# Statistics in Biology

BI-102



Revised 9/07

*customized for MS Office 2007 suite*

The goal of most biological experiments is to determine whether some treatment affected some measurable. In this case, the treatment is the independent variable, and the measurable is the dependent variable. Almost every time we conduct an experiment, there will be some variation in the dependent variable when the independent variable is changed. However, we must use some method to determine whether this variation was simply the result of random experimental error or an actual effect. Statistics is the field that provides us with such tools.

In this course, we will consider two different types of experimental designs and the simple statistical methods used to analyze the results.

Case 1: Qualitative or Discrete Treatment Levels

Consider an example in which we would like to determine whether the height that people can jump is different before and after 8:00am. In this case, the independent variable is "time of day", and it has two different levels, (1) before 8:00am and (2) after 8:00am. The dependent variable is height, which we measure. A good experimental design will have many control variables. These might include gender, age, height, and weight. To control these variables, we might insist that all experimental subjects are of the same gender and fall into small ranges of age, height, and weight. There will inevitably be some confounding variables, which we are unable to control. For example, we may be unable to control athletic ability, because we are only able to find athletes that get up before 8:00am, and these athletes are unable to participate in our study after 8:00am because they have practice. In this case, if we find an effect, it is up to us to admit that we cannot tell whether time of day or athletic ability was responsible for the effect.

Let's say that two groups conduct this experiment. Group 1 obtains the following data:

| Height (inches) | |
|---|---|
| Before 8:00am | After 8:00am |
| 22 | 25 |
| 26 | 28 |
| 27 | 26 |
| Average = 25.00 | Average = 26.33 |

Can this group claim that the "before 8:00am" subjects had a different height than the "after 8:00am" subjects? The answer is no.

Group 2 obtains the following data:

| Height (inches) | |
|---|---|
| Before 8:00am | After 8:00am |
| 22 | 25 |
| 26 | 28 |
| 27 | 26 |
| 22 | 25 |
| 26 | 28 |
| 27 | 26 |
| 22 | 25 |
| 26 | 28 |
| 27 | 26 |
| 22 | 25 |
| 26 | 28 |
| 27 | 26 |
| 22 | 25 |
| 26 | 28 |
| 27 | 26 |
| 22 | 25 |
| 26 | 28 |
| 27 | 26 |
| Average = 25.00 | Average = 26.33 |

Can this group claim that the "before 8:00am" subjects had a different height than the "after 8:00am" subjects? The answer is yes. Notice that both groups obtained the same average heights for both sets of subjects, so how can this be? The answer is that Group 2 has more confidence in their averages, because they tested more subjects; therefore, they are able to detect smaller differences between sets of subjects.

The statistical test used to draw the conclusions above is called the Student's t-test (or just t-test). We start with a hypothesis, usually something that we would like to demonstrate is false. This is called a *null hypothesis*. For a t-test, the null hypothesis states that two sets of data have the same average. The result of a t-test is known is a *p-value* (note that many different statistical tests yield p-values). A p-value reflects the statistical likelihood that we are making a mistake if we reject the null hypothesis. The smaller the p-value, the more confidently we can reject the null hypothesis, and therefore the more likely that there is an actual difference between the two sets of data. In the scientific community, it is generally agreed upon that the cutoff p-value is p= 0.05.

T-tests are easy to perform using Excel. To do this, enter your data in columns as shown above. Find an empty cell, somewhere below the data itself, and type:

=ttest(B4:B6,C4:C6,2,2)

In this formula, B4:B6 describes the cells for the first group of data (in column B in rows 4 through 6), and C4:C6 describes the cells for the second group of data. The values of 2 in the formula refer to parameters in the test that need not concern us here. When you hit enter, the p-value will appear.

For the data from Group 1, the p-value is 0.492 (try it), and for the data from Group 2, the p-value is 0.034. Since only the p-value for Group 2 is less than the cutoff of 0.05, only this group can report a statistical difference between its sets of subjects.

The following is a TRUE statement that Group 2 could make regarding their data:

*Subjects jumping after 8:00am jumped higher than those jumping prior to 8:00am (p=0.034).*

Notice that the claim is supported by a p-value. This is expected in scientific writing. *Never make a quantitative claim without supporting it with a p-value.*

So what can Group 1 say? Unfortunately, if one is unable to reject a null hypothesis (i.e., the p-value is greater than 0.05), there is little to say conclusively. There are two reasons why a null hypothesis might not be rejected:

(1) There may actually be no effect of the independent variable on the dependent variable.
(2) There may have been an effect, but our data were too noisy to detect it.

