Chapter 8

Data Analysis and Interpretation

Objectives

Statistics. Understand how and why statistics are used to analyze data. Understand calculations for the arithmetic mean and standard deviation of a set of values. Understand how and why the *t*-test is used for data analysis and interpretation.

Graphs. Know which type of graph (*e.g.* bar, line, or pie) is used for which type of data. Understand why graphs are useful for analyzing and interpreting data.

Introduction to Statistics

Government, scientists, doctors, lawyers, economists, businesses, and more utilize statistics. Health professionals may ask whether a medication is effective, for example, to lower blood pressure. To answer this question, experiments are performed on patients, and data, such as blood pressure measurements, are collected. Statistics and graphs allow the health professionals conducting the experiment to evaluate the effectiveness of the drug and to communicate their findings. Although health professionals use hundreds or thousands of patients to perform the experiment and would measure blood pressure many times throughout the experiment, producing thousands or hundreds of thousands of measurements, we will use an unrealistically small data set of ten patients and only the initial and final blood-pressure measurements for demonstration.

Basic analysis of Data

Arithmetic Mean. The arithmetic mean ($\bar{x}$) is the average. The arithmetic mean of a set of numbers is calculated by summing the numbers and dividing by the total number of values in the set.

arithmetic mean ($\bar{x}$) = $\dfrac{\sum x_1}{n}$

where x_1 represents each value and n represents the total number of values

For the example experiment designed to test if a medication is effective for lowering blood pressure (considering only diastolic[1]):

Initial diastolic blood pressure of patients who are given a placebo	Final diastolic blood pressure of patients who are given a placebo	Initial diastolic blood pressure of patients who are given experimental medication	Final diastolic blood pressure of patients who are given experimental medication
109	101	104	90
105	96	100	81
99	104	101	92
104	106	100	83
98	99	98	95
93	100	98	84
100	107	103	94
92	97	97	88
107	105	103	89
103	100	102	94

The arithmetic means for the four sets of measurements in the example are:

101	101.5	100.6	89

Standard Deviation

For the example, in which samples of the populations were measured, notice that the arithmetic means of the initial blood pressure measurements are the same, but the measurements varied more in the patients who received a placebo than those that received the medication. An indication of the variance around the arithmetic mean in a set of numbers better describes the set than the mean alone. Calculating the standard deviation around the arithmetic mean takes into account how much each value deviates from the mean of the values and the total number of values as shown in the following equation for standard deviation.

$$\text{standard deviation (s)} = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$$

Thus:

1. Subtract the arithmetic mean from each value and square the difference. (Note that squaring the difference eliminates the direction of the difference.)

2. Sum the differences from step 1.

3. Divide the sum of step 2 by the number of values in the number set minus 1.

4. Calculate the square root of the result of step 3.

[1] Diastolic pressure is a measurement of the lowest pressure in the ventricles and atria when the heart relaxes after contraction in preparation for refilling during the cardiac cycle.

The standard deviations for the four sets of measurements in the example are:

5.7	3.8	2.4	5.0

Thus, the variation around the arithmetic mean of the initial measurements of the patients given placebos was greater than that of the patients given medication. Variation is important for determining if differences in arithmetic means are indicative of differences in the set of values.

t-test

Based on the arithmetic means and the variance, do you think that the blood pressure medication affects diastolic blood pressure? Although the arithmetic mean final blood pressure for patients given medication is lower than that of patients given placebos, the variation in patients given medication is greater than those given placebos. Statistics provides an accepted method, the *t*-test, to determine if there is a difference between two sets of values. The *t*-test takes into account the averages and the standard deviations of the two sets of values being compared. The *t*-test results in acceptance or rejection of the null hypothesis that there is no difference between the two sets of values. The *t*-test is performed as follows mathematically; however, as described in the following section, software is an efficient way of performing the *t*-test.

1. Calculate the *t*-value from the following equation.

$$\frac{\left|\bar{x}_1 - \bar{x}_2\right|}{\sqrt{\left(\frac{s_1{}^2}{n_1}\right)+\left(\frac{s_2{}^2}{n_2}\right)}}$$

where subscripts 1 and 2 represent the two sets of values being compared

2. Calculate the degrees of freedom by $n_1 + n_2 - 2$ (*i.e.* *n*-1 for each set).

3. Use a *t*-distribution table (available on the internet or in any basic statistics textbook) to determine the critical *t*-value. To use the table, the degrees of freedom (above) and the critical *p*-value will need to be known. Determine the critical *p*-value that will be used to determine differences. For most biological experiments (and our purposes in this course), a *p*-value ≤ 0.05 (*i.e.* the probability that the two sets of data are different by chance is ≤ 5%) will be considered different. If the *t*-value for the comparison being made is greater than the critical *t*-value, one concludes a difference and *vice versa*.

