HANDBOOK OF BIOLOGICAL STATISTICS

SECOND EDITION

JOHN H. MCDONALD

University of Delaware

SPARKY HOUSE PUBLISHING

Baltimore, Maryland, U.S.A.

Contents

Basics

Introduction 1
Data analysis steps 4
Kinds of biological variables 7
Probability 13
Hypothesis testing 15
Random sampling 21

Tests for nominal variables

Exact binomial test 24
Power analysis 33
Chi-square test of goodness-of-fit 39
G-test of goodness-of-fit 46
Randomization test of goodness-of-fit 52
Chi-square test of independence 57
G-test of independence 64
Fisher's exact test 70
Randomization test of independence 76
Small numbers in chi-square and G-tests 80
Repeated G-tests of goodness-of-fit 84
Cochran-Mantel-Haenszel test 88

Descriptive statistics

Central tendency 95
Dispersion 102
Standard error 107
Confidence limits 112

Tests for one measurement variable

Student's t-test 118
Introduction to one-way anova 123

Model I vs. Model II anova 127
Testing homogeneity of means 130
Planned comparisons among means 137
Unplanned comparisons among means 141
Estimating added variance components 146
Normality 150
Homoscedasticity 155
Data transformations 160
Kruskal-Wallis test 165
Nested anova 173
Two-way anova 182
Paired t-test 191
Wilcoxon signed-rank test 198
Sign test 202

Tests for multiple measurement variables

Linear regression and correlation 207
Spearman rank correlation 221
Polynomial regression 224
Analysis of covariance 232
Multiple regression 239
Logistic regression 247

Multiple tests

Multiple comparisons 256
Meta-analysis 260

Miscellany

Using spreadsheets for statistics 266
Displaying results in graphs: Excel 274
Displaying results in graphs: Calc 287
Displaying results in tables 297
Introduction to SAS 300
Choosing the right test 308

Introduction

Welcome to the *Handbook of Biological Statistics*! This online textbook evolved from a set of notes for my Biological Data Analysis class at the University of Delaware. My main goal in that class is to teach biology students how to choose the appropriate statistical test for a particular experiment, then apply that test and interpret the results. I spend relatively little time on the mathematical basis of the tests; for most biologists, statistics is just a useful tool, like a microscope, and knowing the detailed mathematical basis of a statistical test is as unimportant to most biologists as knowing which kinds of glass were used to make a microscope lens. Biologists in very statistics-intensive fields, such as ecology, epidemiology, and systematics, may find this handbook to be a bit superficial for their needs, just as a microscopist using the latest techniques in 4-D, 3-photon confocal microscopy needs to know more about their microscope than someone who's just counting the hairs on a fly's back.

You may navigate through these pages using the "Previous topic" and "Next topic" links at the top of each page, or you may skip from topic to topic using the links on the left sidebar. Let me know if you find a broken link anywhere on these pages.

I have provided a spreadsheet to perform almost every statistical test. Each comes with sample data already entered; just download the program, replace the sample data with your data, and you'll have your answer. The spreadsheets were written for Excel, but they should also work using the free program Calc, part of the OpenOffice.org (http://www.openoffice.org/) suite of programs. If you're using OpenOffice.org, some of the graphs may need re-formatting, and you may need to re-set the number of decimal places for some numbers. Let me know if you have a problem using one of the spreadsheets, and I'll try to fix it.

I've also linked to a web page for each test wherever possible. I found most of these web pages using John Pezzullo's excellent list of Interactive Statistical Calculation Pages (http://StatPages.org) , which is a good place to look for information about tests that are not discussed in this handbook.

There are instructions for performing each statistical test in SAS, as well. It's not as easy to use as the spreadsheets or web pages, but if you're going to be doing a lot of advanced statistics, you're going to have to learn SAS or a similar program sooner or later.

Printed version

While this handbook is primarily designed for online use, you may find it convenient to print out some or all of the pages. If you print a page, the sidebar on the left, the banner, and the decorative pictures (cute critters, etc.) should not print. I'm not sure how well printing will work with various browsers and operating systems, so if the pages don't print properly, please let me know.

If you want a spiral-bound, printed copy of the whole handbook (313 pages), you can buy one from Lulu.com (http://www.lulu.com/content/3862228) for $16 plus shipping. I've used this print-on-demand service as a convenience to you, not as a money-making scheme, so don't feel obligated to buy one. You can also download a pdf of the entire handbook from that link and print it yourself. The pdf has page numbers and a table of contents, so it may be a little easier to use than individually printed web pages.

You may cite the printed version as:

McDonald, J.H. 2009. Handbook of Biological Statistics, 2nd ed. Sparky House Publishing, Baltimore, Maryland.

It's better to cite the print version, rather than the web pages, because I plan to extensively revise the web pages once a year or so. I'll keep the free pdf of the print version of each major revision as a separate edition on Lulu.com (http://www.lulu.com/content/386228) , so people can go back and see what you were citing at the time you wrote your paper. The page numbers of each section in the print version are given at the bottom of each web page.

I am constantly trying to improve this textbook. If you find errors or have suggestions for improvement, please e-mail me at mcdonald@udel.edu. If you have statistical questions about your research, I'll be glad to try to answer them. However, I must warn you that I'm not an expert in statistics, so if you're asking about something that goes far beyond what's in this textbook, I may not be able to help you. And please don't ask me for help with your statistics homework (unless you're in my class, of course!).

Further reading

There are lots of statistics textbooks, but most are too elementary to use as a serious reference, too math-obsessed, or not biological enough. The two books I use the most, and see cited most often in the biological literature, are Sokal and Rohlf (1995) and Zar (1999). They cover most of the same topics, at a similar level, and either would serve you well when you want more detail than I provide in this handbook. I've provided references to the appropriate pages in both books on most of these web pages.

There are a number of online statistics manuals linked at StatPages.org. If you're interested in business statistics, time-series analysis, or other topics that I don't cover here, that's an excellent place to start. Wikipedia has some good articles on statistical topics, while others are either short and sketchy, or overly technical.

Sokal, R.R., and F.J. Rohlf. 1995. Biometry: The principles and practice of statistics in biological research. 3rd edition. W.H. Freeman, New York.

Zar, J.H. 1999. Biostatistical analysis. 4th edition. Prentice Hall, Upper Saddle River, NJ.

Acknowledgment

Preparation of this handbook has been supported in part by a grant to the University of Delaware from the Howard Hughes Medical Institute Undergraduate Science Education Program.

Thanks!

Step-by-step analysis of biological data

I find that a systematic, step-by-step approach is the best way to analyze biological data. The statistical analysis of a biological experiment may be broken down into the following steps:

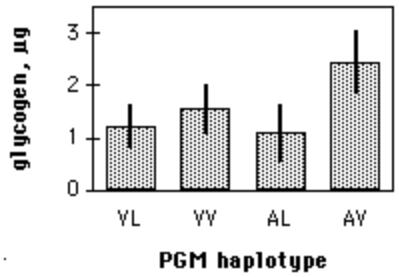
1. Specify the biological question to be answered.
2. Put the question in the form of a biological null hypothesis and alternate hypothesis.
3. Put the question in the form of a statistical null hypothesis and alternate hypothesis.
4. Determine which variables are relevant to the question.
5. Determine what kind of variable each one is.
6. Design an experiment that controls or randomizes the confounding variables.
7. Based on the number of variables, the kind of variables, the expected fit to the parametric assumptions, and the hypothesis to be tested, choose the best statistical test to use.
8. If possible, do a power analysis to determine a good sample size for the experiment.
9. Do the experiment.
10. Examine the data to see if it meets the assumptions of the statistical test you chose (normality, homoscedasticity, etc.). If it doesn't, choose a more appropriate test.
11. Apply the chosen statistical test, and interpret the result.
12. Communicate your results effectively, usually with a graph or table.

Here's an example of how this works. Verrelli and Eanes (2001) measured glycogen content in *Drosophila melanogaster* individuals. The flies were polymorphic at the genetic locus that codes for the enzyme phosphoglucomutase (PGM). At site 52 in the PGM protein sequence, flies had either a valine or an alanine. At site 484, they had either a valine or a leucine. All four combinations of amino acids (V-V, V-L, A-V, A-L) were present.

1. One biological question is "Do the amino acid polymorphisms at the *Pgm* locus have an effect on glycogen content?" The biological question is usually something about biological processes, usually in the form "Does X cause Y?"
2. The biological null hypothesis is "Different amino acid sequences do not affect the biochemical properties of PGM, so glycogen content is not affected by PGM sequence." The biological alternative hypothesis is "Different amino acid sequences do affect the biochemical properties of PGM, so glycogen content is affected by PGM sequence."
3. The statistical null hypothesis is "Flies with different sequences of the PGM enzyme have the same average glycogen content." The alternate hypothesis is "Flies with different sequences of PGM have different average glycogen contents." While the biological null and alternative hypotheses are about biological processes, the statistical null and alternative hypotheses are all about the numbers; in this case, the glycogen contents are either the same or different.
4. The two relevant variables are glycogen content and PGM sequence.
5. Glycogen content is a measurement variable, something that is recorded as a number that could have many possible values. The sequence of PGM that a fly has (V-V, V-L, A-V or A-L) is a nominal variable, something with a small number of possible values (four, in this case) that is usually recorded as a word.
6. Other variables that might be important, such as age and where in a vial the fly pupated, were either controlled (flies of all the same age were used) or randomized (flies were taken randomly from the vials without regard to where they pupated).
7. Because the goal is to compare the means of one measurement variable among groups classified by one nominal variable, and there are more than two classes, the appropriate statistical test is a Model I one-way anova.
8. A power analysis would have required an estimate of the standard deviation of glycogen content, which probably could have been found in the published literature, and a number for the effect size (the variation in glycogen content among genotypes that the experimenters wanted to detect). In this experiment, any difference in glycogen content among genotypes would be interesting, so the experimenters just used as many flies as was practical in the time available.
9. The experiment was done: glycogen content was measured in flies with different PGM sequences.
10. The anova assumes that the measurement variable, glycogen content, is normal (the distribution fits the bell-shaped normal curve) and homoscedastic (the variances in glycogen content of the different PGM sequences are equal), and inspecting histograms of the data shows that the

data fit these assumptions. If the data hadn't met the assumptions of anova, the Kruskal–Wallis test or Welch's test might have been better.

11. The one-way anova was done, using a spreadsheet, web page, or computer program, and the result of the anova is a P-value less than 0.05. The interpretation is that flies with some PGM sequences have different average glycogen content than flies with other sequences of PGM.
12. The results could be summarized in a table, but a more effective way to communicate them is with a graph:



Glycogen content in *Drosophila melanogaster*. Each bar represents the mean glycogen content (in micrograms per fly) of 12 flies with the indicated PGM haplotype. Narrow bars represent +/-2 standard errors of the mean.

Reference

Verrelli, B.C., and W.F. Eanes. 2001. The functional impact of PGM amino acid polymorphism on glycogen content in *Drosophila melanogaster*. Genetics 159: 201-210. (Note that for the purposes of this handbook, I've used a different statistical test than Verrelli and Eanes did. They were interested in interactions among the individual amino acid polymorphisms, so they used a two-way anova.)

Types of variables

One of the first steps in deciding which statistical test to use is determining what kinds of variables you have. When you know what the relevant variables are, what kind of variables they are, and what your null and alternative hypotheses are, it's usually pretty easy to figure out which test you should use. For our purposes, it's important to classify variables into three types: measurement variables, nominal variables, and ranked variables.

Similar experiments, with similar null and alternative hypotheses, will be analyzed completely differently depending on which of these three variable types are involved. For example, let's say you've measured *variable X* in a sample of 56 male and 67 female isopods (*Armadillidium vulgare*, commonly known as pillbugs or roly-polies), and your null hypothesis is "Male and female *A. vulgare* have the same values of variable X." If variable X is width of the head in millimeters, it's a measurement variable, and you'd analyze it with a t-test or a Model I one-way analysis of variance (anova). If variable X is a genotype (such as *AA*, *Aa*, or *aa*), it's a nominal variable, and you'd compare the genotype frequencies with a Fisher's exact test, chi-square test or G-test of independence. If you shake the isopods until they roll up into little balls, then record which is the first isopod to unroll, the second to unroll, etc., it's a ranked variable and you'd analyze it with a Kruskal–Wallis test.

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.

The mathematical theories underlying statistical tests involving measurement variables assume that they could have an infinite number of possible values. In practice, the number of possible values of a measurement variable is limited by the precision of the measuring device. For example, if you measure isopod head widths using an ocular micrometer that has a precision of 0.01 mm, the possible values for adult isopods whose heads range from 3 to 5 mm wide would be 3.00, 3.01, 3.02, 3.03... 5.00 mm, or only 201 different values. As long as there are a large number of possible values of the variable, it doesn't matter that there aren't really

an infinite number. However, if the number of possible values of a variable is small, this violation of the assumption could be important. For example, if you measured isopod heads using a ruler with a precision of 1 mm, the possible values could be 3, 4 or 5 mm, and it might not be a good idea to use the statistical tests designed for continuous measurement variables on this data set.

Variables that require counting a number of objects, such as the number of bacteria colonies on a plate or the number of vertebrae on an eel, are known as meristic variables. They are considered measurement variables and are analyzed with the same statistics as continuous measurement variables. Be careful, however; when you count something, it is sometimes a nominal variable. For example, the number of bacteria colonies on a plate is a measurement variable; you count the number of colonies, and there are 87 colonies on one plate, 92 on another plate, etc. Each plate would have one data point, the number of colonies; that's a number, so it's a measurement variable. However, if the plate has red and white bacteria colonies and you count the number of each, it is a nominal variable. Each colony is a separate data point with one of two values of the variable, "red" or "white"; because that's a word, not a number, it's a nominal variable. In this case, you might summarize the nominal data with a number (the percentage of colonies that are red), but the underlying data are still nominal.

