Excerpted and adapted from: McDonald, J.H. 2014. Handbook of Biological Statistics (3rd ed.). Sparky House Publishing, Baltimore, MD. (http://www.biostathandbook.com/graph.html)

Guide to fairly good graphs

General tips for all graphs

Do start with a descriptive title or caption. Your graph should be able to stand on its own to represent the findings of your investigation.

Don't clutter up your graph with unnecessary junk. Grid lines, background patterns, 3-D effects, unnecessary legends, excessive tick marks, etc. all distract from the message of your graph.

Do include all necessary information. Clearly label both axes of your graph, including measurement units if appropriate. You should identify symbols and patterns in a legend on the graph, or in the caption. If the graph has "error bars," you should say in the caption whether they're 95% confidence interval, standard error, standard deviation, comparison interval, or something else. Most of the You must use graph paper if constructing a graph by hand.

Choosing the right kind of graph

There are many kinds of graphs—bubble graphs, pie graphs, doughnut graphs, radar graphs—and each may be the best for some kinds of data. But by far the most common graphs in scientific publications are scatter graphs and bar graphs.

Use a **scatter graph** (also known as an X-Y graph) for graphing data sets consisting of pairs of numbers. These could be *measurement variables* (see below), or they could be *nominal variables* (see below) summarized as percentages. Plot the independent variable on the X axis (the horizontal axis), and plot the dependent variable on the Y axis.

The independent variable is the one that you manipulate, and the dependent variable is the one that you observe. For example, you might manipulate salt content in the diet and observe the effect this has on blood pressure. Sometimes you don't really manipulate either variable, you observe them both. In that case, if you are testing the hypothesis that changes in one variable cause changes in the other, put the variable that you think causes the changes on the X axis. For example, you might plot "height, in cm" on the X axis and "number of head-bumps per week" on the Y axis if you are investigating whether being tall causes people to bump their heads more often. Finally, there are times when there is no cause-and-effect relationship, in which case you can plot either variable on the X axis; an example would be a graph showing the correlation between arm length and leg length.

There are a few situations where it is common to put the independent variable on the Y axis. For example, oceanographers often put "distance below the surface of the ocean" on the Y axis, with the top of the ocean at the top of the graph, and the dependent variable (such as chlorophyll concentration, salinity, fish abundance, etc.) on the X axis. Don't do this unless you're really sure that it's a strong tradition in your field.

Use a **bar graph** for plotting means or percentages for different values of a nominal variable, such as mean blood pressure for people on four different diets. Usually, the mean or percentage is on the Y axis, and the different values of the nominal variable are on the X axis, yielding vertical bars.

In general, I recommend using a bar graph when the variable on the X axis is nominal, and a scatter graph when the variable on the X axis is measurement. Sometimes it is not clear whether the variable on the X axis is a measurement or nominal variable, and thus whether the graph should be a scatter graph or a bar graph. This is most common with measurements taken at different times. In this case, I think a good rule is that if you could have had additional data points in between the values on your X axis, then you should use a scatter graph; if you couldn't have additional data points, a bar graph is appropriate. For example, if you sample the pollen content of the air on January 15, February 15, March 15, etc., you should use a scatter graph, with "day of the year" on the X axis. Each point represents the pollen content on a single day, and you could have sampled on other days; there could be points in between January 15 and February 15. However, if you sampled the pollen every day of the year and then calculated the mean pollen content for each month, you should plot a bar graph, with a separate bar for each month. This is because you have one mean for January, and one mean for February, and of course there are no months

between January and February. This is just a recommendation on my part; if most people in your field plot this kind of data with a scatter graph, you probably should too.

Measurement variables

Measurement variables are, as the name implies, things you can measure. An individual observation of a measurement variable is always a number. Examples include length, weight, pH, and bone density. Other names for them include "numeric" or "quantitative" variables.