In the example experiment, the *t*-value for the comparison of the initial and final blood pressures of the patients given placebos is 0.23 and for the patients given experimental medication is 6.75. Based on the *t*-distribution table, the critical *t* value is 2.1; therefore, there is not a difference between the initial and final blood pressures of those patients given the placebo, but there is a difference between the initial and final blood pressures of the patients given the experimental medication. From the *t*-test, we can conclude that the medication is effective at lowering diastolic blood pressure.

When the *p*-value is presented, as often given by software calculations (see below), significance is determined based on the critical *p*-value. If the critical *p*-value is 0.05, a *p*-value ≤ 0.05 (*i.e.* the probability that the two sets of data are different by chance is ≤ 5%) will be considered different. In the example experiment, the *p*-value for comparing the initial and final for patients who are given placebos is 0.82, and the *p*-value for comparing the initial and final for patients who are given medication is 0.000003. Thus, we would conclude that the medication has an effect on blood pressure, but the placebo does not. Note that the lower the *p*-value, the greater the difference is between the two sets of values (*i.e.* the greater the effect).

Although they are not described here, several important assumptions (*e.g.* random sampling) regarding the two sets of numbers that are being compared that must be considered when utilizing this test.

*Using software (*e.g. *Microsoft Excel) for basic statistics*[2]

Use of software to perform statistical analyses is accurate and efficient; however, it is important to understand the premises of the computations performed by the software and to be able to provide accurate information to the software regarding the experimental design and criteria. One common program that performs basic calculations and creates basic graphical representations is Microsoft Excel. To perform calculations such as those presented above the following procedure can be followed. In addition, utilize the Help menu.

Calculating the average and standard deviation with Microsoft Excel

1. Enter the data in table format (similar to the table in the above example).

2. Highlight the cell in which you want the result of the calculation to be displayed.

3. Go to the “Insert” menu and choose “Function.”

4. To display all of the calculations that the program can perform, select “All” from the “Function category” in the left box.

5. To calculate, for example, the mean average, choose “AVERAGE” from the list.

6. A box will appear that requires the input information for the calculation. Ways to input the values for which you want to calculate the average follow. One, you can click the arrow to the right of the field and then select the cells in the spreadsheet that contain the values; by holding the mouse button and drag across all of the cells containing values to be included, you can select many cells at one time. (Then press enter.) Two, you can enter the numerical values into each number field. Three, you can enter individual cell coordinates into each number field.

7. Select “OK” and the result of the calculation appears in the selected cell.

8. Similarly, the standard deviation, “STDEV,” can be calculated. (Follow the above, but substitute “STDEV” for “AVERAGE.”)

Calculating the *p*-value with Microsoft Excel

1. Follow the first four steps above and choose “TTEST” from the available functions.

[2] a good website for instructions on using excel: http://www.georgetown.edu/departments/psychology/researchmethods/computer/excel2.htm

2. A box will appear that requires the input information for the calculation. See step six above for ways to input information from your table of values. (It is best to follow the first way). Remember the *t*-test is used to compare two sets of values (*e.g.* placebo and medication). For "Array 1," input the first set of values (*e.g.* placebo or control). For "Array 2," input the second set of values (*e.g.* medication or experimental). For tails, input "2" and for type, input "2."

3. The "TTEST" function results in the *p*-value. See the "*t*-test" section above for how to interpret the *p*-value.

Note: When calculating *t*-values and *p*-values using statistical programs, we will be performing two-tails and assume equal variance. Tails and types of *t*-tests are beyond the scope of this course; however, for more information about tails and types in these analyses, many resources are on the web or in a basic statistics textbook.

Graphs

Data are represented graphically in many different ways. The type of graph chosen to represent data depends on the type of data and the comparisons or relationships necessary for interpretation. In the example experiment involving blood-pressure medication, the data includes initial and final diastolic blood pressures for patients given either a placebo or the experimental medication; thus four averages and four standard deviations. To determine if the medication is effective a comparison between initial and final blood pressure would aid in interpreting the data. A line graph would not be appropriate because the blood pressure was not tracked throughout the experiment. A bar graph would be appropriate; each of the four averages (initial and final of placebo and initial and final of medication) will be represented by a bar. To create a bar graph in Microsoft Excel, the following procedure can be followed.

1. With the data in table format in Excel, choose "Chart" from the "Insert" menu.

2. Select the appropriate type of chart, in this case "Column," and select "Next."

3. Select "Series" at the top of the next box.

4. Select "Add" and name the series, *e.g.* "Initial Diastolic Blood Pressure." The series name can be entered by typing the name in the field or by using the arrow next to the field to select a cell of the spreadsheet containing the name of the series. Enter the values for the series by using the arrow next to the field and selecting the cell, or cells, that contain the numbers to be represented. (Use the control key to select multiple cells.) In the example, the values will be the arithmetic mean initial diastolic blood pressures.

