

Graphing and Basic Statistics

JOHN D. NAGY

Department of Life Sciences, Scottsdale Community College
9000 E. Chaparral Rd., Scottsdale, AZ 85256

Department of Mathematics and Statistics, Arizona State University
PO Box 874501, Tempe AZ, 85287-4501

1 Introduction

This little handout is meant to be a terse introduction to the use of spreadsheet software in science to perform calculations and make figures (graphs). It is a practical guide specifically for college and university students in majors-level biology. A companion handout—*Basic Statistics Every Scientist Should Know*—goes into the statistics in more detail, but is generally more advanced than necessary for such an introductory course.

2 Using a spreadsheet as a calculating device

2.1 Calculating averages

In most well-designed experiments, the procedures will be repeated under identical conditions at least twice, usually more often. The number of times an experiment is repeated is typically represented as n , which is often called the *sample size*.

The average (also called the mean) is usually easy to calculate using a spreadsheet. For example, suppose I wanted to calculate the mean of the numbers in a spreadsheet's cells A1, A2, A3 and A4 and place the result in cell A5. To do that I'd follow these procedures, which are standard for essentially all spreadsheets:

1. Click on cell A5.
2. Type “=AVERAGE(” exactly this way. (The = sign tells the spreadsheet that what you are entering is a formula, not a literal value.)
3. Highlight cells A1 through A4. Automatically cell A5 will now read, “=AVERAGE(A1:A4”.
4. Hit enter. (Most spreadsheets do not require you to close the parentheses, but if you want to be sure, then close the parentheses before hitting enter.)

The mean of the numbers in cells A1 through A4 will now appear in cell A5.

2.2 Calculating standard deviations

When an experimental treatment is repeated more than once, the results almost always vary. Therefore, in addition to the mean we also have to measure the *variation* in the measurements; that is, we need to find a way to express how “spread out” the values are. One common measurement of variation is called the *standard deviation*. To calculate it using a spreadsheet, follow the procedures for calculating an average, but type “=STDEV(” instead of “=AVERAGE(”.

2.3 Calculating standard errors of the mean

A typically more useful number (not quite the same concept as variance, but related to it) is called the *standard error of the mean* or just *standard error*. It is very easy to calculate once you have the standard deviation, because if we represent the standard error as SE and the standard deviation as s , then

$$SE = \frac{s}{\sqrt{n}}.$$

So, suppose you wanted to calculate the standard error of the numbers in cells A1 through A4, you already had the standard deviation in A5, and you wanted to place your result in cell A6. Then, to calculate the standard error,

1. Click on cell A6.
2. Type “=” and then click on cell A5. Your entry will now read, “=A5”.
3. Type “/SQRT(4)” and hit enter. (NOTE: we take the $\sqrt{4}$ because there were 4 measurements. In general we would take \sqrt{n} .)

3 What type of graph should I use?

The basic purpose of any graph is to express relationships among variables, as indicated by data, in the simplest, most clear way possible. Therefore,

Rule 1 *Always favor a simpler graph over a more complicated one that expresses the same relationship.*

3.1 First, determine the variable types

1. **Independent variables** are called independent because we think of them as logically prior to, typically determining, the dependent variable. You can recognize them in an experiment because *they are the variables being altered and varied by the experimenter*.
2. **Dependent variables** logically depend on the independent variable. In an experiment, *they are the variables being measured*.
3. **Controlled variables** are variables that could affect the dependent variable but that the experimenter is not interested in or not testing. Therefore, *controlled variables are held constant in all replications of the experiment*.

Most experiments test the effects of the independent variables on the dependent variables while holding the controlled variables constant.

3.2 Determine which variables go on which axis

Rule 2 *The horizontal (x) axis almost always represents the independent variable, and the vertical (y) axis almost always represents the dependent variable. NOTE: I use the phrase “almost always” simply to be completely accurate. But interpret this as “always.” No exceptions to any of these rules are likely to arise in an introductory biology course.*

Rule 3 *If time is one of the variables, it is almost always the independent variable.*

3.3 Choose the correct type of graph

To do this you need to classify variables in another way:

1. **Continuous variables** are variables that can be expressed as a real number. (Practically speaking, any number you normally use is a real number. All the integers, like 1 or 567 are real, as are all fractions like $22/7$ or 2.54, and all irrational numbers like π (3.14159 . . .), which can't be expressed as a ratio of integers or a finite decimal fraction.) For example, a person's weight is continuous because it can be essentially any positive real number.
2. **Categorical variables** are usually expressed by some qualitative attribute rather than a number. For example, the color of a car is a categorical variable, as is the type of lung cancers commonly observed—small cell, large cell, glandular and skin-like (squamous).