Some authors divide measurement variables into two types. One type is continuous variables, such as length of an isopod's antenna, which in theory have an infinite number of possible values. The other is discrete (or meristic) variables, which only have whole number values; these are things you count, such as the number of spines on an isopod's antenna. The mathematical theories underlying statistical tests involving measurement variables assume that the variables are continuous. Luckily, these statistical tests work well on discrete measurement variables, so you usually don't need to worry about the difference between continuous and discrete measurement variables. The only exception would be if you have a very small number of possible values of a discrete variable, in which case you might want to treat it as a nominal variable instead.

Nominal variables

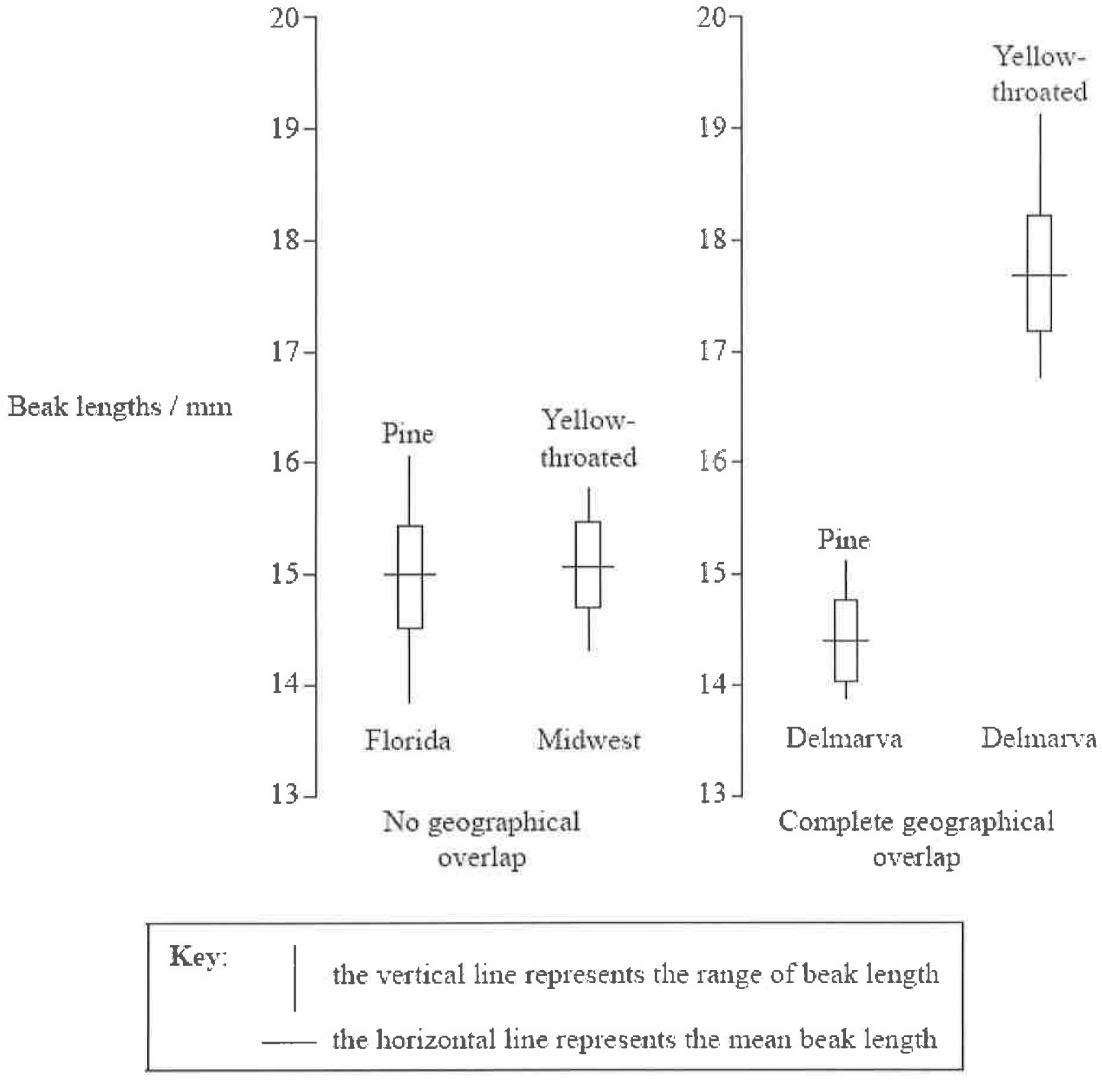
Nominal variables (sometimes known as categorical) classify observations into discrete categories. Examples of nominal variables include sex (e.g. male, female), genotype (values are AA, Aa, or aa), or ankle condition (values are normal, sprained, torn ligament, or broken). A good rule of thumb is that an individual observation of a nominal variable can be expressed as a word, not a number. If you have just two values of what would normally be a measurement variable, it's nominal instead: think of it as "present" vs. "absent" or "low" vs. "high." Nominal variables are often used to divide individuals up into categories, so that other variables may be compared among the categories. In the comparison of head width in male vs. female isopods, the isopods are classified by sex, a nominal variable, and the measurement variable head width is compared between the sexes.