It is impossible to determine whether choice (1) or (2) was the actual case. Therefore, we must be extremely careful about the conclusions we make when we cannot reject our null hypothesis. The following are examples of FALSE statements made by Group 1:

*Subjects jumping after 8:00am did not jump a different height than those jumping before 8:00am.*

*Subjects jumping after 8:00am jumped higher than those jumping prior to 8:00am.*

Make sure you understand why each of these statements is false; you will receive deductions on your lab report if you make such false statements. Here is something that Group 1 could claim that would not be false:

*Subjects jumping after 8:00am jumped higher than those jumping prior to 8:00am, but not to a level of statistical significance.*

Finally, be careful with statements you make regarding cause and effect. While Group 2 could state that subjects jumping after 8:00am jumped higher than those jumping prior to 8:00am, it would be a stretch to say that time of day had an effect on jumping height. While we would like to make this statement, some of the confounding variables (like athletic ability) may have also been affecting jumping height. Therefore, we could not possibly make such a statement unless we demonstrated that we have eliminated the effects of all possible confounding variables (which is impossible to do). The Discussion section of a report is usually a good place to talk about all of the possible contributions of confounding variables, whether you found an effect or not.

It is also desirable to give a visual demonstration of how two sets of data compare. This is commonly done using a bar graph in which the average values of the two sets of data are graphed. Figure 1 shows such a bar graph for the data from Group 2. There are a few things to note in this bar graph. Error bars have been added to demonstrate how variable the data was. These error bars represent a measure known as *standard deviation.* Some scientists plot a measure known as *standard error of the mean (S.E.M)* instead. S.E.M. values are smaller, but some statisticians argue that this is not a valid representation of the spread of the data, so showing standard deviation is the more conservative route to take. Next, note the caption in the figure. This caption is short, but yet it tells exactly what is shown in the figure. One important item in the caption is "n = 19". The *n* value tells how many replicates were performed. It is the number of data points that were averaged for each bar. It is also important that the reader know this value to help make judgments regarding differences. Also, note that the y-axis gives a range that starts at zero. By default, Excel will zoom in on a range that is covered by the data (perhaps 24 to 30 inches). However, zooming in on one's data to make differences appear more significant is considered misleading. Finally, note that the y-axis contains both the measured quantity, height, and its units, inches. Do not forget to include units. Bar graphs are nice visual representations of data, but remember that no conclusions can be drawn from them in contrast to the statistics discussed above.

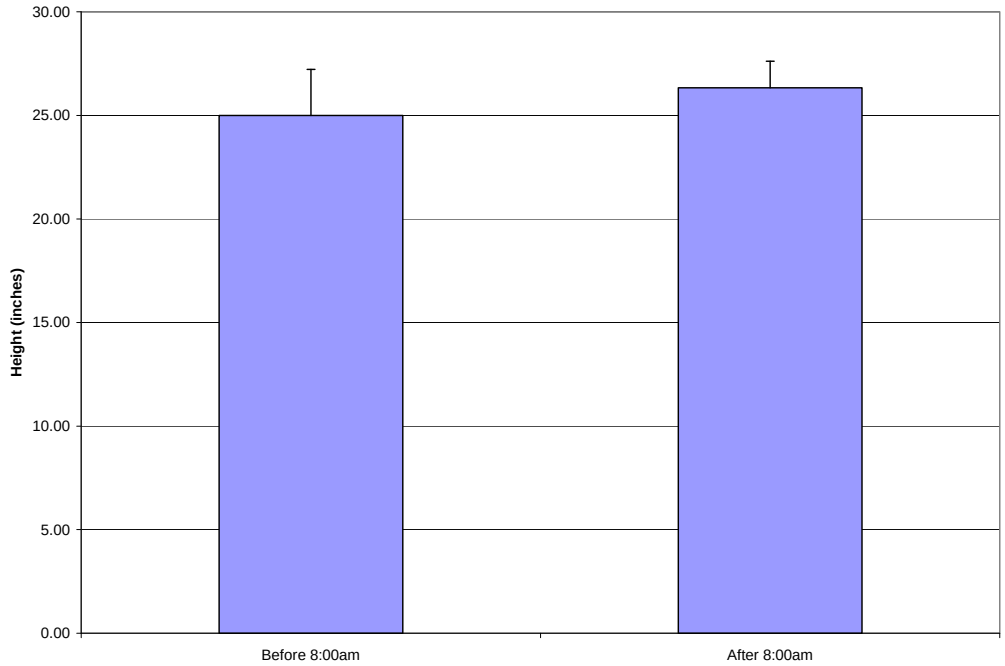


**Figure 1:** Jumping heights measured in subjects before and after 8:00am. Error bars represent standard deviation. n = 19.