Something that could be measured is a measurement variable, even when the values are controlled by the experimenter. For example, if you grow bacteria on one plate with medium containing 10 mM mannose, another plate with 20 mM mannose, etc. up to 100 mM mannose, the different mannose concentrations are a measurement variable, even though you made the media and set the mannose concentration yourself.

Nominal variables

These variables, also called "attribute variables" or "categorical variables," classify observations into a small number of categories. A good rule of thumb is that an individual observation of a nominal variable is usually a word, not a number. Examples of nominal variables include sex (the possible values are male or female), genotype (values are *AA*, *Aa*, or *aa*), or ankle condition (values are normal, sprained, torn ligament, or broken). Nominal variables are often used to divide individuals up into classes, so that other variables may be compared among the classes. 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 I count the number of male and female *A. vulgare* in a sample from Newark and a sample from Baltimore, I might say that 52.3 percent of the isopods in Newark and 62.1 percent of the isopods in Baltimore are female. These percentages may look like a measurement variable, but they really represent a

nominal variable, sex. I determined the value of the nominal variable (male or female) on 65 isopods from Newark, of which 34 were female and 31 were male. I might plot 52.3 percent on a graph as a simple way of summarizing the data, but I would 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."

It is possible to convert a measurement variable to a nominal variable, dividing individuals up into a small number of classes based on ranges of the variable. For example, if you are studying levels of HDL (the "good cholesterol") and blood pressure, you could measure the HDL level, then divide people into two groups, "low HDL" (less than 40 mg/dl) and "normal HDL" (40 or more mg/dl) and compare the mean blood pressures of the two groups, using a nice simple t-test.

Converting measurement variables to nominal variables ("categorizing") is common in epidemiology and some other fields. It is a way of avoiding some statistical problems when constructing complicated regression models involving lots of variables. I think it's better for most biological experiments if you don't do this. One problem with categorizing measurement variables is that you'd be discarding a lot of information; in our example, you'd be lumping together everyone with HDL from 0 to 39 mg/dl into one group, which could decrease your chances of finding a relationship between the two variables if there really is one. Another problem is that it would be easy to consciously or subconsciously choose the dividing line between low and normal HDL that gave an "interesting" result. For example, if you did the experiment thinking that low HDL caused high blood pressure, and a couple of people with HDL between 40 and 45 happened to have high blood pressure, you might put the dividing line between low and normal at 45 mg/dl. This would be cheating, because it would increase the chance of getting a "significant" difference if there really isn't one. If you are going to categorize variables, you should decide on the categories by some objective means; either use categories that other people have used previously, or have some predetermined rule such as dividing the observations into equally-sized groups.

Ranked variables

Ranked variables, also called ordinal variables, are those for which the individual observations can be put in order from smallest to largest, even though the exact values are unknown. If you shake a bunch of *A. vulgare* up, they roll into balls, then after a little while start to unroll and walk around. If you wanted to

know whether males and females unrolled at the same average time, you could pick up the first isopod to unroll and put it in a vial marked "first," pick up the second to unroll and put it in a vial marked "second," and so on, then sex the isopods after they've all unrolled. You wouldn't have the exact time that each isopod stayed rolled up (that would be a measurement variable), but you would have the isopods in order from first to unroll to last to unroll, which is a ranked variable. While a nominal variable is recorded as a word (such as "male") and a measurement variable is recorded as a number (such as "4.53"), a ranked variable can be recorded as a rank (such as "seventh").

You could do a lifetime of biology and never use a true ranked variable. The reason they're important is that the statistical tests designed for ranked variables (called "non-parametric tests," for reasons you'll learn later) make fewer assumptions about the data than the statistical tests designed for measurement variables. Thus the most common use of ranked variables involves converting a measurement variable to ranks, then analyzing it using a non-parametric test. For example, let's say you recorded the time that each isopod stayed rolled up, and that most of them unrolled after one or two minutes. Two isopods, who happened to be male, stayed rolled up for 30 minutes. If you analyzed the data using a test designed for a measurement variable, those two sleepy isopods would cause the average time for males to be much greater than for females, and the difference might look statistically significant. When converted to ranks and analyzed using a non-parametric test, the last and next-to-last isopods would have much less influence on the overall result, and you would be less likely to get a misleadingly "significant" result if there really isn't a difference between males and females.

Some variables are impossible to measure objectively with instruments, so people are asked to give a subjective rating. For example, pain is often measured by asking a person to put a mark on a 10-cm scale, where 0 cm is "no pain" and 10 cm is "worst possible pain." This is a measurement variable, even though the "measuring" is done by the person's brain. For the purpose of statistics, the important thing is that it is measured on an "interval scale"; ideally, the difference between pain rated 2 and 3 is the same as the difference between pain rated 7 and 8. Pain would be a ranked variable if the pains at different times were compared with each other; for example, if someone kept a pain diary and then at the end of the week said "Tuesday was the worst pain, Thursday was second worst, Wednesday was third, etc...." These rankings are not an interval scale; the difference between Tuesday and Thursday may be much bigger, or much smaller, than the difference between Thursday and Wednesday.

Circular variables

A special kind of measurement variable is a circular variable. These have the property that the highest value and the lowest value are right next to each other; often, the zero point is completely arbitrary. The most common circular variables

in biology are time of day, time of year, and compass direction. If you measure time of year in days, Day 1 could be January 1, or the spring equinox, or your birthday; whichever day you pick, Day 1 is adjacent to Day 2 on one side and Day 365 on the other.

If you are only considering part of the circle, a circular variable becomes a regular measurement variable. For example, if you're doing a regression of the number of geese in a corn field vs. time of year, you might treat Day 1 to be March 28, the day you planted the corn; the fact that the year circles around to March 27 would be irrelevant, since you would chop the corn down in September.

If your variable really is circular, there are special, very obscure statistical tests designed just for circular data; see chapters 26 and 27 in Zar.

Ambiguous variables

When you have a measurement variable with a small number of values, it may not be clear whether it should be considered a measurement or a nominal variable. For example, if you compare bacterial growth in two media, one with 0 mM mannose and one with 20 mM mannose, and you have several measurements of bacterial growth at each concentration, you should consider mannose to be a nominal variable (with the values "mannose absent" or "mannose present") and analyze the data using a t-test or a one-way anova. If there are 10 different mannose concentrations, you should consider mannose concentration to be a measurement variable and analyze the data using linear regression (or perhaps polynomial regression).

But what if you have three concentrations of mannose, or five, or seven? There is no rigid rule, and how you treat the variable will depend in part on your null and alternative hypotheses. If your alternative hypothesis is "different values of mannose have different rates of bacterial growth," you could treat mannose concentration as a nominal variable. Even if there's some weird pattern of high growth on zero mannose, low growth on small amounts, high growth on intermediate amounts, and low growth on high amounts of mannose, the one-way anova can give a significant result. If your alternative hypothesis is "bacteria grow faster with more mannose," it would be better to treat mannose concentration as a measurement variable, so you can do a regression. In my class, we use the following rule of thumb:

—a measurement variable with only two values should be treated as a nominal variable;

—a measurement variable with six or more values should be treated as a measurement variable;

—a measurement variable with three, four or five values does not exist.

Of course, in the real world there are experiments with three, four or five values of a measurement variable. Your decision about how to treat this variable will depend in part on your biological question. You can avoid the ambiguity

when you design the experiment--if you want to know whether a dependent variable is related to an independent variable that could be measurement, it's a good idea to have at least six values of the independent variable.

The same rules apply to ranked variables. If you put 10 different bandages on a person's arm, rip them off, then have the person rank them from most painful to least painful, that is a ranked variable. You could do Spearman's rank correlation to see if the pain rank is correlated with the amount of adhesive on the bandage.If you do the same experiment with just two bandages and ask "Which hurts worse, bandage A or bandage B?", that's a nominal variable; it just has two possible values (A or B), or three if you allow ties.

Ratios

Some biological variables are ratios of two measurement variables. If the denominator in the ratio has no biological variation and a small amount of measurement error, such as heartbeats per minute or white blood cells per ml of blood, you can treat the ratio as a regular measurement variable. However, if both numerator and denominator in the ratio have biological variation, it is better, if possible, to use a statistical test that keeps the two variables separate. For example, if you want to know whether male isopods have relatively bigger heads than female isopods, you might want to divide head width by body length and compare this head/body ratio in males vs. females, using a t-test or a one-way anova. This wouldn't be terribly wrong, but it could be better to keep the variables separate and compare the regression line of head width on body length in males to that in females using an analysis of covariance.

Sometimes treating two measurement variables separately makes the statistical test a lot more complicated. In that case, you might want to use the ratio and sacrifice a little statistical rigor in the interest of comprehensibility. For example, if you wanted to know whether there was a relationship between obesity and high-density lipoprotein (HDL) levels in blood, you could do multiple regression with height and weight as the two X variables and HDL level as the Y variable. However, multiple regression is a complicated, advanced statistical technique, and if you found a significant relationship, it could be difficult to explain to your fellow biologists and very difficult to explain to members of the public who are concerned about their HDL levels. In this case it might be better to calculate the body mass index (BMI), the ratio of weight over squared height, and do a simple linear regression of HDL level and BMI.

Further reading

Sokal and Rohlf, pp. 10-13.

Zar, pp. 2-5 (measurement, nominal and ranked variables); pp. 592-595 (circular variables).

Probability

The basic idea of a statistical test is to identify a null hypothesis, collect some data, then estimate the probability of getting the observed data if the null hypothesis were true. If the probability of getting a result like the observed one is low under the null hypothesis, you conclude that the null hypothesis is probably not true. It is therefore useful to know a little about probability.

One way to think about probability is as the proportion of individuals in a population that have a particular characteristic. (In this case, both "individual" and "population" have somewhat different meanings than they do in biology.) The probability of sampling a particular kind of individual is equal to the proportion of that kind of individual in the population. For example, in fall 2003 there were 21,121 students at the University of Delaware, and 16,428 of them were undergraduates. If a single student were sampled at random, the probability that they would be an undergrad would be 16,428 / 21,121, or 0.778. In other words, 77.8% of students were undergrads, so if you'd picked one student at random, the probability that they were an undergrad would have been 77.8%.

When dealing with probabilities in biology, you are often working with theoretical expectations, not population samples. For example, in a genetic cross of two individual *Drosophila melanogaster* that are heterozygous at the *white* locus, Mendel's theory predicts that the probability of an offspring individual being a recessive homozygote (having white eyes) is one-fourth, or 0.25. This is equivalent to saying that one-fourth of a population of offspring will have white eyes.

Multiplying probabilities

You could take a semester-long course on mathematical probability, but most biologists just need a few basic principles. The probability that an individual has one value of a nominal variable AND another value is estimated by multiplying the probabilities of each value together. For example, if the probability that a *Drosophila* in a cross has white eyes is one-fourth, and the probability that it has legs where its antennae should be is three-fourths, the probability that it has white eyes AND leg-antennae is one-fourth times three-fourths, or 0.25 X 0.75, or 0.1875. This estimate assumes that the two values are independent, meaning that the probability of one value is not affected by the other value. In this case,

independence would require that the two genetic loci were on different chromosomes, among other things.

Adding probabilities

The probability that an individual has one value OR another, MUTUALLY EXCLUSIVE, value is found by adding the probabilities of each value together. "Mutually exclusive" means that one individual could not have both values. For example, if the probability that a flower in a genetic cross is red is one-fourth, the probability that it is pink is one-half, and the probability that it is white is one-fourth, then the probability that it is red OR pink is one-fourth plus one-half, or three-fourths.

More complicated situations

When calculating the probability that an individual has one value OR another, and the two values are NOT MUTUALLY EXCLUSIVE, it is important to break things down into combinations that are mutually exclusive. For example, let's say you wanted to estimate the probability that a fly from the cross above had white eyes OR leg-antennae. You could calculate the probability for each of the four kinds of flies: red eyes/normal antennae (0.75 X 0.25 = 0.1875), red eyes/leg-antennae (0.75 X 0.75 = 0.5625), white eyes/normal antennae (0.25 X 0.25 = 0.0625), and white eyes/leg-antennae (0.25 X 0.75 = 0.1875). Then, since the last three kinds of flies are the ones with white eyes or leg-antennae, you'd add those probabilities up (0.5625 + 0.0625 + 0.1875 = 0.8125).

When to calculate probabilities

While there are some kind of probability calculations underlying all statistical tests, it is rare that you'll have to use the rules listed above. About the only time you'll actually calculate probabilities by adding and multiplying is when figuring out the expected values for a goodness-of-fit test.

Further reading

Sokal and Rohlf, pp. 62-71.

Zar, pp. 48-63.

Basic concepts of hypothesis testing

Null hypothesis

The null hypothesis is a statement that you want to test. In general, the null hypothesis is that things are the same as each other, or the same as a theoretical expectation. For example, if you measure the size of the feet of male and female chickens, the null hypothesis could be that the average foot size in male chickens is the same as the average foot size in female chickens. If you count the number of male and female chickens born to a set of hens, the null hypothesis could be that the ratio of males to females is equal to the theoretical expectation of a 1:1 ratio.

The alternative hypothesis is that things are different from each other, or different from a theoretical expectation. For example, one alternative hypothesis would be that male chickens have a different average foot size than female chickens; another would be that the sex ratio is different from 1:1.