Nominal variables are often summarized as proportions or percentages. For example, if you count the number of male and female *A. vulgare* in a sample from Newark and a sample from Baltimore, you might say that 52.3% of the isopods in Newark and 62.1% of the isopods in Baltimore are female. These percentages may look like a measurement variable, but they really represent a nominal variable, sex. You determined the value of the nominal variable (male or female) on 65 isopods from Newark, of which 34 were female and 31 were male. You might plot 52.3% on a graph as a simple way of summarizing the data, but you should use the 34 female and 31 male numbers in all statistical tests.

It may help to understand the difference between measurement and nominal variables if you imagine recording each observation in a lab notebook. If you are measuring head widths of isopods, an individual observation might be "3.41 mm." That is clearly a measurement variable. An individual observation of sex might be "female," which clearly is a nominal variable. Even if you don't record the sex of each isopod individually, but just counted the number of males and females and wrote those two numbers down, the underlying variable is a series of observations of "male" and “female."

Examples:

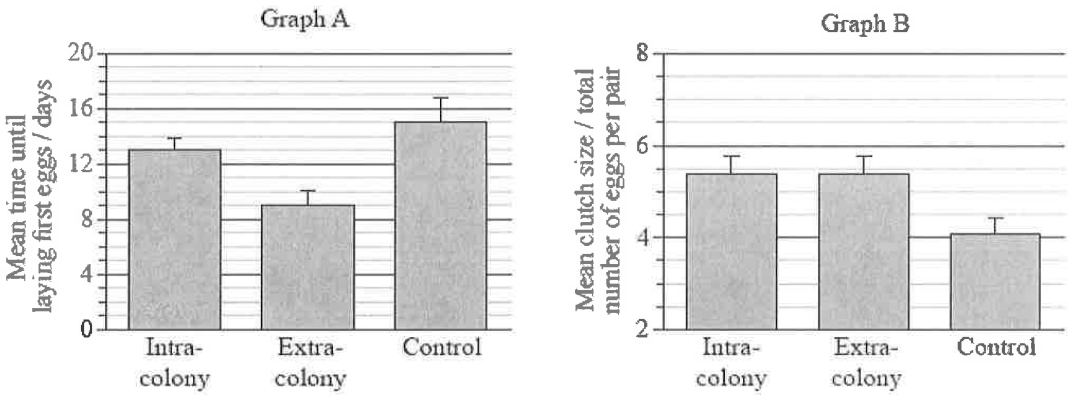
1. Competition between genetically similar species of birds may lead to changes of one or more characteristics. One characteristic that results from this kind of selection is differences in the beaks. Researchers studied the beak lengths of two species of warblers. The graphs below show the beak lengths of Pine Warblers (*Dendroica pinus*) and Yellow-throated Warblers (*Dendroica dominica*) from three geographically isolated areas in the USA.