Here are the steps taken in Excel to construct Figure 1:

- Data were listed in two columns, as above. Headings (*Before 8:00am* and *After 8:00am*) were placed at the top of the columns.
- Below each set of data, the average and standard deviation were calculated. To calculated an average, type:

  =average(E3:E20)

  where E3:E20 is the range of cells containing the data. To calculate the standard deviation, type:

  =stdev(E3:E20)
- Select Insert…Column (under "Charts")…select the first 2-D Column option.
- Under Data, choose Select Data.
- Hit the Remove button until all series are removed.
- Hit the Add button.
- In the Values field, enter the range of cells containing the averages (ex. E21:F21).
- Hit OK.
- Under Horizontal (Category) Axis Labels, hit Edit.
- Enter the range of cells containing the headings (ex. E2:F2).
- Hit OK twice until the menus disappear.
- Clean up the graph:
  - Click on the legend (where it says "series 1") and hit the delete key to delete it
  - Right click on one of the numbers on the y-axis, and select Format Axis
    - Next to Minimum, select "Fixed" and enter 0 in the field.
  - Select the Layout tab in the main menu bar, select Axis Titles, Primary Vertical Axis Title, Rotated Title. Click inside that title (where it says "Axis Title") and replace the text with "Height (inches)".

- Add error bars:
    - Still under the Layout tab, in the Analysis menu, select Error Bars, followed by More Error Bars Options…
    - Select Custom, and hit Specify Value.
    - In the Positive Error Value, enter the cells containing the Standard Deviation values (ex. E22:F22).
    - Repeat for Negative Error Value.
    - Hit OK and Close to get rid of all of the menus.

To copy your chart into an MS Word document, click somewhere near the exterior of the chart, and select Copy from the Home tab. Open up the MS Word document and place the cursor at the point of insertion. From the Home tab, hit the small arrow under Paste. Select Paste Special…/Bitmap/OK. This inserts a slightly cleaner version of the graph than would the normal Paste command.

In this example, the independent variable, time of day, had two discrete levels, before 8:00am and after 8:00am. If the independent variable has more than two levels, there are two different ways of analyzing the data. One way to analyze it is to make separate comparisons between the categories using the procedure discussed in this section. For example, if the three categories were before 8:00am (I), between 8:00am and noon (II), and after noon (III), then we could make three separate comparisons: I vs. II, I vs. III, and II vs. III. For each comparison, we would perform a t-test as discussed above. Technically, since we are making three comparisons, we are increasing our likelihood of making an error in at least one of the three, so instead of using a cutoff p-value of p=0.05, this cutoff p-value should be divided by the number of comparisons we are making, 3. This is called the *Bonferroni correction*. In practice, scientists often neglect this correction and continue to use p=0.05 as the cutoff for every comparison. The second way to analyze data associated with multiple independent variable levels will be discussed next.

Case 2: Continuous Independent Variable With a Predicted Linear Effect

Another way to analyze the relationship between two variables is to construct a mathematical model that relates them. The simplest such model is a linear relationship, and the method used to assess linear relationships between two variables is called *linear regression*. Linear regression is a very useful tool, but it should only be used when two conditions are met: (1) there are more than two levels of the independent variable and it is continuous, and (2) you expect the dependent variable to vary linearly with the independent variable. For the three categories discussed above (before 8:00am, between 8:00am and noon, and after noon), linear regression would not be appropriate, because the three levels are not from a continuous scale.

One example in which we might use regression analysis is if we were looking at the relationship between jumping height and body weight. This is appropriate because body weight comes from a continuous scale (it can be any number between zero and infinity). Furthermore, it might be sensible to predict that jumping height will vary linearly with body weight, although we would probably predict a negative slope (jumping height will go down as body weight goes up). Note that in such an experiment it would be important to control variables such as height and age to maximize the likelihood that body weight is indeed accounting for any effect that we observe.

The generic linear regression model is given by:

$$Y = \beta_0 + \beta_1 X$$

You may recognize this as looking similar to Y = mX + b equation of a line. Here, $\beta_1$ is equivalent to *m*, the slope of the line. $\beta_0$ is equivalent to *b*, the y-intercept of the line. X is the independent variable, and Y is the dependent variable.