Usually, the null hypothesis is boring and the alternative hypothesis is interesting. Finding that male chickens have bigger feet than female chickens might lead to all kinds of exciting discoveries about developmental biology, endocrine physiology, or sexual selection in chickens. Finding that male and female chickens have the same size feet wouldn't lead to anything except a boring paper in the world's most obscure chicken journal. It's therefore tempting to look for patterns in your data that support the exciting alternative hypothesis. For example, you might measure the feet of 10 male chickens and 10 female chickens and find that the mean is 0.1 mm longer for males. You're almost certain to get some difference in the means, just due to chance, so before you get all happy and start buying formal wear for the Nobel Prize ceremony, you need to ask "What's the probability of getting a difference in the means of 0.1 mm, just by chance, if the boring null hypothesis is really true?" Only when that probability is low can you reject the null hypothesis. The goal of statistical hypothesis testing is to estimate the probability of getting your observed results under the null hypothesis.

Biological vs. statistical null hypotheses

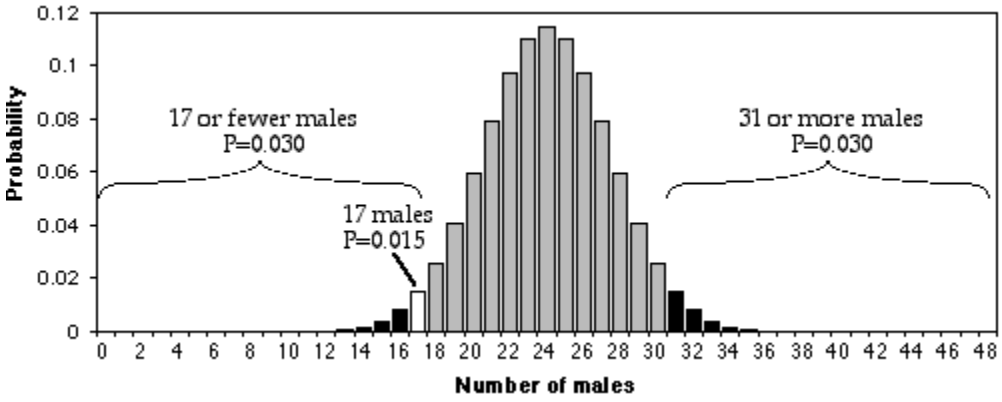
It is important to distinguish between *biological* null and alternative hypotheses and *statistical* null and alternative hypotheses. "Sexual selection by females has caused male chickens to evolve bigger feet than females" is a biological alternative hypothesis; it says something about biological processes, in this case sexual selection. "Male chickens have a different average foot size than females" is a statistical alternative hypothesis; it says something about the numbers, but nothing about what caused those numbers to be different. The biological null and alternative hypotheses are the first that you should think of, as they describe something interesting about biology; they are two possible answers to the biological question you are interested in ("What affects foot size in chickens?"). The statistical null and alternative hypotheses are statements about the data that should follow from the biological hypotheses: if sexual selection favors bigger feet in male chickens (a biological hypothesis), then the average foot size in male chickens should be larger than the average in females (a statistical hypothesis). If you reject the statistical null hypothesis, you then have to decide whether that's enough evidence that you can reject your biological null hypothesis. For example, if you don't find a significant difference in foot size between male and female chickens, you could conclude "There is no significant evidence that sexual selection has caused male chickens to have bigger feet." If you do find a statistically significant difference in foot size, that might not be enough for you to conclude that sexual selection caused the bigger feet; it might be that males eat more, or that the bigger feet are a developmental byproduct of the roosters' combs, or that males run around more and the exercise makes their feet bigger. When there are multiple biological interpretations of a statistical result, you need to think of additional experiments to test the different possibilities.

Testing the null hypothesis

The primary goal of a statistical test is to determine whether an observed data set is so different from what you would expect under the null hypothesis that you should reject the null hypothesis. For example, let's say you've given up on chicken feet and now are studying sex determination in chickens. For breeds of chickens that are bred to lay lots of eggs, female chicks are more valuable than male chicks, so if you could figure out a way to manipulate the sex ratio, you could make a lot of chicken farmers very happy. You've tested a treatment, and you get 25 female chicks and 23 male chicks. Anyone would look at those numbers and see that they could easily result from chance; there would be no reason to reject the null hypothesis of a 1:1 ratio of females to males. If you tried a different treatment and got 47 females and 1 male, most people would look at those numbers and see that they would be extremely unlikely to happen due to luck, if the null hypothesis were true; you would reject the null hypothesis and conclude that your treatment

really changed the sex ratio. However, what if you had 31 females and 17 males? That's definitely more females than males, but is it really so unlikely to occur due to chance that you can reject the null hypothesis? To answer that, you need more than common sense, you need to calculate the probability of getting a deviation that large due to chance.

P-values



Probability of getting different numbers of males out of 48, if the parametric proportion of males is 0.5.

In the figure above, the BINOMDIST function of Excel was used to calculate the probability of getting each possible number of males, from 0 to 48, under the null hypothesis that 0.5 are male. As you can see, the probability of getting 17 males out of 48 total chickens is about 0.015. That seems like a pretty small probability, doesn't it? However, that's the probability of getting *exactly* 17 males. What you want to know is the probability of getting 17 *or fewer* males. If you were going to accept 17 males as evidence that the sex ratio was biased, you would also have accepted 16, or 15, or 14,… males as evidence for a biased sex ratio. You therefore need to add together the probabilities of all these outcomes. The probability of getting 17 or fewer males out of 48, under the null hypothesis, is 0.030. That means that if you had an infinite number of chickens, half males and half females, and you took a bunch of random samples of 48 chickens, 3.0% of the samples would have 17 or fewer males.

This number, 0.030, is the P-value. It is defined as the probability of getting the observed result, or a more extreme result, if the null hypothesis is true. So "P=0.030" is a shorthand way of saying "The probability of getting 17 or fewer male chickens out of 48 total chickens, *IF* the null hypothesis is true that 50 percent of chickens are male, is 0.030."

Significance levels

Does a probability of 0.030 mean that you should reject the null hypothesis, and conclude that your treatment really caused a change in the sex ratio? The convention in most biological research is to use a significance level of 0.05. This means that if the probability value (P) is less than 0.05, you reject the null hypothesis; if P is greater than or equal to 0.05, you don't reject the null hypothesis. There is nothing mathematically magic about 0.05; people could have agreed upon 0.04, or 0.025, or 0.071 as the conventional significance level.

The significance level you use depends on the costs of different kinds of errors. With a significance level of 0.05, you have a 5 percent chance of rejecting the null hypothesis, even if it is true. If you try 100 treatments on your chickens, and none of them really work, 5 percent of your experiments will give you data that are significantly different from a 1:1 sex ratio, just by chance. This is called a "Type I error," or "false positive." If there really is a deviation from the null hypothesis, and you fail to reject it, that is called a "Type II error," or "false negative." If you use a higher significance level than the conventional 0.05, such as 0.10, you will increase your chance of a false positive to 0.10 (therefore increasing your chance of an embarrassingly wrong conclusion), but you will also decrease your chance of a false negative (increasing your chance of detecting a subtle effect). If you use a lower significance level than the conventional 0.05, such as 0.01, you decrease your chance of an embarrassing false positive, but you also make it less likely that you'll detect a real deviation from the null hypothesis if there is one.

You must choose your significance level before you collect the data, of course. If you choose to use a different signifigance level than the conventional 0.05, be prepared for some skepticism; you must be able to justify your choice. If you were screening a bunch of potential sex-ratio-changing treatments, the cost of a false positive would be the cost of a few additional tests, which would show that your initial results were a false positive. The cost of a false negative, however, would be that you would miss out on a tremendously valuable discovery. You might therefore set your significance value to 0.10 or more. On the other hand, once your sex-ratio-changing treatment is undergoing final trials before being sold to farmers, you'd want to be very confident that it really worked, not that you were just getting a false positive. Otherwise, if you sell the chicken farmers a sex-ratio treatment that turns out to not really work (it was a false positive), they'll sue the pants off of you. Therefore, you might want to set your significance level to 0.01, or even lower. **Throughout this handbook, I will always use P<0.05 as the significance level.**

One-tailed vs. two-tailed probabilities

The probability that was calculated above, 0.030, is the probability of getting 17 or fewer males out of 48. It would be significant, using the conventional P<0.05

criterion. However, what about the probability of getting 17 or fewer females? If your null hypothesis is "The proportion of males is 0.5 or more" and your alternative hypothesis is "The proportion of males is less than 0.5," then you would use the P=0.03 value found by adding the probabilities of getting 17 or fewer males. This is called a one-tailed probability, because you are adding the probabilities in only one tail of the distribution shown in the figure. However, if your null hypothesis is "The proportion of males is 0.5", then your alternative hypothesis is "The proportion of males is different from 0.5." In that case, you should add the probability of getting 17 or fewer females to the probability of getting 17 or fewer males. This is called a two-tailed probability. If you do that with the chicken result, you get P=0.06, which is not quite significant.

You should decide whether to use the one-tailed or two-tailed probability before you collect your data, of course. A one-tailed probability is more powerful, in the sense of having a lower chance of false negatives, but you should only use a one-tailed probability if you really, truly have a firm prediction about which direction of deviation you would consider interesting. In the chicken example, you might be tempted to use a one-tailed probability, because you're only looking for treatments that decrease the proportion of worthless male chickens. But if you accidentally found a treatment that produced 87 percent male chickens, would you really publish the result as "The treatment did not cause a significant decrease in the proportion of male chickens"? Probably not. You'd realize that this unexpected result, even though it wasn't what you and your farmer friends wanted, would be very interesting to other people; by leading to discoveries about the fundamental biology of sex-determination in chickens, in might even help you produce more female chickens someday. Any time a deviation in either direction would be interesting, you should use the two-tailed probability. In addition, people are skeptical of one-tailed probabilities, especially if a one-tailed probability is significant and a two-tailed probability would not be significant (as in the chicken example). Unless you provide a very convincing explanation, people may think you decided to use the one-tailed probability *after* you saw that the two-tailed probability wasn't quite significant. It may be easier to always use two-tailed probabilities. **For this handbook, I will always use two-tailed probabilities, unless I make it very clear that only one direction of deviation from the null hypothesis would be interesting.**

Reporting your results

In the olden days, when people looked up P-values in printed tables, they would report the results of a statistical test as "P<0.05", "P<0.01", "P>0.10", etc. Nowadays, almost all computer statistics programs give the exact P value resulting from a statistical test, such as P=0.029, and that's what you should report in your publications. You will conclude that the results are either significant or they're not significant; they either reject the null hypothesis (if P is below your pre-determined

significance level) or don't reject the null hypothesis (if P is above your significance level). But other people will want to know if your results are "strongly" significant (P much less than 0.05), which will give them more confidence in your results than if they were "barely" significant (P=0.043, for example). In addition, other researchers will need the exact P value if they want to combine your results with others into a meta-analysis.

Further reading

Sokal and Rohlf, pp. 157-169.

Zar, pp. 79-85.

Confounding variables and random sampling

Confounding variables

Due to a variety of genetic, developmental, and environmental factors, no two organisms are exactly alike. This means that when you design an experiment to try to see whether variable X causes a difference in variable Y, you should always ask yourself, is there some variable Z that could cause an apparent relationship between X and Y?

As an example of such a confounding variable, imagine that you want to compare the amount of insect damage on leaves of American elms (which are susceptible to Dutch elm disease) and Princeton elms, a strain of American elms that is resistant to Dutch elm disease. You find 20 American elms and 20 Princeton elms, pick 50 leaves from each, and measure the area of each leaf that was eaten by insects. Imagine that you find significantly more insect damage on the Princeton elms than on the American elms (I have no idea if this is true).

It could be that the genetic difference between the types of elm directly causes the difference in the amount of insect damage. However, there are likely to be some important confounding variables. For example, many American elms are many decades old, while the Princeton strain of elms was made commercially available only recently and so any Princeton elms you find are probably only a few years old. American elms are often treated with fungicide to prevent Dutch elm disease, while this wouldn't be necessary for Princeton elms. American elms in some settings (parks, streetsides, the few remaining in forests) may receive relatively little care, while Princeton elms are expensive and are likely planted by elm fanatics who take good care of them (fertilizing, watering, pruning, etc.). It is easy to imagine that any difference in insect damage between American and Princeton elms could be caused, not by the genetic differences between the strains, but by a confounding variable: age, fungicide treatment, fertilizer, water, pruning, or something else.

Designing an experiment to eliminate differences due to confounding variables is critically important. One way is to control all possible confounding variables. For example, you could plant a bunch of American elms and a bunch of Princeton

elms, then give them all the same care (watering, fertilizing, pruning, fungicide treatment). This is possible for many variables in laboratory experiments on model organisms.

When it isn't practical to keep all the possible confounding variables constant, another solution is to statistically control for them. You could measure each confounding variable you could think of (age of the tree, height, sunlight exposure, soil chemistry, soil moisture, etc.) and use a multivariate statistical technique to separate the effects of the different variables. This is common in epidemiology, because carefully controlled experiments on humans are often impractical and sometimes unethical. However, the analysis, interpretation, and presentation of complicated multivariate analyses are not easy.