[Source: R Ficken et al. 1968. *Evolution*. Vol 27. Pp 307–314. Republished with the permission of Wiley-Blackwell.]

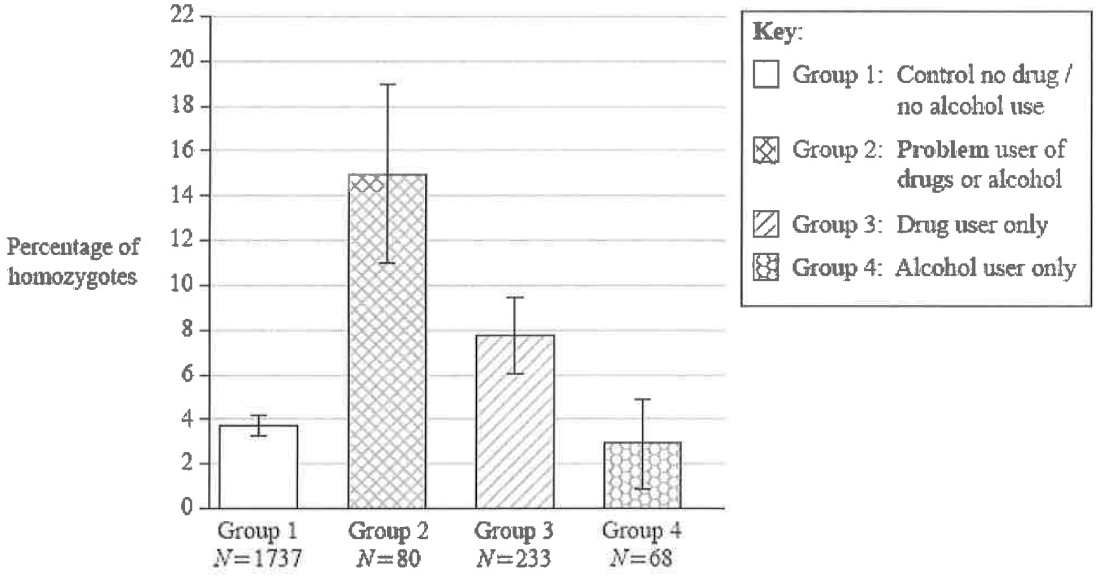
2. The effect of social stimulation on the reproductive patterns of egg-laying female Zebra finches (*Taeniopygia guttata*) was studied. The sounds of the same colony (intra-colony) and of a different colony (extra-colony) were recorded and played to different pairs of Zebra finches.

Graph A shows the mean time until the laying of the first eggs. Graph B shows the mean clutch size (total number of eggs per pair). The control pairs had no recordings played to them.