A good way to determine if indeed there is a linear relationship between X and Y is to determine whether the slope is different from zero. If the slope is equal to zero, then that means that Y will have the same value for any value of X, so it does not depend on X. Statistically, it is common for a null hypothesis to be formed that states that the slope, $\beta_1$, is equal to zero. As above, this null hypothesis is one that we would like to demonstrate is false, so we seek to reject it. Regression output (using a statistical test beyond the scope of this tutorial) will give a p-value associated with this null hypothesis. This p-value has the same interpretation as above. If $p < 0.05$, we can confidently reject the null hypothesis, thereby concluding that indeed there is a linear relationship between the two variables.

A measure of how well the data fit a regression model is called the *R-squared* value. This value corresponds to the fraction of the variability of the data that is accounted for by the regression model. In other words, it is a measure of how close the experimental data come to the regression line. If the data fall right on the line, R-squared will be equal to one; the further from the line they fall, the closer R-squared gets to zero. The interpretation of R-squared is sometimes quite vague. In some cases the model is considered good only if R-squared is greater than 0.99. In other cases, such as with biological data, an R-squared value as low as 0.50 is considered acceptable. Note that it is possible for a relationship to be very weak (having a slope close to zero) but yet still obtain a very high R-squared value if the experimental data fall close to this nearly horizontal regression line. For this reason, the R-squared value should only be used secondarily, and never as a substitute for the p-value of the slope. If the slope is found to be significantly different from zero ($p < 0.05$), then the R-squared value can be given to demonstrate how well the data fit the model.

To conduct the experiment between jumping height and body weight, we collect jumping height and body weight data from several individuals of a common gender, age, and height. In Excel, the data are placed in two columns.

Before beginning, you will need to make sure the Analysis Toolpak is installed for your Excel.

- Click on the Microsoft Office Button, and then click Excel Options
- Click Add-Ins, and then in the Manage box, select Excel Add-Ins
- Click Go
- In the Add-Ins available box, select the Analysis ToolPak check box, and then click OK.
- After you load the Analysis ToolPak, the Data Analysis command is available I the Analysis group in the Data tab.

To perform the regression:

- Under the Data tab, select Data Analysis from the Analysis group.
- Select Regression
- Hit OK
- In the Regression dialog box, hit the small icon next to the field box for Input Y Range
- In the spreadsheet, highlight the data for the dependent variable (jumping height)
- Hit the icon to take you back to the dialog box.
- Hit the icon next to the field box for Input X Range
- In the spreadsheet, highlight the data for the independent variable (body weight)
- Hit the icon to take you back to the dialog box.
- Check the circle next to Output Range.
- Hit the icon next to the field box for Output Range.
- In the spreadsheet, click where you would like the upper-left corner of the regression table to be placed. The table is quite large (see below).
- Hit the icon to take you back to the dialog box.
- Hit OK.

In the spreadsheet, your regression table should appear. The regression table should look like that shown in Fig. 2. The second value under the P-value column heading (outlined) is the p-value for the slope of the regression line. Notice that, in the figure, this has a value of $1.26 \times 10^{-8}$. Since this is much less than 0.05, we can say with confidence that the slope of the line is different from zero. The second value under the Coefficients column heading is -0.104. This is the actual slope of the regression line. Since it is negative, we can say that there is a *negative linear relationship* between jumping height and body weight. Finally, the R Square value (outlined) is 0.825. Since this is fairly close to one, we could conclude that the data fit the linear model quite well.

| Body Weight (lbs) | Jumping Height (inches) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 125 | 30 | SUMMARY OUTPUT | | | | | | | | |
| 128 | 26 | *Regression Statistics* | | | | | | | | |
| 135 | 32 | Multiple R | 0.90832657 | | | | | | | |
| 136 | 29 | R Square | 0.825057158 | | | | | | | |
| 139 | 28 | Adjusted R Square | 0.81584964 | | | | | | | |
| 146 | 29 | Standard Error | 1.414292641 | | | | | | | |
| 152 | 27 | Observations | 21 | | | | | | | |
| 155 | 28 | ANOVA | | | | | | | | |
| 158 | 26 | | *df* | *SS* | *MS* | *F* | *Significance F* | | | |
| 160 | 25 | Regression | 1 | 179.2338454 | 179.2338454 | 89.6069013 | 1.26425E-08 | | | |
| 161 | 28 | Residual | 19 | 38.00424982 | 2.000223675 | | | | | |
| 161 | 27 | Total | 20 | 217.2380952 | | | | | | |
| 165 | 25 | | | | | | | | | |
| 170 | 24 | | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 95.0%* | *Upper 95.0%* |
| 173 | 26 | Intercept | 42.91464437 | 1.833154795 | 23.41026763 | 1.78798E-15 | 39.07780729 | 46.75148144 | 39.07780729 | 46.75148144 |
| 179 | 24 | X Variable 1 | -0.10357772 | 0.010941972 | -9.466092187 | 1.26425E-08 | -0.126479531 | -0.080675908 | -0.126479531 | -0.080675908 |
| 181 | 25 | | | | | | | | | |
| 190 | 23 | | | | | | | | | |
| 198 | 21 | | | | | | | | | |
| 214 | 21 | | | | | | | | | |
| 242 | 18 | | | | | | | | | |