5. Enter the x-axis labels by using the arrow next to the field to select the cells of the spreadsheet that correspond to the averages, *e.g.* Placebo and Medication.

6. If there is another series, *e.g.* Final Diastolic Blood Pressure in the example, follow steps 4 and 5 again.

7. When all the series have been entered, select "Next."

8. In the next box, enter titles for the graph and the axes and adjust any of the other parameter choices.

9. Select “Next” and then name the file, select the preferred location, and select “Finish.”

10. Graphs can be formatted by double clicking on the elements (*e.g.* axes, background, bars, *etc*.).

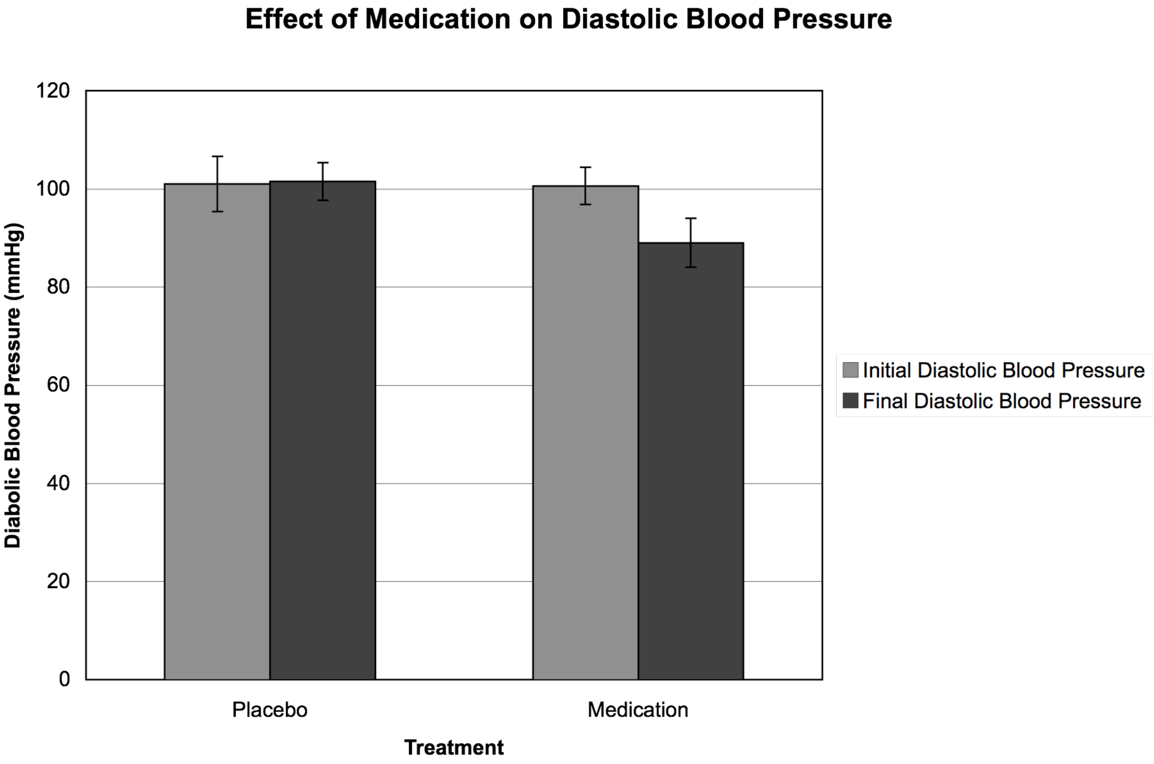
Now, the bars represent the arithmetic means, but it is important to represent the variation around the arithmetic means.

1. Right click on the bars and select “Format data series.”

2. Select “Y Error Bars” from the top of the window.

3. Use the arrow next to “Custom +” to select the cells that correspond the standard deviation (the cells containing the standard deviations must be selected in the same order as the corresponding bars). Repeat with the “Custom – “ field.

Finally, it is important to refer to the graphical representation and the statistical analyses of your data when discussing the experiment.



Review Questions

1. What type of graph would best represent the data from an experiment that was aimed to determine the effect of a plant-growth regulator on plant height if the height were measured every other day for 14 days? Without using numbers, sketch or use Excel to create the graph of measured control plants (without gibberellic acid) and experimental plants (with

gibberellic acid). Include a graph title, axes titles, legend, data lines, and standard deviation bars.

2. If you needed to be treated for a condition, such as high blood pressure, with medication, would you choose a medication that, in studies, has a *p*-value of 0.05 and costs 2 cents per pill or a medication that has a *p*-value of 0.001 and costs 10 cents per pill? Why?