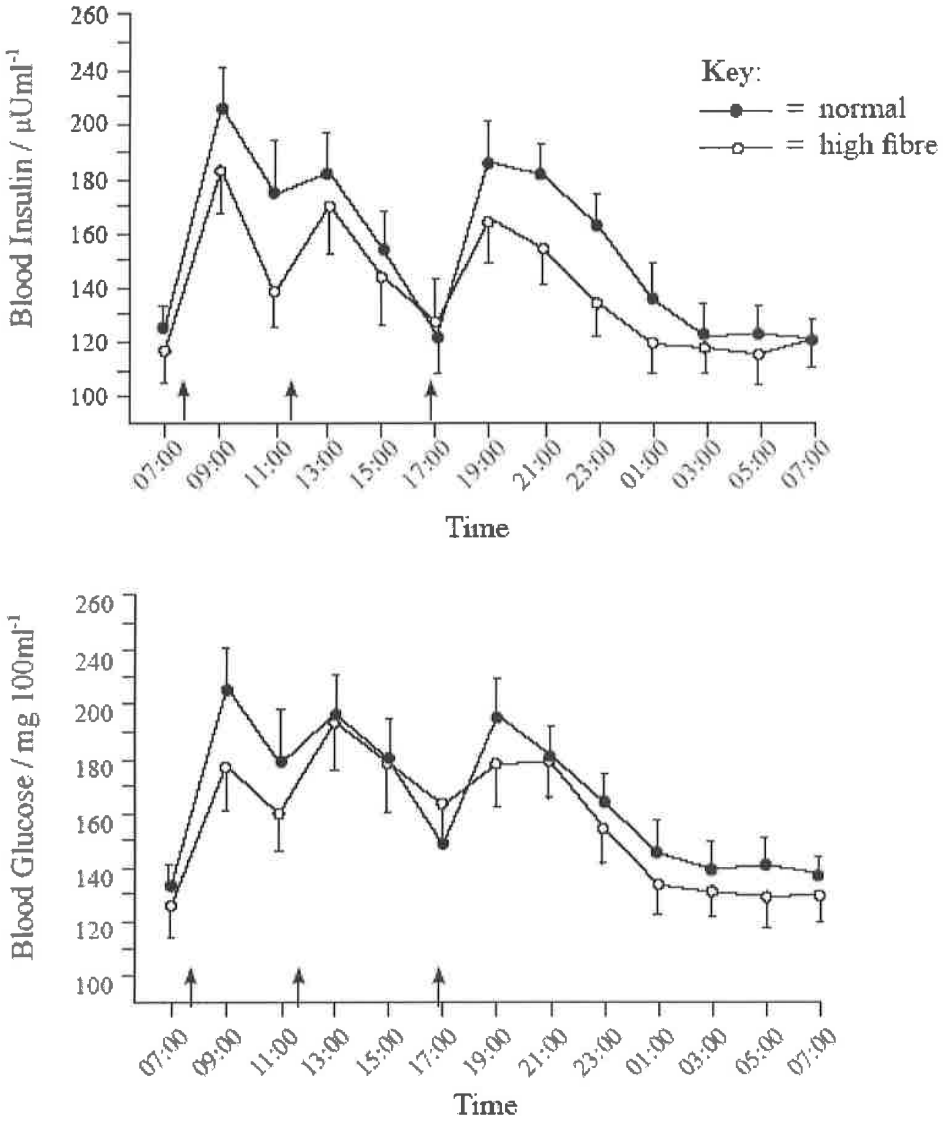
[Source: J Waas et al. 2005. *Proceedings of the Royal Society.* Vol 272. Pp 383–388. Reproduced with permission.]

3. Drug abuse and alcohol abuse are neurobehavioural disorders of complex origin. A human gene has been identified that encodes the main enzyme (FAAH) for inactivating cannabinoid (THC). A mutation in this gene can occur and the homozygous mutation allows normal catalytic activity of FAAH but makes the FAAH more likely to be broken down. A study was conducted to test for the presence of the homozygous FAAH mutation in relation to drug and alcohol abuse. Four different groups were formed based on their use of drugs and alcohol.



