the top of the screen, and then select either 'XY Scatter' or 'Line'.
iii. Right click on the top of the graph and choose 'Select Data' to modify which column(s) of your data should be the horizontal (independent) vs vertical (dependent) variables.
iv. Note also that by right-clicking on the graph, you can modify the chart type or format the plot area.
v. Drag graph into desired position and manipulate size by clicking and dragging on corners of graph.

b. Bar (in MS Excel)
i. Select desired data columns by dragging mouse over all desired cells of data.
ii. In Excel, click on 'Insert' at the top of the screen, and then select 'Bar'. In the drop-down that appears, you have the option of selecting several different types of bars, or you can select 'All Chart Types' at the bottom for more options (including bars that appear vertically—which is generally what you should use in Biology).
iii. Right click on the top of the graph and choose 'Select Data' to modify which column(s) of your data should be the horizontal (independent) vs vertical (dependent) variables.
iv. Note also that by right-clicking on the graph, you can modify the chart type or format the plot area.
v. Drag graph into desired position and manipulate size by clicking and dragging on corners of graph.
vi. **TO ADD CUSTOM STANDARD ERROR (SE) BARS**:
i. Make sure you have your means in 2 different cells (in 2 different columns) that you used to make your bar graph. Type in the corresponding

SE in the cell directly below the means (if you still need to calculate your SE, follow the directions for Summary Statistics under A2 above).

ii. Click on the chart area of your bar graph once, and you should see a little PLUS symbol pop up near the upper right corner of the chart area—click on this.

iii. Click on Error Bars, then click on the little arrow that appears to the right, then click on Standard Error, then click on More Options, then click on Custom at the bottom of the list. For Positive Error Value, click on the little red arrow to the right and then drag your cursor over the 2 error bars cells on your spreadsheet. Repeat for Negative Error Value—dragging your cursor over the same 2 cells on your spreadsheet.