The third way to control confounding variables is to randomize them. For example, if you are planting a bunch of elm trees in a field and are carefully controlling fertilizer, water, pruning, etc., there may still be some confounding variables you haven't thought of. One side of the field might be closer to a forest and therefore be exposed to more herbivorous insects. Or parts of the field might have slightly different soil chemistry, or drier soil, or be closer to a fence that insect-eating birds like to perch on. To control for these variables, you should mix the American and Princeton elms throughout the field, rather than planting all the American elms on one side and all the Princeton elms on the other. There would still be variation among individuals in your unseen confounding variables, but because it was randomized, it would not cause a consistent difference between American and Princeton elms.

Random sampling

An important aspect of randomizing possible confounding variables is taking random samples of a population. "Population," in the statistical sense, is different from a biological population of individuals; it represents all the possible measurements of a particular variable. For example, if you are measuring the fluorescence of a pH-sensitive dye inside a kidney cell, the "population" could be the fluorescence at all possible points inside that cell. Depending on your experimental design, the population could also be the fluorescence at all points inside all of the cells of one kidney, or even the fluorescence at all points inside all of the cells of all of the kidneys of that species of animal.

A random sample is one in which all members of a population have an equal probability of being sampled. If you're measuring fluorescence inside kidney cells, this means that all points inside a cell, and all the cells in a kidney, and all the kidneys in all the individuals of a species, would have an equal chance of being sampled.

A perfectly random sample of observations is difficult to collect, and you need to think about how this might affect your results. Let's say you've used a confocal microscope to take a two-dimensional "optical slice" of a kidney cell. It would be

easy to use a random-number generator on a computer to pick out some random pixels in the image, and you could then use the fluorescence in those pixels as your sample. However, if your slice was near the cell membrane, your "random" sample would not include any points deep inside the cell. If your slice was right through the middle of the cell, however, points deep inside the cell would be over-represented in your sample. You might get a fancier microscope, so you could look at a random sample of the "voxels" (three-dimensional pixels) throughout the volume of the cell. But what would you do about voxels right at the surface of the cell? Including them in your sample would be a mistake, because they might include some of the cell membrane and extracellular space, but excluding them would mean that points near the cell membrane are under-represented in your sample.

As another example, let's say you want to estimate the amount of physical activity the average University of Delaware undergrad gets. You plan to attach pedometers to 50 students and count how many steps each student takes during a week. If you stand on a sidewalk and recruit students, one confounding variable would be where the sidewalk is. If it's on North College Avenue, the primary route between the main campus and the remote Christiana Towers dorms, your sample will include students who do more walking than students who live closer to campus. Recruiting volunteers on a sidewalk near a student parking lot, a bus stop, or the student health center could get you more sedentary students. It would be better to pick students at random from the student directory and ask them to volunteer for your study. However, motivation to participate would be a difficult confounding variable to randomize; I'll bet that particularly active students who were proud of their excellent physical condition would be more likely to volunteer for your study than would students who spend all their time looking at great musicians on MySpace and searching YouTube for videos of cats. To get a truly random sample, you'd like to be able to make everyone you chose randomly participate in your study, but they're people, so you can't. Designing a perfectly controlled experiment involving people can be very difficult. Maybe you could put pedometers on cats, instead--that would be pretty funny looking.

Exact test for goodness-of-fit

The main goal of a statistical test is to answer the question, "What is the probability of getting a result like my observed data, if the null hypothesis were true?" If it is very unlikely to get the observed data under the null hypothesis, you reject the null hypothesis.

Most statistical tests take the following form:

1. Collect the data.
2. Calculate a number, the *test statistic,* that measures how far the observed data deviate from the expectation under the null hypothesis.
3. Use a mathematical function to estimate the probability of getting a test statistic as extreme as the one you observed, if the null hypothesis were true. This is the *P-value*.

Exact tests, such as the exact test for goodness-of-fit, are different. There is no test statistic; instead, the probability of obtaining the observed data under the null hypothesis is calculated directly. This is because the predictions of the null hypothesis are so simple that the probabilities can easily be calculated.

When to use it

You use the exact binomial test when you have one nominal variable with only two values (such as male or female, left or right, green or yellow). The observed data are compared with the expected data, which are some kind of theoretical expectation (such as a 1:1 sex ratio or a 3:1 ratio in a genetic cross) that is determined before the data are collected. If the total number of observations is too high (around a thousand), computers may not be able to do the calculations for the exact test, and a G-test or chi-square test of goodness-of-fit must be used instead (and will give almost exactly the same result).

You can do exact multinomial tests of goodness-of-fit when the nominal variable has more than two values. The basic concepts are the same as for the exact binomial test. Here I'm limiting the explanation to the binomial test, because it's more commonly used and easier to understand.

Null hypothesis

For a two-tailed test, which is what you almost always should use, the null hypothesis is that the number of observations in each category is equal to that predicted by a biological theory, and the alternative hypothesis is that the observed data are different from the expected. For example, if you do a genetic cross in which you expect a 3:1 ratio of green to yellow pea pods, and you have a total of 50 plants, your null hypothesis is that there are 37.5 plants with green pods and 12.5 with yellow pods.

If you are doing a one-tailed test, the null hypothesis is that the observed number for one category is equal to or less than the expected; the alternative hypothesis is that the observed number in that category is greater than expected.

How the test works

Let's say you want to know whether our cat, Gus, has a preference for one paw or uses both paws equally. You dangle a ribbon in his face and record which paw he uses to bat at it. You do this 10 times, and he bats at the ribbon with his right paw 8 times and his left paw 2 times. Then he gets bored with the experiment and leaves. Can you conclude that he is right-pawed, or could this result have occurred due to chance under the null hypothesis that he bats equally with each paw?

The null hypothesis is that each time Gus bats at the ribbon, the probability that he will use his right paw is 0.5. The probability that he will use his right paw on the first time is 0.5. The probability that he will use his right paw the first time AND the second time is 0.5 x 0.5, or 0.5^2, or 0.25. The probability that he will use his right paw all ten times is 0.5^{10}, or about 0.001.

For a mixture of right and left paws, the calculation is more complicated. Where n is the total number of trials, k is the number of "successes" (statistical jargon for whichever event you want to consider), p is the expected proportion of successes if the null hypothesis is true, and Y is the probability of getting k successes in n trials, the equation is:

$$Y = \frac{p^k(1-p)^{(n-k)}n!}{k!(n-k)!}$$

Fortunately, there's an spreadsheet function that does the calculation for you. To calculate the probability of getting exactly 8 out of 10 right paws, you would enter

```
=BINOMDIST(2, 10, 0.5, FALSE)
```

The first number, 2, is whichever event there are fewer than expected of; in this case, there are only two uses of the left paw, which is fewer than the expected 5.

The second number is the total number of trials. The third number is the expected proportion of whichever event there were fewer than expected of. And FALSE tells it to calculate the exact probability for that number of events only. In this case, the answer is P=0.044, so you might think it was significant at the P<0.05 level.

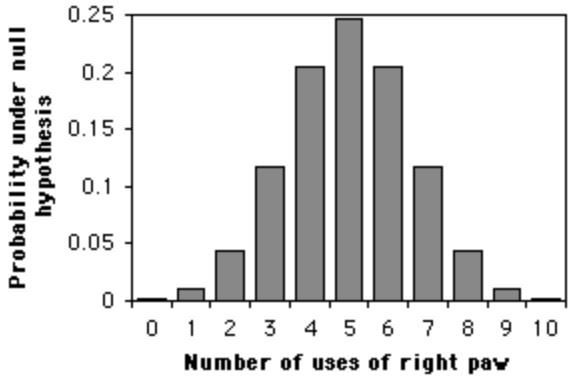
However, it would be incorrect to only calculate the probability of getting exactly 2 left paws and 8 right paws. Instead, you must calculate the probability of getting a deviation from the null expectation as large as, *or larger than,* the observed result. So you must calculate the probability that Gus used his left paw 2 times out of 10, or 1 time out of 10, or 0 times out of ten. Adding these probabilities together gives P=0.055, which is not quite significant at the P<0.05 level. You do this in a spreadsheet by entering

```
=BINOMDIST(2, 10, 0.5, TRUE).
```

The "TRUE" parameter tells the spreadsheet to calculate the sum of the probabilities of the observed number and all more extreme values; it's the equivalent of

```
=BINOMDIST(2, 10, 0.5, FALSE)+BINOMDIST(1, 10, 0.5,
      FALSE)+BINOMDIST(0, 10, 0.5, FALSE).
```

There's one more thing. The above calculation gives the total probability of getting 2, 1, or 0 uses of the left paw out of 10. However, the alternative hypothesis is that the number of uses of the right paw is not equal to the number of uses of the left paw. If there had been 2, 1, or 0 uses of the right paw, that also would have been an equally extreme deviation from the expectation. So you must add the probability of getting 2, 1, or 0 uses of the right paw, to account for both tails of the probability distribution; you are doing a two-tailed test. This gives you P=0.109, which is not very close to being significant. (If the null hypothesis had been 0.50 or more uses of the left paw, and the alternative hypothesis had been less than 0.5 uses of left paw, you could do a one-tailed test and use P=0.054. But you almost never have a situation where a one-tailed test is appropriate.)



Graph showing the probability distribution for the binomial with 10 trials.

The most common use of an exact binomial test is when the null hypothesis is that numbers of the two outcomes are equal. In that case, the meaning of a two-tailed test is clear, and the two-tailed P-value is found by multiplying the one-tailed P-value times two.

When the null hypothesis is not a 1:1 ratio, but something like a 3:1 ratio, the meaning of a two-tailed exact binomial test is not agreed upon; different statisticians, and different statistical programs, have slightly different interpretations and sometimes give different results for the same data. My spreadsheet adds the probabilities of all possible outcomes that are less likely than the observed numbers; this method of small P-values is preferred by most statisticians.

Examples

Mendel crossed pea plants that were heterozygotes for green pod/yellow pod; pod color is the nominal variable, with "green" and "yellow" as the values. If this is inherited as a simple Mendelian trait, with green dominant over yellow, the expected ratio in the offspring is 3 green: 1 yellow. He observed 428 green and 152 yellow. The expected numbers of plants under the null hypothesis are 435 green and 145 yellow, so Mendel observed slightly fewer green-pod plants than expected. The P-value for an exact binomial test is 0.533, indicating that the null hypothesis cannot be rejected; there is no significant difference between the observed and expected frequencies of pea plants with green pods.

Roptrocerus xylophagorum is a parasitoid of bark beetles. To determine what cues these wasps use to find the beetles, Sullivan et al. (2000) placed female wasps in the base of a Y-shaped tube, with a different odor in each arm of the Y, then counted the number of wasps that entered each arm of the tube. In one experiment, one arm of the Y had the odor of bark being eaten by adult beetles, while the other arm of the Y had bark being eaten by larval beetles. Ten wasps entered the area with the adult beetles, while 17 entered the area with the larval beetles. The difference from the expected 1:1 ratio is not significant (P=0.248). In another experiment that compared infested bark with a mixture of infested and uninfested bark, 36 wasps moved towards the infested bark, while only 7 moved

towards the mixture; this is significantly different from the expected ratio ($P=9\times10^{-6}$).

Yukilevich and True (2008) mixed 30 male and 30 female *Drosophila melanogaster* from Alabama with 30 male and 30 females from Grand Bahama Island. They observed 246 matings; 140 were homotypic (male and female from the same location), while 106 were heterotypic (male and female from different locations). The null hypothesis is that the flies mate at random, so that there should be equal numbers of homotypic and heterotypic matings. There were significantly more homotypic matings (exact binomial test, P=0.035) than heterotypic.

Graphing the results

You plot the results of an exact test the same way would any other goodness-of-fit test.

Similar tests

A G-test or chi-square goodness-of-fit test could also be used for the same data as the exact test of goodness-of-fit. Where the expected numbers are small, the exact test will give more accurate results than the G-test or chi-squared tests. Where the sample size is large (over a thousand), attempting to use the exact test may give error messages (computers have a hard time calculating factorials for large numbers), so a G-test or chi-square test must be used. For intermediate sample sizes, all three tests give approximately the same results. I recommend that you use the exact test when *n* is less than 1000; see the web page on small sample sizes for further discussion.

The exact test and randomization test should give you the same result, if you do enough replicates for the randomization test, so the choice between them is a matter of personal preference. The exact test sounds more "exact"; the randomization test may be easier to understand and explain.

The sign test is a particular application of the exact binomial test. It is usually used when observations of a measurement variable are made in pairs (such as right-vs.-left or before-vs.-after), and only the direction of the difference, not the size of the difference, is of biological interest.

The exact test for goodness-of-fit is not the same as Fisher's exact test of independence. A test of independence is used for two nominal variables, such as sex and location. If you wanted to compare the ratio of males to female students at Delaware to the male:female ratio at Maryland, you would use a test of independence; if you want to compare the male:female ratio at Delaware to a theoretical 1:1 ratio, you would use a goodness-of-fit test.

How to do the test

Spreadsheet

I have set up a spreadsheet that performs the exact binomial test for sample sizes up to 1000. It is self-explanatory.

Web page

Richard Lowry has set up a web page (http://faculty.vassar.edu/lowry/binomialX.html) that does the exact binomial test. I'm not aware of any web pages that will do exact multinomial tests.

SAS

Here is a sample SAS program, showing how to do the exact binomial test on the Gus data. The `p=0.5` gives the expected proportion of whichever value of the nominal variable is alphabetically first; in this case, it gives the expected proportion of "left."