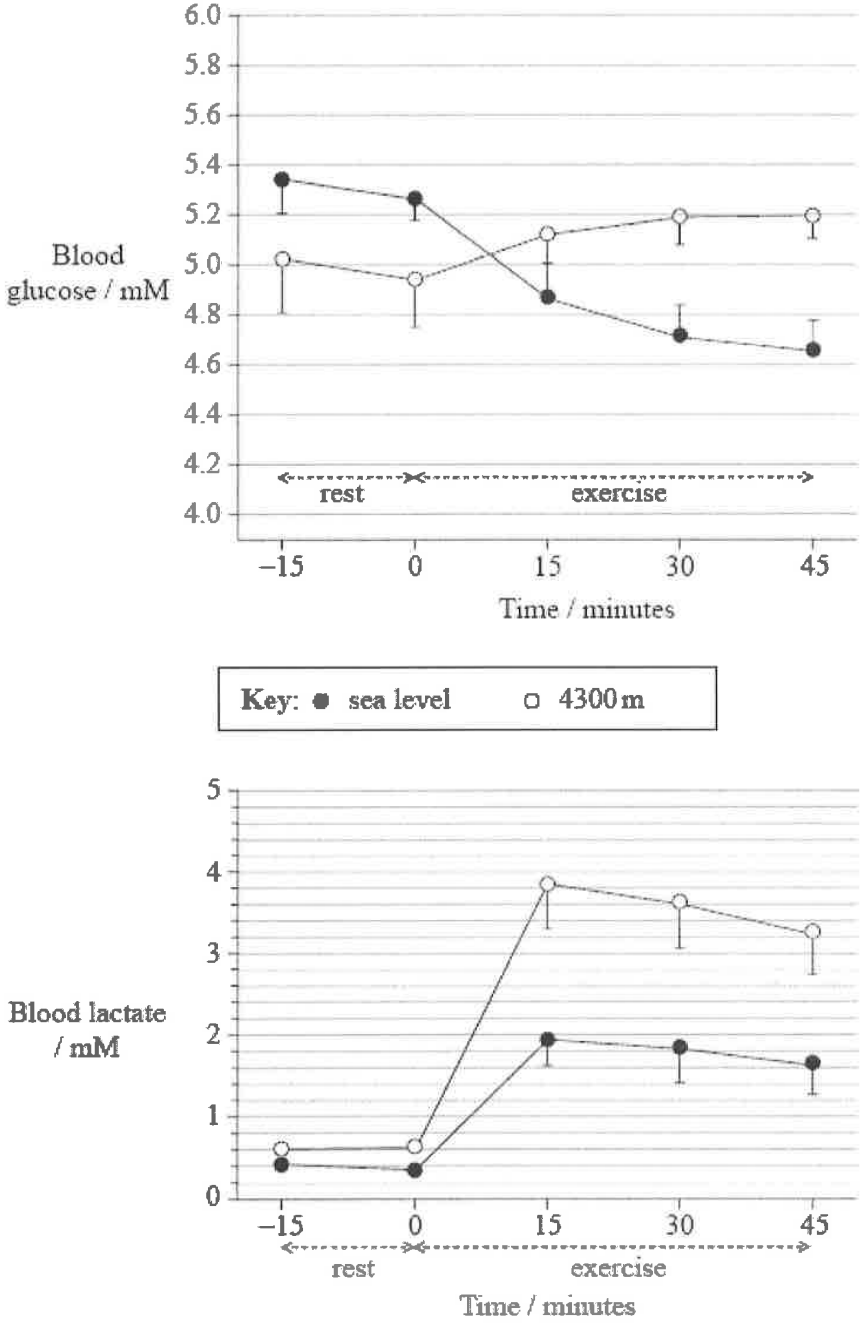
[Adapted from Sipe, J C et al. 2002. *PNAS.* Pp 8394–8399. Copyright 2002 National Academy of Sciences, USA.]

In North America normal diets are traditionally low in fibre. Scientists have proposed the hypothesis that people with diabetes benefit from a high fibre diet. The diagrams show the effect on blood glucose and blood insulin concentrations of a group of diabetic patients fed on a high fibre diet compared to a group of diabetic patients fed on a normal diet. The bars indicate standard error.



[Source: M Chandalia, *et al.*, *The New England Journal of Medicine*, (2000), **342**, pages 1392–1398]

Sixteen women were studied to evaluate blood glucose and blood lactate levels while exercising at sea level and at high altitude, 4300 m above sea level.



[Source: Barry Braun et al. 2000. "Women at altitude: carbohydrate utilization during exercise at 4,300 m". *Journal of Applied Physiology*. Vol 88, issue 1. Pp 246–256 (Figure 6). © American Physiological Society. Reproduced with permission.]