**Figure 2:** Spreadsheet showing the regression analysis between body weight and jumping height.

As a visual demonstration of the linear regression, we can plot the data. To do this, with the spreadsheet containing the data opened:

- Under the Insert tab, from the Charts group, select Scatter, and pick the upper left chart.
- Under the Design tab, select Select Data.
- Hit the Remove button continuously until all series are removed
- Hit the Add button

- In the Series X-Values field, enter the range of data for the independent variable (body weight, ex. A3:A23).
- In the Y-Values field, enter the range of data for the dependent variable (jumping height, ex. B3:B23).
- Hit OK twice to clear the menu boxes.
- Under the Layout tab, select Legend and click None.
- Click Axis Titles and add appropriate Horizontal and Vertical Axis titles. Don't forget units.
- Right click somewhere in the data on the graph, and select Format Data Series…
    - Under Line Color, select No Line
    - Hit Close
- Once again, right click somewhere in the data. This time, select Add Trendline…
- Hit Close

You can export your graph to a Word document as explained above. Figure 3 shows the regression plot for this example.

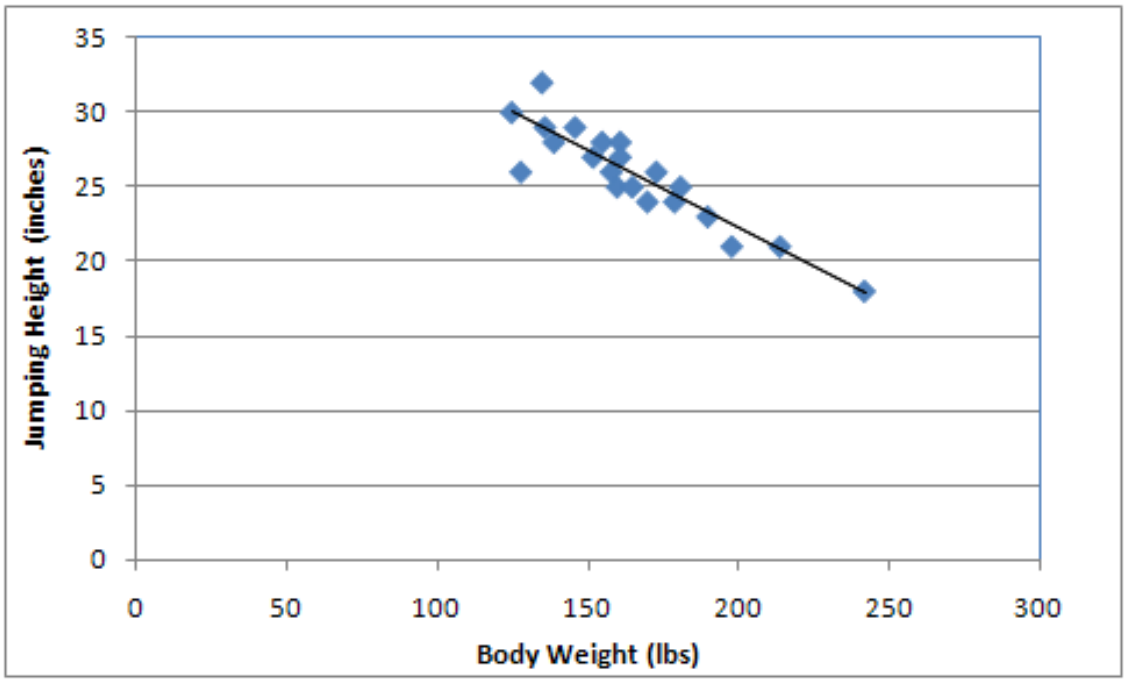


**Figure 3:** Relationship between jumping height and body weight. The solid line is the linear regression fit to the data collected.