The SAS exact binomial function finds the two-tailed P-value by doubling the P-value of one tail. The binomial distribution is not symmetrical when the expected proportion is other than 50 percent, so the technique SAS uses isn't as good as the method of small P-values. I don't recommend doing the exact binomial test in SAS when the expected proportion is anything other than 50 percent.

```
data gus;
   input paw $;
   cards;
right
left
right
right
right
right
left
right
right
right
;
proc freq data=gus;
   tables paw / binomial(p=0.5);
exact binomial;
run;
```

Near the end of the output is this:

```
Exact Test
One-sided Pr <=  P            0.0547
Two-sided = 2 * One-sided     0.1094
```

The "Two-sided=2*One-sided" number is the two-tailed P-value that you want.

If you have the total numbers, rather than the raw values, you'd use a "weight" parameter in PROC FREQ. The `zeros` option tells it to include observations with counts of zero, for example if Gus had used his left paw 0 times; it doesn't hurt to always include the `zeros` option.

```
data gus;
   input paw $ count;
   cards;
right 10
left 2
;
proc freq data=gus;
   weight count / zeros;
   tables paw / binomial(p=0.5);
exact binomial;
run;
```

This example shows how do to the exact multinomial test. The numbers are Mendel's data from a genetic cross in which you expect a 9:3:3:1 ratio of peas that are round+yellow, round+green, wrinkled+yellow, and wrinkled+green. The `order=data` option tells SAS to analyze the data in the order they are input (rndyel, rndgrn, wrnkyel, wrnkgrn, in this case), not alphabetical order. The `testp=(0.5625 0.1875 0.0625 0.1875)` lists the expected proportions in the same order.

```
data peas;
   input color $ count;
   cards;
rndyel 315
rndgrn 108
wrnkyel 101
wrnkgrn 32
;
proc freq data=peas order=data;
   weight count / zeros;
   tables color / chisq testp=(0.5625 0.1875 0.1875 0.0625);
   exact chisq;
run;
```

The P-value you want is labelled "Exact Pr >= ChiSq":

```
     Chi-Square Test
for Specified Proportions
-----------------------------------
Chi-Square                    0.4700
DF                                 3
Asymptotic Pr >  ChiSq        0.9254
Exact      Pr >= ChiSq        0.9272
```

Power analysis

For the exact binomial test, you can do the power analysis with the form at http://udel.edu/~mcdonald/statexactbin.html. Enter the probability of one of the two outcomes under the null hypothesis; the probability under the alternative hypothesis; the significance level of the test (you will almost always use 0.05); and the power (0.80, 0.90, and 0.50 are common values). You should almost always use the two-tailed test.

As an example, let's say I wanted to do an experiment to see if Gus the cat really did use one paw more than the other for getting my attention. The null hypothesis is that the probability that he uses his left paw is 0.50. I decide that if the probability of him using his left paw is 0.40, I want my experiment to have an 80% probability of getting a significant (P<0.05) result. If he uses his left paw 60% of the time, I'll accept that as a significant result too, so it's a two-tailed test. Entering 0.50, 0.40, 0.05, and 0.80 in the boxes, and choosing two-tailed test, the result is 169. This means that if Gus really is using his left paw 40% (or 60%) of the time, a sample size of 169 observations will have an 80% probability of giving me a significant (P<0.05) exact binomial test.

Note that if the null expectation is not 0.50, you will get different results, depending on whether you make the alternative proportion smaller or larger than the expected. For example, if you do a genetic cross in which 25% of the offspring are expected to be yellow, it would take 117 observations to give you 80% power if the alternative hypothesis is 15% yellow, but 141 observations if the alternative hypothesis is 35% yellow. In this situation, you should probably use the larger number.

This form uses exact probability calculations for sample sizes less than 500, and the normal approximation for larger sample sizes. Techniques used by other programs may give somewhat different results (Chernick and Liu 2002).

If your nominal variable has more than two values, use this power and sample size page (http://www.stat.uiowa.edu/~rlenth/Power/index.html) . It is designed for chi-square tests, not exact tests, but the sample sizes will be very close. Choose "Generic chi-square test" from the box on the left side of the page (if you don't see the list of tests, make sure your web browser has Java turned on). Under "Prototype data," enter the chi-square value and sample size for some fake

data. For example, if you're doing a genetic cross with an expected 1:2:1 ratio, and your minimum effect size is 10 percent more heterozygotes than expected, use the chi-square spreadsheet to do a chi-square test on observed numbers of 20:60:20 compared to expected proportions of 1:2:1. The spreadsheet gives you a chi-square value of 4.00 and an *n* of 100, which you enter under "Prototype data". Then set *d* (the degrees of freedom) equal to 2, and leave alpha at 0.05. The sliders can then be slid back and forth to yield the desired result. For example, if you slide the Power to 0.90, *n* is equal to 316. Note that the absolute values of the prototype data don't matter, only their relative relationship; you could have used 200:600:200, which would give you a chi-square value of 40.0 and an n of 1000, and gotten the exact same result.

Further reading

Sokal and Rohlf, pp. 686-687.

Zar, pp. 533-538.

References

Chernick, M.R., and C.Y. Liu. 2002. The saw-toothed behavior of power versus sample size and software solutions: single binomial proportion using exact methods. Amer. Stat. 56: 149-155.

Mendel, G. 1865. Experiments in plant hybridization. available at MendelWeb. (http://www.mendelweb.org/Mendel.html)

Sullivan, B.T., E.M. Pettersson, K.C. Seltmann, and C.W. Berisford. 2000. Attraction of the bark beetle parasitoid *Roptrocerus xylophagorum* (Hymenoptera: Pteromalidae) to host-associated olfactory cues. Env. Entom. 29: 1138-1151.

Yukilevich, R., and J.R. True. 2008. Incipient sexual isolation among cosmopolitan *Drosophila melanogaster* populations. Evolution 62: 2112-2121.

Power analysis

When you are designing an experiment, it is a good idea to estimate the sample size you'll need. This is especially true if you're proposing to do something painful to humans or other vertebrates, where it is particularly important to minimize the number of individuals (without making the sample size so small that the whole experiment is a waste of time and suffering), or if you're planning a very time-consuming or expensive experiment. Methods have been developed for many statistical tests to estimate the sample size needed to detect a particular effect, or to estimate the size of the effect that can be detected with a particular sample size.

In order to do a power analysis, you need to specify an effect size. This is the size of the difference between your null hypothesis and the alternative hypothesis that you hope to detect. For applied and clinical biological research, there may be a very definite effect size that you want to detect. For example, if you're testing a new dog shampoo, the marketing department at your company may tell you that producing the new shampoo would only be worthwhile if it made dogs' coats at least 25% shinier, on average. That would be your effect size, and you would use it it deciding how many dogs you would need to put through the canine reflectometer.

When doing basic biological research, you often don't know how big a difference you're looking for, and the temptation may be to just use the biggest sample size you can afford, or use a similar sample size to other research in your field. You should still do a power analysis before you do the experiment, just to get an idea of what kind of effects you could detect. For example, some anti-vaccination kooks have proposed that the U.S. government conduct a large study of unvaccinated and vaccinated children to see whether vaccines cause autism. It is not clear what effect size would be interesting: 10% more autism in one group? 50% more? twice as much? However, doing a power analysis shows that even if the study included every unvaccinated child in the United States aged 3 to 6, and an equal number of vaccinated children, there would have to be 25% more autism in one group in order to have a high chance of seeing a significant difference. A more plausible study, of 5,000 unvaccinated and 5,000 vaccinated children, would detect a significant difference with high power only if there were three times more autism in one group than the other. Because it is unlikely that there is such a big difference in autism between vaccinated and unvaccinated children, and because failing to find a relationship with such a study would not convince anti-

vaccination kooks that there was no relationship (*nothing* would convince them there's no relationship—that's what makes them kooks), the power analysis tells you that such a study would not be worthwhile.

Parameters

There are four or five numbers involved in a power analysis. The minimum effect size is the minimum deviation from the null hypothesis that you hope to detect. For example, if you are treating hens with something that you hope will change the sex ratio of their chicks, you might decide that the minimum change in the proportion of sexes that you're looking for is 10 percent. You might have a good economic reason for choosing the effect size; if not, you might want to see what kind of effects other people have found in similar experiments. If you don't have a particular effect size in mind, you might want to try different effect sizes and produce a graph of effect size vs. sample size.

Alpha is the significance level of the test (the P-value), the probability of rejecting the null hypothesis even though it is true (a false positive). The usual value is alpha=0.05. Some power calculators use the one-tailed alpha, which is confusing, since the two-tailed alpha is much more common. Be sure you know which you're using.

Beta, in a power analysis, is the probabilty of accepting the null hypothesis, even though it is false (a false negative), when the real difference is equal to the minimum effect size. The power of a test is the probability of rejecting the null hypothesis when the real difference is equal to the minimum effect size, or 1−beta. There is no clear consensus on the value to use; a power of 80% (equivalent to a beta of 20%) is probably the most common, while powers of 50% and 90% are also sometimes used. The cost to you of a false negative should influence your choice of power; if you really, really want to be sure that you detect your effect size, you'll want to use a higher value for power (lower beta), which will result in a bigger sample size. Some power calculators ask you to enter beta, while others ask for power (1−beta); be very sure you understand which you need to use.

For measurement variables, you also need an estimate of the standard deviation. This can come from pilot experiments or from similar experiments in the published literature. Your standard deviation once you do the experiment is unlikely to be exactly the same, so your experiment will actually be somewhat more or less powerful than you had predicted. For nominal variables, the standard deviation is a simple function of the sample size, so you don't need to estimate it separately.

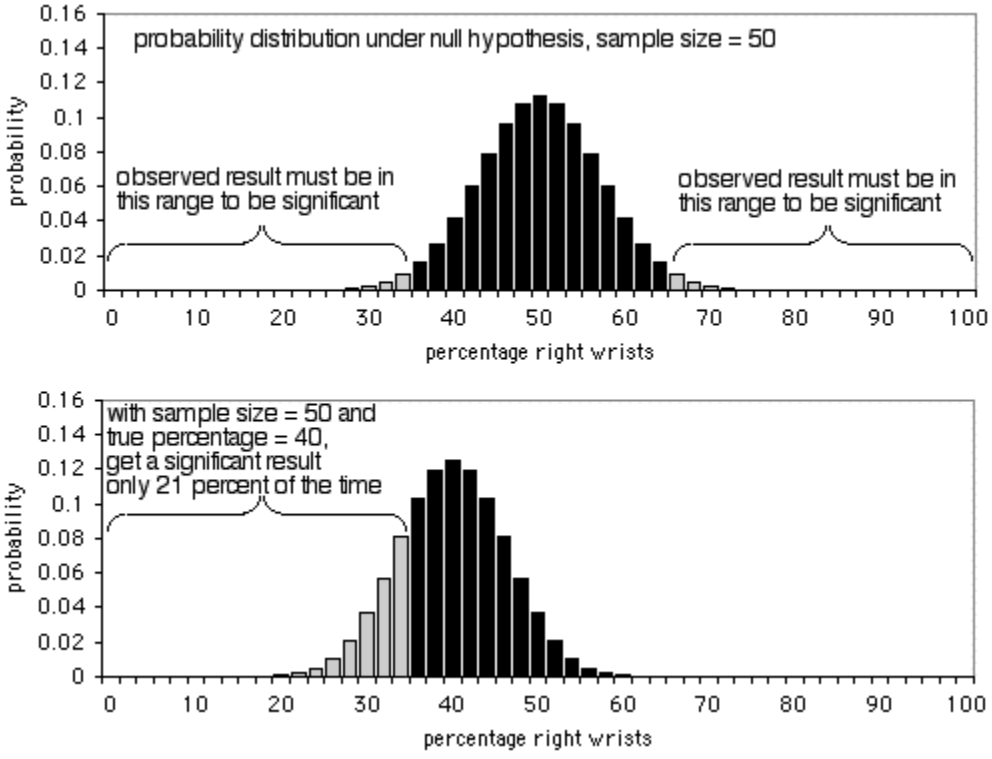
On most web pages that do power analyses, you can either enter the desired power and estimate the sample size needed, or enter the sample size and estimate the power. If the effect size is really the minimum specified, and the standard deviation is as specified, the probability that this sample size will give a significant

result (at the P<alpha level) is 1−beta, and the probability that it won't give a significant result is beta.

The equations for estimating sample size from alpha, beta, standard deviation, and minimum effect size can be quite complicated. Fortunately, there are online calculators for doing power analyses for many statistical tests. I'll try to put a link for power analysis on the web page for each statistical test.