The two most common types of graphs in scientific communication are **scatter plots** and **bar plots**.

Rule 4 *If both independent and dependent variables are continuous, then use a scatter plot.*

For example, take a look at Fig. 1. This figure shows the relationship between human body weight and basal metabolic rate represented by 44 different people. Both variables graphed—body weight and metabolic rate—are continuous, so a scatter plot is the proper way to express the relationship. Each point on the graph represents the body weight and metabolic rate of a single person.

Rule 5 *If the independent variable is a set of categories, then use a bar plot.*

Figure 2 shows an example of this rule. In this case, we wish to show the percentage of adults who currently smoke broken out by age group. The independent variable is age class (see rule 3), which is categorical. So, a bar plot is the best choice here. Note also that this figure has error bars representing the mean ± 1 standard error.

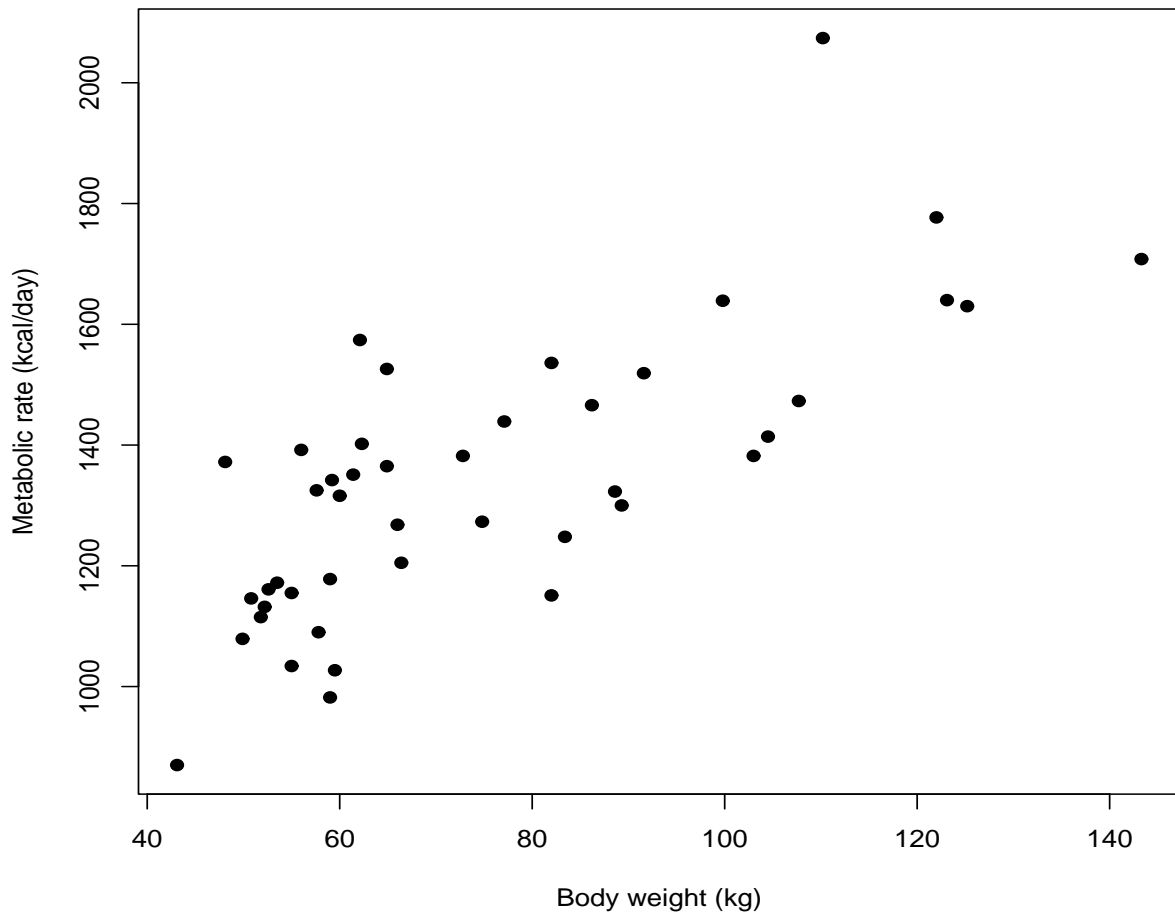


Figure 1: A scatter plot of body weights and metabolic rates for 44 human subjects. There are no error bars since each point represents a single measurement. Data from the ISWR package of the standard R statistical software installation, as described in D.G. Altman's, *Practical Statistics for Medical Research* (1991).