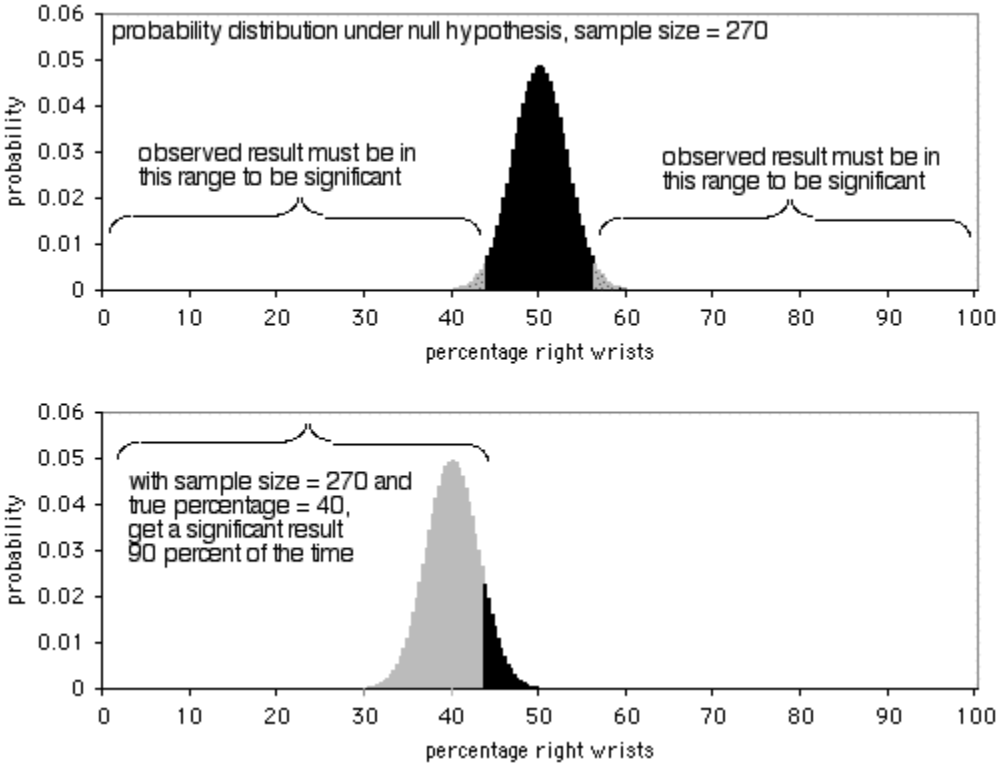
How it works

The details of a power analysis are different for different statistical tests, but the basic concepts are similar; here I'll use the exact binomial test as an example. Imagine that you are studying wrist fractures, and your null hypothesis is that half the people who break one wrist break their right wrist, and half break their left. You decide that the minimum effect size is 10 percent; if the percentage of people who break their right wrist is 60 percent or more, or 40 percent or less, you want to have a significant result from the exact binomial test. Alpha is 5 percent, as usual. You want power to be 90 percent, which means that if the percentage of broken right wrists is 40 percent or 60 percent, you want a sample size that will yield a significant (P<0.05) result 90 percent of the time, and a non-significant result (which would be a false negative in this case) only 10 percent of the time.



The first graph shows the probability distribution under the null hypothesis, with a sample size of 50 individuals. In order to be significant at the P<0.05 level,

the observed result would have to be less than 36 percent or more than 64 percent of people breaking their right wrists. As the second graph shows, if the true percentage is 40 percent, the sample data will be this extreme only 21 percent of the time. Obviously, a sample size of 50 is too small for this experiment; it would only yield a significant result 21 percent of the time, even if there's a 40:60 ratio of broken right wrists to left wrists.



The next graph shows the probability distribution under the null hypothesis, with a sample size of 270 individuals. In order to be significant at the P<0.05 level, the observed result would have to be less than 43.7 percent or more than 56.3 percent of people breaking their right wrists. As the second graph shows, if the true percentage is 40 percent, the sample data will be this extreme 90 percent of the time. A sample size of 270 is pretty good for this experiment; it would yield a significant result 90 percent of the time if there's a 40:60 ratio of broken right wrists to left wrists.

Examples

You plan to cross peas that are heterozygotes for Yellow/green pea color, where Yellow is dominant. The expected ratio in the offspring is 3 Yellow: 1 green. You want to know whether yellow peas are actually more or less fit, which might show up as a different proportion of yellow peas than expected. You *arbitrarily*

decide that you want a sample size that will detect a significant (P<0.05) difference if there are 3 percent more or fewer yellow peas than expected, with a power of 90 percent. You will test the data using the exact binomial test for goodness-of-fit if the sample size is small enough, or a G-test for goodness-of-fit if the sample size is larger. The power analysis is the same for both tests.

Go to the power calculator on the exact binomial test web page. Enter 0.75 for "Proportion if null hypothesis is true" and enter 0.72 for "Proportion if alternative hypothesis is true" (3 percent fewer yellow peas). Enter 0.05 for alpha and 0.90 for power, then click on "Run." The result is 2253. That's a lot of peas! Note that, because the confidence intervals on a percentage are not symmetrical, the results are different if you enter 0.78 for "Proportion if null hypothesis is true"; you should try it both ways and use the larger sample size result.

The example data for Student's t-test shows that the average height in the 2 p.m. section of Biological Data Analysis was 66.6 inches and the average height in the 5 p.m. section was 64.6 inches, but the difference is not significant (P=0.207). You want to know how many students you'd have to sample to have an 80 percent chance of a difference this large being significant. Go to the power calculator on the Student's t-test web page. Enter 2.0 for the difference in means. Using the STDEV function in Excel, calculate the standard deviation for each sample in the original data; it is 4.8 for sample 1 and 3.6 for sample 2. Enter 0.05 for alpha and 0.80 for power. The result is 72, meaning that if 5 p.m. students really were two inches shorter than 2 p.m. students, you'd need 72 students in each class to detect a significant difference 80 percent of the time, if the true difference really is 2.0 inches.

How to do power analyses

Web pages

I have put sample size calculators on the web pages in this handbook for some of the simpler tests, including the exact binomial test, Student's t-test, and paired t-test. Russ Lenth has put together a more extensive set of power analyses (http://www.stat.uiowa.edu/~rlenth/Power/index.html) .

G*Power

G*Power (http://www.psycho.uni-duesseldorf.de/abteilungen/aap/gpower3/) is an excellent free program, available for Mac and Windows, that will do power analyses for a large variety of tests. Calculating the effect size parameter can be the most difficult part of a power analysis, so one advantage of G*Power is that it allows you to calculate the effect size parameter using sample data; click on the "Determine" button next to the box labelled "effect size," fill in data that looks like what you want to detect, and it will calculate the effect size.

SAS

SAS has a PROC POWER that you can use for power analyses. You enter the needed parameters (which vary depending on the test) and enter a period (which symbolizes missing data in SAS) for the parameter you're solving for (usually `ntotal`, the total sample size, or `npergroup`, the number of samples in each group). I find that G*Power is easier to use than SAS for this purpose, so I don't recommend using SAS for your power analyses.

Further reading

Sokal and Rohlf, pp. 167-169.

Zar, p. 83.

Chi-square test for goodness-of-fit

The chi-square test for goodness-of-fit is an alternative to the G-test for goodness-of-fit. Most of the information on this page is identical to that on the G-test page. You should read the section on "Chi-square vs. G-test" near the bottom of this page, pick either chi-square or G-test, then stick with that choice for the rest of your life.

When to use it

Use the chi-square test for goodness-of-fit when you have one nominal variable with two or more values (such as red, pink and white flowers). The observed counts of numbers of observations in each category are compared with the expected counts, which are calculated using some kind of theoretical expectation (such as a 1:1 sex ratio or a 1:2:1 ratio in a genetic cross).

If the expected number of observations in any category is too small, the chi-square test may give inaccurate results, and an exact test or a randomization test should be used instead. See the web page on small sample sizes for further discussion.

Null hypothesis

The statistical null hypothesis is that the number of observations in each category is equal to that predicted by a biological theory, and the alternative hypothesis is that the observed numbers are different from the expected. The null hypothesis is usually an extrinsic hypothesis, one for which the expected proportions are determined before doing the experiment. Examples include a 1:1 sex ratio or a 1:2:1 ratio in a genetic cross. Another example would be looking at an area of shore that had 59% of the area covered in sand, 28% mud and 13% rocks; if seagulls were standing in random places, your null hypothesis would be that 59% of the seagulls were standing on sand, 28% on mud and 13% on rocks.

In some situations, an intrinsic hypothesis is used. This is a null hypothesis in which the expected proportions are calculated after the experiment is done, using some of the information from the data. The best-known example of an intrinsic

hypothesis is the Hardy-Weinberg proportions of population genetics: if the frequency of one allele in a population is p and the other allele is q, the null hypothesis is that expected frequencies of the three genotypes are p^2, $2pq$, and q^2. This is an intrinsic hypothesis, because p and q are estimated from the data after the experiment is done, not predicted by theory before the experiment.

How the test works

The test statistic is calculated by taking an observed number (O), subtracting the expected number (E), then squaring this difference. The larger the deviation from the null hypothesis, the larger the difference between observed and expected is. Squaring the differences makes them all positive. Each difference is divided by the expected number, and these standardized differences are summed. The test statistic is conventionally called a "chi-square" statistic, although this is somewhat confusing (it's just one of many test statistics that follows the chi-square distribution). The equation is

$$\mathrm{chi}^2 = \sum (O-E)^2/E$$

As with most test statistics, the larger the difference between observed and expected, the larger the test statistic becomes.

The distribution of the test statistic under the null hypothesis is approximately the same as the theoretical chi-square distribution. This means that once you know the chi-square test statistic, you can calculate the probability of getting that value of the chi-square statistic.

The shape of the chi-square distribution depends on the number of degrees of freedom. For an extrinsic null hypothesis (the much more common situation, where you know the proportions predicted by the null hypothesis before collecting the data), the number of degrees of freedom is simply the number of values of the variable, minus one. Thus if you are testing a null hypothesis of a 1:1 sex ratio, there are two possible values (male and female), and therefore one degree of freedom. This is because once you know how many of the total are females (a number which is "free" to vary from 0 to the sample size), the number of males is determined. If there are three values of the variable (such as red, pink, and white), there are two degrees of freedom, and so on.

An intrinsic null hypothesis is one in which you estimate one or more parameters from the data in order to get the numbers for your null hypothesis. As described above, one example is Hardy-Weinberg proportions. For an intrinsic null hypothesis, the number of degrees of freedom is calculated by taking the number of values of the variable, subtracting 1 for each parameter estimated from the data, then subtracting 1 more. Thus for Hardy-Weinberg proportions with two alleles and three genotypes, there are three values of the variable (the three

genotypes); you subtract one for the parameter estimated from the data (the allele frequency, *p*); and then you subtract one more, yielding one degree of freedom.

Examples: extrinsic hypothesis

Mendel crossed peas that were heterozygotes for Smooth/wrinkled, where Smooth is dominant. The expected ratio in the offspring is 3 Smooth: 1 wrinkled. He observed 423 Smooth and 133 wrinkled.

The expected frequency of Smooth is calculated by multiplying the sample size (556) by the expected proportion (0.75) to yield 417. The same is done for green to yield 139. The number of degrees of freedom when an extrinsic hypothesis is used is the number of values of the nominal variable minus one. In this case, there are two values (Smooth and wrinkled), so there is one degree of freedom.

The result is chi-square=0.35, 1 d.f., P=0.557, indicating that the null hypothesis cannot be rejected; there is no significant difference between the observed and expected frequencies.

Mannan and Meslow (1984) studied bird foraging behavior in a forest in Oregon. In a managed forest, 54% of the canopy volume was Douglas fir, 40% was ponderosa pine, 5% was grand fir, and 1% was western larch. They made 156 observations of foraging by red-breasted nuthatches; 70 observations (45% of the total) in Douglas fir, 79 (51%) in ponderosa pine, 3 (2%) in grand fir, and 4 (3%) in western larch. The biological null hypothesis is that the birds forage randomly, without regard to what species of tree they're in; the statistical null hypothesis is that the proportions of foraging events are equal to the proportions of canopy volume. The difference in proportions is significant (chi-square=13.593, 3 d.f., P=0.0035).

The expected numbers in this example are pretty small, so it would be better to analyze it with an exact test or a randomization test. I'm leaving it here because it's a good example of an extrinsic hypothesis that comes from measuring something (canopy volume, in this case), not a mathematical theory.

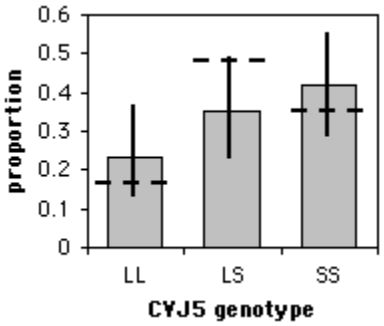
Example: intrinsic hypothesis

McDonald et al. (1996) examined variation at the CVJ5 locus in the American oyster, *Crassostrea virginica*. There were two alleles, L and S, and the genotype frequencies in Panacea, Florida were 14 LL, 21 LS, and 25 SS. The estimate of the L allele proportion from the data is 49/120=0.408. Using the Hardy-Weinberg formula and this estimated allele proportion, the expected genotype proportions are 0.167 LL, 0.483 LS, and 0.350 SS. There are three classes (LL, LS and SS) and one parameter estimated from the data (the L allele proportion), so there is one degree of freedom. The result is chi-square=4.54, 1 d.f., P=0.033, which is significant at the 0.05 level. We can reject the null hypothesis that the data fit the expected Hardy-Weinberg proportions.

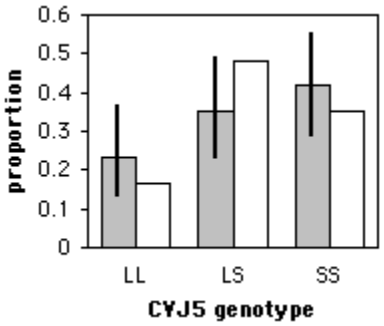
Graphing the results

If there are just two values of the nominal variable, you wouldn't display the result in a graph, as that would be a bar graph with just one bar. Instead, you just report the proportion; for example, Mendel found 23.9% wrinkled peas in his cross.

With more than two values of the nominal variable, you'd usually present the results of a goodness-of-fit test in a table of observed and expected proportions. If the expected values are obvious (such as 50%) or easily calculated from the data (such as Hardy–Weinberg proportions), you can omit the expected numbers from your table. For a presentation you'll probably want a graph showing both the observed and expected proportions, to give a visual impression of how far apart they are. You should use a bar graph for the observed proportions; the expected can be shown with a horizontal dashed line, or with bars of a different pattern.



Genotype proportions at the CVJ5 locus in the American oyster. Horizontal dashed lines indicate the expected proportions under Hardy–Weinberg equilibrium; error bars indicate 95% confidence intervals.



Genotype proportions at the CVJ5 locus in the American oyster. Gray bars are observed proportions, with 95% confidence intervals; white bars are expected proportions under Hardy–Weinberg equilibrium.

One way to get the horizontal lines on the graph is to set up the graph with the observed proportions and error bars, set the scale for the Y-axis to be fixed for the minimum and maximum you want, and get everything formatted (fonts, patterns, etc.). Then replace the observed proportions with the expected proportions in the spreadsheet; this should make the columns change to represent the expected values. Using the spreadsheet drawing tools, draw horizontal lines at the top of the columns. Then put the observed proportions back into the spreadsheet. Of course, if the expected proportion is something simple like 25%, you can just draw the horizontal line all the way across the graph.

Similar tests

The chi-square test of independence is used for two nominal variables, not one.

There are several tests that use chi-square statistics. The one described here is formally known as Pearson's chi-square. It is by far the most common chi-square test, so it is usually just called the chi-square test.

You have a choice of four goodness-of-fit tests: the exact binomial test or exact multinomial test, the G-test of goodness-of-fit,, the chi-square test of goodness-of-fit, or the randomization test. For small values of the expected numbers, the chi-square and G-tests are inaccurate, because the distribution of the test statistics do not fit the chi-square distribution very well.

The usual rule of thumb is that you should use the exact test or randomization test when the smallest expected value is less than 5, and the chi-square and G-tests are accurate enough for larger expected values. This rule of thumb dates from the olden days when statistics were done by hand, and the calculations for the exact test were very tedious and to be avoided if at all possible. Nowadays, computers make it just as easy to do the exact test or randomization test as the computationally simpler chi-square or G-test. I recommend that you use the exact test when the total sample size is less than 1000. With sample sizes between 50 and 1000, it generally doesn't make much difference which test you use, so you shouldn't criticize someone for using the chi-square or G-test (as I have in the examples above). See the web page on small sample sizes for further discussion.

Chi-square vs. G-test

The chi-square test gives approximately the same results as the G-test. Unlike the chi-square test, G-values are additive, which means they can be used for more elaborate statistical designs, such as repeated G-tests of goodness-of-fit. G-tests are a subclass of likelihood ratio tests, a general category of tests that have many uses for testing the fit of data to mathematical models; the more elaborate versions of likelihood ratio tests don't have equivalent tests using the Pearson chi-square statistic. The G-test is therefore preferred by many, even for simpler designs. On the other hand, the chi-square test is more familiar to more people, and it's always a good idea to use statistics that your readers are familiar with when possible. You may want to look at the literature in your field and see which is more commonly used.

How to do the test

Spreadsheet

I have set up a spreadsheet for the chi-square test of goodness-of-fit. It is largely self-explanatory. It will calculate the degrees of freedom for you if you're using an extrinsic null hypothesis; if you are using an intrinsic hypothesis, you must enter the degrees of freedom into the spreadsheet.

Web pages

There are also web pages that will perform this test here (http://www.graphpad.com/quickcalcs/chisquared2.cfm) , here (http://faculty.vassar.edu/lowry/csfit.html) , or here (http://quantrm2.psy.ohio-state.edu/kris/chisq/chisq.htm) . None of these web pages lets you set the degrees of freedom to the appropriate value for testing an intrinsic null hypothesis.

SAS

Here is a SAS program that uses PROC FREQ for a chi-square test. It uses the Mendel pea data from above, and it assumes you've already counted the number of smooth and wrinkled peas. The `weight count` command tells SAS that the 'count' variable is the number of times each value of 'texture' was observed. The `zeros` option tells it to include observations with counts of zero, for example if you had 20 smooth peas and 0 wrinkled peas; it doesn't hurt to always include the `zeros` option. `chisq` tells SAS to do a chi-square test, and `testp=(75 25);` tells it the expected percentages. The expected percentages must add up to 100. The expected percentages are given for the values of 'texture' in alphabetical order: 75 percent 'smooth', 25 percent 'wrinkled'.

```
data peas;
   input texture $ count / zeros;
   cards;
smooth 423
wrinkled 133
;
proc freq data=peas;
   weight count;
   tables texture / chisq testp=(75 25);
run;
```

Here's a SAS program that uses PROC FREQ for a chi-square test on raw data. I've used two dots to indicate that I haven't shown the complete data set.

```
data peas;
   input texture $;
   cards;
smooth
wrinkled
   .
   .
smooth
;
proc freq data=peas;
   tables texture / chisq testp=(75 25);
run;
```

The output includes the following:

```
     Chi-Square Test
for Specified Proportions
-------------------------
Chi-Square         0.3453
DF                      1
Pr > ChiSq         0.5568
```

You would report this as "chi-square=0.3453, 1 d.f., P=0.5568."

Power analysis

If your nominal variable has just two values, you can use the power calculator on the exact binomial page.

If your nominal variable has more than two values, use the free G*Power (http://www.psycho.uni-duesseldorf.de/abteilungen/aap/gpower3/) program. Choose "Goodness-of-fit tests: Contingency tables" from the Statistical Test menu, then choose "Chi-squared tests" from the Test Family menu. To calculate effect size, click on the Determine button and enter the null hypothesis proportions in the first column and the proportions you hope to see in the second column. Then click on the Calculate and Transfer to Main Window button. Set your alpha and power, and be sure to set the degrees of freedom (Df); for an extrinsic null hypothesis, that will be the number of rows minus one.

As an example, let's say you want to do a genetic cross of snapdragons with an expected 1:2:1 ratio, and you want to be able to detect a pattern with 5 percent more heterozygotes that expected. Enter 0.25, 0.50, and 0.25 in the first column, enter 0.225, 0.55, and 0.225 in the second column, click on Calculate and Transfer to Main Window, enter 0.05 for alpha, 0.80 for power, and 2 for degrees of freedom. The result is a total sample size of 964.

Further reading

Sokal and Rohlf, p. 701.

Zar, pp. 462-466.

References

Mannan, R.W., and E.C. Meslow. 1984. Bird populations and vegetation characteristics in managed and old-growth forests, northeastern Oregon. J. Wildl. Manage. 48: 1219-1238.

McDonald, J.H., B.C. Verrelli and L.B. Geyer. 1996. Lack of geographic variation in anonymous nuclear polymorphisms in the American oyster, *Crassostrea virginica.* Molecular Biology and Evolution 13: 1114-1118.

G-test for goodness-of-fit

The G-test for goodness-of-fit, also known as a likelihood ratio test for goodness-of-fit, is an alternative to the chi-square test of goodness-of-fit. Most of the information on this page is identical to that on the chi-square page. You should read the section on "Chi-square vs. G-test" near the bottom of this page, pick either chi-square or G-test, then stick with that choice for the rest of your life.

When to use it

Use the G-test for goodness-of-fit when you have one nominal variable with two or more values (such as red, pink and white flowers). The observed counts of numbers of observations in each category are compared with the expected counts, which are calculated using some kind of theoretical expectation (such as a 1:1 sex ratio or a 1:2:1 ratio in a genetic cross).

If the expected number of observations in any category is too small, the G-test may give inaccurate results, and an exact test or a randomization test should be used instead. See the web page on small sample sizes for further discussion.

Null hypothesis

The statistical null hypothesis is that the number of observations in each category is equal to that predicted by a biological theory, and the alternative hypothesis is that the observed numbers are different from the expected. The null hypothesis is usually an extrinsic hypothesis, one for which the expected proportions are determined before doing the experiment. Examples include a 1:1 sex ratio or a 1:2:1 ratio in a genetic cross. Another example would be looking at an area of shore that had 59% of the area covered in sand, 28% mud and 13% rocks; if seagulls were standing in random places, your null hypothesis would be that 59% of the seagulls were standing on sand, 28% on mud and 13% on rocks.

In some situations, an intrinsic hypothesis is used. This is a null hypothesis in which the expected proportions are calculated after the experiment is done, using some of the information from the data. The best-known example of an intrinsic hypothesis is the Hardy-Weinberg proportions of population genetics: if the frequency of one allele in a population is p and the other allele is q, the null hypothesis is that expected frequencies of the three genotypes are p^2, $2pq$, and q^2.

This is an intrinsic hypothesis, because p and q are estimated from the data after the experiment is done, not predicted by theory before the experiment.

How the test works

The test statistic is calculated by taking an observed number (O), dividing it by the expected number (E), then taking the natural log of this ratio. The natural log of 1 is 0; if the observed number is larger than the expected, ln(O/E) is positive, while if O is less than E, ln(O/E) is negative. Each log is multiplied by the observed number, then these products are summed and multiplied by 2. The test statistic is usually called G, and thus this is a G-test, although it is also sometimes called a log-likelihood test or a likelihood ratio test. The equation is

$$G=2\sum[O\times\ln(O/E)]$$

As with most test statistics, the larger the difference between observed and expected, the larger the test statistic becomes.

The distribution of the G-statistic under the null hypothesis is approximately the same as the theoretical chi-square distribution. This means that once you know the G-statistic, you can calculate the probability of getting that value of G using the chi-square distribution.

The shape of the chi-square distribution depends on the number of degrees of freedom. For an extrinsic null hypothesis (the much more common situation, where you know the proportions predicted by the null hypothesis before collecting the data), the number of degrees of freedom is simply the number of values of the variable, minus one. Thus if you are testing a null hypothesis of a 1:1 sex ratio, there are two possible values (male and female), and therefore one degree of freedom. This is because once you know how many of the total are females (a number which is "free" to vary from 0 to the sample size), the number of males is determined. If there are three values of the variable (such as red, pink, and white), there are two degrees of freedom, and so on.

An intrinsic null hypothesis is one in which you estimate one or more parameters from the data in order to get the numbers for your null hypothesis. As described above, one example is Hardy-Weinberg proportions. For an intrinsic null hypothesis, the number of degrees of freedom is calculated by taking the number of values of the variable, subtracting 1 for each parameter estimated from the data, then subtracting 1 more. Thus for Hardy-Weinberg proportions with two alleles and three genotypes, there are three values of the variable (the three genotypes); you subtract one for the parameter estimated from the data (the allele frequency, p); and then you subtract one more, yielding one degree of freedom.

Examples: extrinsic hypothesis

Mendel crossed peas that were heterozygotes for Smooth/wrinkled, where Smooth is dominant. The expected ratio in the offspring is 3 Smooth: 1 wrinkled. He observed 423 Smooth and 133 wrinkled.

The expected frequency of Smooth is calculated by multiplying the sample size (556) by the expected proportion (0.75) to yield 417. The same is done for green to yield 139. The number of degrees of freedom when an extrinsic hypothesis is used is the number of classes minus one. In this case, there are two classes (Smooth and wrinkled), so there is one degree of freedom.

The result is G=0.35, 1 d.f., P=0.555, indicating that the null hypothesis cannot be rejected; there is no significant difference between the observed and expected frequencies.

Mannan and Meslow (1984) studied bird foraging behavior in a forest in Oregon. In a managed forest, 54% of the canopy volume was Douglas fir, 40% was ponderosa pine, 5% was grand fir, and 1% was western larch. They made 156 observations of foraging by red-breasted nuthatches; 70 observations (45% of the total) in Douglas fir, 79 (51%) in ponderosa pine, 3 (2%) in grand fir, and 4 (3%) in western larch. The biological null hypothesis is that the birds forage randomly, without regard to what species of tree they're in; the statistical null hypothesis is that the proportions of foraging events are equal to the proportions of canopy volume. The difference in proportions between observed and expected is significant (G=13.145, 3 d.f., P=0.0043).

The expected numbers in this example are pretty small, so it would be better to analyze it with an exact test or a randomization test. I'm leaving it here because it's a good example of an extrinsic hypothesis that comes from measuring something (canopy volume, in this case), not a mathematical theory.

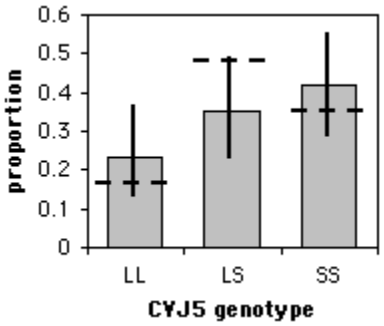
Example: intrinsic hypothesis

McDonald et al. (1996) examined variation at the CVJ5 locus in the American oyster, *Crassostrea virginica*. There were two alleles, L and S, and the genotype frequencies in Panacea, Florida were 14 LL, 21 LS, and 25 SS. The estimate of the L allele proportion from the data is 49/120=0.408. Using the Hardy-Weinberg formula and this estimated allele proportion, the expected genotype proportions are 0.167 LL, 0.483 LS, and 0.350 SS. There are three classes (LL, LS and SS) and one parameter estimated from the data (the L allele proportion), so there is one degree of freedom. The result is G=4.56, 1 d.f., P=0.033, which is significant at the 0.05 level. You can reject the null hypothesis that the data fit the expected Hardy-Weinberg proportions.

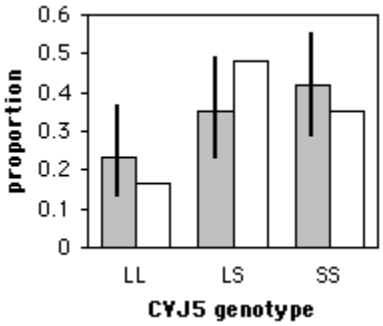
Graphing the results

If there are just two values of the nominal variable, you wouldn't display the result in a graph, as that would be a bar graph with just one bar. Instead, you just report the proportion; for example, Mendel found 23.9% wrinkled peas in his cross.