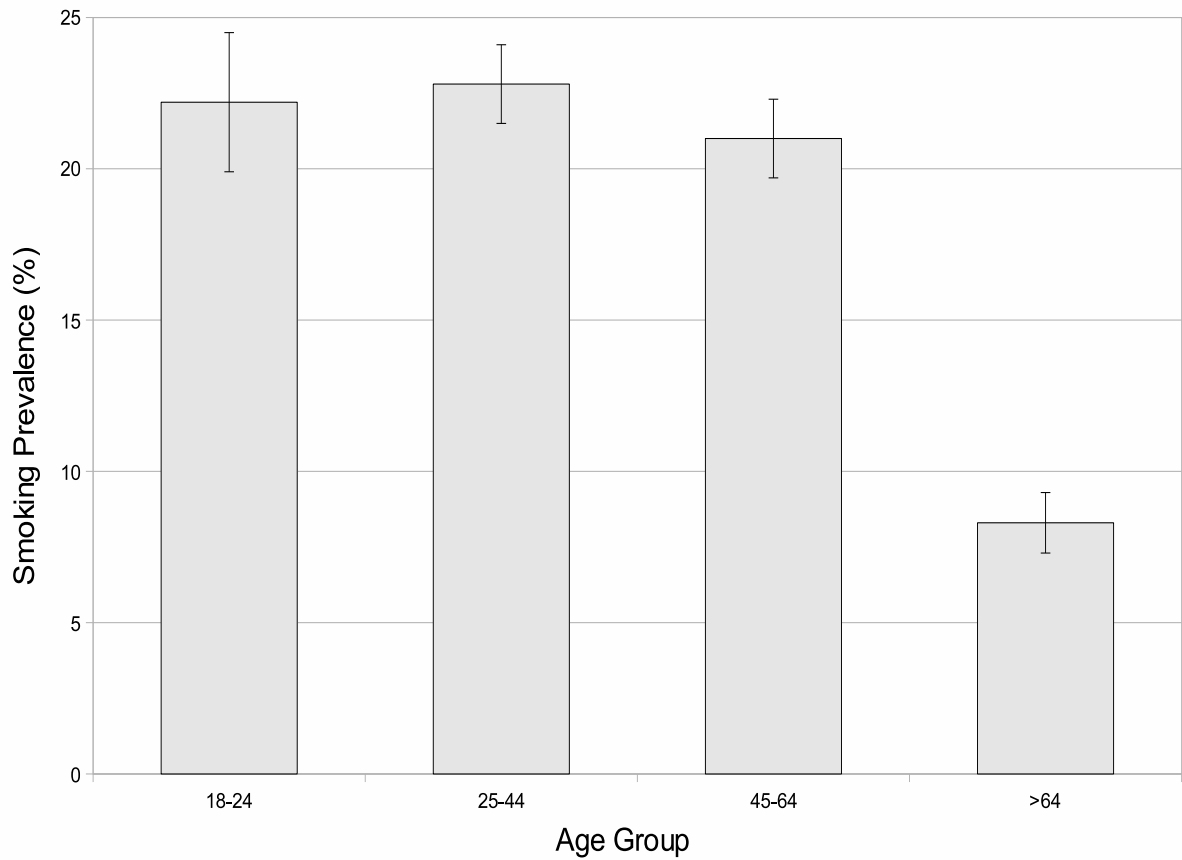


Figure 2: Estimated percentage of U.S. adults aged ≥ 18 years who are current smokers. Data from the Center for Disease Control and Prevention's National Health Interview Survey, 2007. Error bars represent ± 1 standard error of the mean.

Warning 1 *If you use standard spreadsheet software like Excel, Gnumeric or Open Office Calc, be aware that these products are made primarily for business, and so use a different terminology. In particular, for some reason what scientists call “bar plots” or “bar graphs” are called “column graphs” by the business community. **What the business community calls bar graphs are not used in science, so be very careful to use the scientific bar graph, not the business one!***

Warning 2 *Beware of the **line graph** in Excel and other common spreadsheets. This type of graph is essentially a bar graph made to look like a scatter plot with lines connecting the points. It’s rarely useful.*

4 How do I construct the graph?

Most published scientific graphs are produced either by spreadsheet software (Excel, Open Office Calc or Gnumeric, for example), or specialized statistical software (S-plus, SPSS, SAS or R, for example). Spreadsheets are often easier for beginners, and so we will focus on them here. But all professional scientists must at least know someone who is completely familiar with the more advanced statistics packages.

4.1 What do I plot?

Rule 6 *If $n > 1$, then always plot the average (also called the mean) of all n replications. Do not plot all n replications individually.*

Rule 7 *If you plot averages, **always include error bars**. These error bars typically represent ± 1 standard error of the mean or a confidence interval. An average without an error bar is nearly useless.*

4.2 Constructing the plots using spreadsheet software

The precise commands to build a plot vary slightly by software package. However, all packages (except Microsoft’s Excel 2007/8 for Vista) are essentially the same. (The 2007/8 version of Excel is quite odd; the free packages Gnumeric and Open Office Calc are superior, anyway, especially for scientific applications.) The steps below are essentially common to all packages (Gnumeric, Open Office Calc and old Excel; if you use Excel 2007/8 I’m afraid you’re on your own).

1. Highlight *just* the data you wish to plot. Usually you’ll just highlight the means. If you highlight too much, you’ll get a mess.
2. Click the graphing wizard or insert a graph from the main menu bar. (The wizard button looks like a little bar plot for Gnumeric, Calc and old Excel.)
3. On the dialogue box that pops up, click the type of plot you want and then “Next” or “OK” or “Forward.”
4. You should now see a draft of the figure. Look closely at this draft and make sure it’s organized correctly. In particular:

- (a) Make sure each data series is plotted correctly. In the spreadsheet lingo, a data series is the collection of data from one set of experiments or observations. For example, both Figs. 1 and 2 of this handout have only one series. If, for example, we had separated data for men and women, we'd have 2 series, one for each sex.
 - (b) Make sure nothing extraneous is plotted.
 - (c) Make sure each series is correctly labeled. The computer will always default to "Series 1," "Series 2," etc. unless it thinks you've included series labels in what you highlighted. ***If you only have one series, then no legend is required.***
5. If you find any mistakes in how the series are represented or named, go to the dialogue box that allows you to choose the data and labels for each series. Unfortunately, this is where spreadsheets vary the most. In Open Office Calc, highlight the step called "Data Series" and follow the directions. In Gnumeric, click "Series 1," or whichever you need to format, in the upper left-hand pane of the dialogue box, and then click the "Data" tab. From here it's pretty obvious what to do next.
 6. Once the data are properly graphed, label the axes. Again, how this is done varies by software package. In Open Office Calc, go to "Chart Elements" and type in labels for X and Y axes. In Gnumeric, click on the proper axis in the upper left-hand window of the dialogue box, click "Add: Label" and type in the appropriate label.

Rule 8 *Axis labels always include the name of the variable followed by the units in parentheses. For example, "Time (min)" is correct. "Time in min" is not correct. See Figs. 1 and 2 for more examples. NOTE: Categorical variables have no unit, which is why the X-axis in Fig. 2 has no units.*

Warning 3 *Do not add a title to the chart even though the software calls for one. Bolded titles above graphs are a device used in business, but are rarely used in scientific communication.*

7. Complete the graph by adding a caption **below** it. The caption may be any length, and often is an entire paragraph.

Rule 9 *In scientific communication, figure captions always go below the figure (see Figs. 1 and 2 for example) whereas captions for tables always go above the table itself.*

Rule 10 *The first sentence of the caption should be a summary or title of the figure. The rest of the caption **describes** where the data came from. The caption need not include any explanation or interpretation of the data shown. You just need to describe what's in the figure, not say in the caption what it means. In particular, make sure it is clear from where the data come. See Figs. 1 and 2 for examples.*

4.3 How do I add error bars?

Again, the different software packages vary in how error bars are added to the data.

1. For Open Office:

- (a) Insert the graph into the spreadsheet. Then right-click on one of the data points in a scatter plot or one of the bars in a bar plot.
- (b) In the menu that pops up, choose “Insert Y Error Bars.”
- (c) On the dialogue box that pops up, click the “Cell Range” button.
- (d) Under “Parameters” click the button next to the “Positive(+)” data entry line and highlight the standard errors (or other value on which the error bar is based) that you calculated in the spreadsheet.
- (e) Check the “Same value for both” box and then click “OK.”

2. For Gnumeric:

- (a) After making the chart but *before* inserting it into the spreadsheet, highlight the name of the series to which you want to add error bars in the upper left-hand pane of the graph wizard.
- (b) Click on the “Error bars” tab in the main section of the dialogue box.
- (c) Under “Error category” choose “Absolute.”
- (d) Under “Values” place your cursor in the input line labeled “(+)” by clicking anywhere within the input box.
- (e) On the spreadsheet, highlight the standard errors (or other value on which the error bar is based) that you calculated.
- (f) Error bars should now appear in the upper right window showing the current status of the graph.