With more than two values of the nominal variable, you'd usually present the results of a goodness-of-fit test in a table of observed and expected proportions. If the expected values are obvious (such as 50%) or easily calculated from the data (such as Hardy–Weinberg proportions), you can omit the expected numbers from your table. For a presentation you'll probably want a graph showing both the observed and expected proportions, to give a visual impression of how far apart they are. You should use a bar graph for the observed proportions; the expected can be shown with a horizontal dashed line, or with bars of a different pattern.



Genotype proportions at the CVJ5 locus in the American oyster. Horizontal dashed lines indicate the expected proportions under Hardy–Weinberg equilibrium; error bars indicate 95% confidence intervals.



Genotype proportions at the CVJ5 locus in the American oyster. Gray bars are observed proportions, with 95% confidence intervals; white bars are expected proportions under Hardy–Weinberg equilibrium.

One way to get the horizontal lines on the graph is to set up the graph with the observed proportions and error bars, set the scale for the Y-axis to be fixed for the minimum and maximum you want, and get everything formatted (fonts, patterns, etc.). Then replace the observed proportions with the expected proportions in the spreadsheet; this should make the columns change to represent the expected values. Using the spreadsheet drawing tools, draw horizontal lines at the top of the columns. Then put the observed proportions back into the spreadsheet. Of course, if the expected proportion is something simple like 25%, you can just draw the horizontal line all the way across the graph.

Similar tests

The G-test of independence is used for two nominal variables, not one.

You have a choice of four goodness-of-fit tests: the exact binomial test or exact multinomial test, the G-test of goodness-of-fit, the chi-square test of goodness-of-fit, or the randomization test. For small values of the expected numbers, the chi-square and G-tests are inaccurate, because the distribution of the test statistics do not fit the chi-square distribution very well.

The usual rule of thumb is that you should use the exact test or randomization test when the smallest expected value is less than 5, and the chi-square and G-tests are accurate enough for larger expected values. This rule of thumb dates from the olden days when statistics were done by hand, and the calculations for the exact test were very tedious and to be avoided if at all possible. Nowadays, computers make it just as easy to do the exact test or randomization test as the computationally simpler chi-square or G-test. I recommend that you use the exact test when the total sample size is less than 1000. With sample sizes between 50 and 1000, it generally doesn't make much difference which test you use, so you shouldn't criticize someone for using the chi-square or G-test (as I have in the examples above). See the web page on small sample sizes for further discussion.

Chi-square vs. G-test

The chi-square test gives approximately the same results as the G-test. Unlike the chi-square test, the G-values are additive, which means they can be used for more elaborate statistical designs, such as repeated G-tests of goodness-of-fit. G-tests are a subclass of likelihood ratio tests, a general category of tests that have many uses for testing the fit of data to mathematical models; the more elaborate versions of likelihood ratio tests don't have equivalent tests using the Pearson chi-square statistic. The G-test is therefore preferred by many, even for simpler designs. On the other hand, the chi-square test is more familiar to more people, and it's always a good idea to use statistics that your readers are familiar with when possible. You may want to look at the literature in your field and see which is more commonly used.

How to do the test

Spreadsheet

I have set up a spreadsheet that does the G-test of goodness-of-fit. It is largely self-explanatory. It will calculate the degrees of freedom for you if you're using an extrinsic null hypothesis; if you are using an intrinsic hypothesis, you must enter the degrees of freedom into the spreadsheet.

I'm not aware of any web pages that will do a G-test of goodness-of-fit.

SAS

Surprisingly, SAS does not have an option to do a G-test of goodness-of-fit; the manual says the G-test is defined only for tests of independence, but this is incorrect.

Power analysis

If your nominal variable has just two values, use the power calculator on the exact binomial page.

If your nominal variable has more than two values, use the power analysis for chi-squared tests of goodness-of-fit.

Further reading

Sokal and Rohlf, pp. 699-701 (extrinsic hypothesis) and pp. 706-707 (intrinsic hypothesis).

Zar, pp. 473-475.

References

Mannan, R.W., and E.C. Meslow. 1984. Bird populations and vegetation characteristics in managed and old-growth forests, northeastern Oregon. J. Wildl. Manage. 48: 1219-1238.

McDonald, J.H., B.C. Verrelli and L.B. Geyer. 1996. Lack of geographic variation in anonymous nuclear polymorphisms in the American oyster, *Crassostrea virginica.* Molecular Biology and Evolution 13: 1114-1118.

Randomization test of goodness-of-fit

When to use it

Use the randomization test of goodness of fit when you have one nominal variable with three or more values (such as red vs. pink vs. white flowers), and the sample size is too small to do the chi-square test or the G-test of goodness-of-fit. An exact multinomial test would be just as good as a randomization test; I include the randomization test here because you'll find it difficult to do an exact multinomial test if you don't have access to SAS or another statistical package, and because it provides a simple introduction to the important topic of randomization-based tests (also known as Monte Carlo simulations).

The first step in doing a randomization test is to calculate the test statistic, in this case the chi-square statistic. This is a measure of how far the observed numbers are from the expected; a bigger deviation from the expected leads to a bigger chi-square statistic. When doing a chi-square test, you use the relationship that under the null hypothesis, the chi-square statistic approximately follows the mathematical chi-square distribution. With small expected numbers, this approximate relationship is not very accurate, which is why the randomization test is necessary.

For the spreadsheet and web page described here, the null hypothesis must be extrinsic (such as an expected 1: 2: 1 ratio in a genetic cross), not intrinsic (such as the p^2: 2pq: q^2 Hardy-Weinberg proportions of population genetics). If you want to do a randomization test with an intrinsic hypothesis, you will probably have to write a program yourself.

Null hypothesis

The statistical null hypothesis is that the number of observations in each category is equal to that predicted by a biological theory, and the alternative hypothesis is that the observed numbers are different from the expected. The null hypothesis is usually an extrinsic hypothesis, one for which the expected proportions are determined before doing the experiment. An example is a 1: 2: 1 ratio in a genetic cross.

How the test works

Imagine you did a cross in which you expected a 1:2:1 ratio of red to pink to white snapdragon flowers. You got only 8 offspring, so you expect 2 red, 4 pink, and 2 white, if the null hypothesis is true; you actually got 5 red, 2 pink, and 1 white.

You calculate the chi-square statistic, which is 6.00. That is significant (P=0.0498), but you know that the chi-square test can be inaccurate with such small expected numbers. So you put one red ball, two pink balls, and one white ball in a hat. Without looking, you reach in, grab a ball, and write down what color it is. You put the ball back in and repeat this process until you've sampled 8 balls from a known 1:2:1 ratio. You calculate the chi-square statistic for these numbers, and see whether it's as big or bigger than your observed chi-square statistic of 6.00.

You repeat this process of randomly sampling 8 balls from a known 1:2:1 ratio, and see how often you get a chi-square of 6.00 or larger. If you get a chi-square that large more than 5 percent of the time, it tells you that if the null hypothesis is true, you'll get your observed result (or something even more deviant from the null) more than 5 percent of the time, so you can't reject the null hypothesis. If the randomization trials produce a chi-square of 6.00 or larger less than 5 percent of the time, you reject the null hypothesis. For these numbers (5 red, 2 pink, 1 white), I get P=0.0576 after 10,000 randomizations (done on a computer, not with a hat), which is not significant.

Because it is taking a random sample of all possible combinations, the randomization test will give slightly different estimates of the P-value every time you run it. The more replicates you run, the more accurate your estimate of the P-value will be. You might want to start with a small number of replicates, such as 1,000, to be sure everything is working properly, then change the number of replicates to 100,000 or even 1,000,000 for your final result.

This randomization test of goodness-of-fit is an example of an important technique in statistics. Sometimes you want to estimate the probability of getting an observed result if a null hypothesis is true (the P-value), and you have a test statistic that measures how far the observations are from the expected, but there is no theoretical relationship between the test statistic and the P-value. If you can simulate on a computer taking random samples from a population that fits the null hypothesis, you can see how often the observed value of the test statistic occurs, and therefore estimate the P-value. This technique is often called "Monte Carlo simulation," because it's like selecting a bunch of random numbers with a roulette wheel in the casino there. More elaborate Monte Carlo simulations usually require writing a computer program or using specialized mathematical software, so they are beyond the scope of this handbook, but you should be aware of the general concept.

Example

The red-breasted nuthatch example from the chi-square and G-test of goodness-of-fit pages has some rather small expected numbers; under the null hypothesis, you'd only expect 7.8 foraging events in grand fir and 1.6 events in western larch. The chi-square and G-tests might therefore be a little inaccurate, so it would be better to use a randomization test. Using SAS to run one MILLION replicate randomizations, the proportion of chi-square values for the randomly sampled data that were equal to or greater than the observed chi-squared value (13.59) was only 0.0069; in other words, P=0.0069. This is somewhat higher than the results for the chi-square test (P=0.035) or G-test (P=0.043), but it doesn't change the conclusion, that the foraging events are signficantly different from randomly distributed among the tree species.

Graphing the results

You plot the results of a randomization test of goodness-of-fit the same way you would a chi-square test of goodness-of-fit.

Similar tests

You have a choice of four goodness-of-fit tests: the exact binomial test or exact multinomial test, the G-test of goodness-of-fit,, the chi-square test of goodness-of-fit, or the randomization test. For small values of the expected numbers, the chi-square and G-tests are inaccurate, because the distribution of the test statistics do not fit the chi-square distribution very well.

The usual rule of thumb is that you should use the exact test or randomization test when the smallest expected value is less than 5, and the chi-square and G-tests are accurate enough for larger expected values. This rule of thumb dates from the olden days when statistics were done by hand, and the calculations for the exact test were very tedious and to be avoided if at all possible. Nowadays, computers make it just as easy to do the exact test or randomization test as the computationally simpler chi-square or G-test. I recommend that you use the exact test or randomization test when the total sample size is less than 1000. See the web page on small sample sizes for further discussion.

The exact test and randomization test should give you the same result, if you do enough replicates for the randomization test, so the choice between them is a matter of personal preference. The exact test sounds more "exact"; the randomization test may be easier to understand and explain. You can do the randomization test with a spreadsheet, web page, or simple homemade computer program; the exact test may require a sophisticated statistics program such as SAS.

If some of your expected numbers are too small to use the chi-square or G-test, but your total sample size is too big for an exact test, you may have to use a randomization test.

How to do the test

Spreadsheet

I've put together a spreadsheet that will perform the randomization test of goodness-of-fit for up to 10 categories and up to 100 observations. It does 200 replicates at a time; to get a decent number of replicates, you should copy the numbers from the cell labelled "reps. with greater chi-sq." into the row labelled "enter reps. with greater chi-sq." By entering these numbers in this row alongside each other, you can do up to 10,000 total replicates.

Web page

Go to this web page (http://faculty.vassar.edu/lowry/csfit.html) and enter your observed numbers in the first column. If you had 5 red flowers, 2 pink flowers and 1 white flower in a genetic cross, you would enter those numbers. You may then either enter the expected numbers (2, 4 and 2) in the second column, or enter the expected proportions (0.25, 0.50, and 0.25) in the third column.

Hit the Calculate button, and you'll get the chi-square test results; in this case, the chi-square statistic is 6.00. Then scroll down and hit the 200 Random Samples button. The web page will then use a random-number generator to choose a flower at random. The probability of it choosing a red flower is 0.25, a pink flower 0.50, and a white flower 0.25. It will do this for 8 flowers, then calculate the chi-square statistic for this simulated data set. Then it will repeat this, for a total of 200 simulated data sets. You'll probably want to do more that 200 replicates, to get a more accurate estimate of the P value; you should do at least 1000, by hitting the 200 Random Samples button four more times, or maybe 10,000 if you want publication-quality data.

SAS

To conduct a randomization test of goodness-of-fit in SAS, use the TABLES and EXACT commands with the CHISQ and MC options. Here's an example using the snapdragons. The `testp=(25 50 25)` option gives the expected percentages, which must add up to 100; you could also use the proportions, 0.25, 0.50 and 0.25. The `order=data` option tells SAS that the expected values are given for the values of "color" in the order they are input (red, pink, white).

```
data snapdragons;
   input color $ observed;
   cards;
red   5
pink  2
white 1
;
proc freq data=snapdragons order=data;
   weight observed;
   tables color / chisq testp=(25 50 25);
```

```
   exact chisq / mc n=100000;
run;
```

The output includes two p-values, one for the regular chi-square test and one for the randomization test.

```
          Chi-Square Test
     for Specified Proportions
-------------------------------------
Chi-Square                      6.0000
DF                                   2
Asymptotic Pr > ChiSq           0.0498 Chi-square P-value

Monte Carlo Estimate for the Exact Test

 Pr >= ChiSq                    0.0594 Randomization P-value
 99% Lower Conf Limit           0.0575
 99% Upper Conf Limit           0.0613

 Number of Samples              100000
```

Power analysis

I don't know how to do a precise power analysis for this test. Unless the numbers are very, very small, the P-values are fairly similar to those for a chi-square test of goodness-of-fit, so the power analysis described there should be adequate.

Further reading

Sokal and Rohlf, p. 810.

Chi-square test of independence

The chi-square test may be used both as a test of goodness-of-fit (comparing frequencies of one nominal variable to theoretical expecations) and as a test of independence (comparing frequencies of one nominal variable for different values of a second nominal variable). The underlying arithmetic of the test is the same; the only difference is the way the expected values are calculated. However, goodness-of-fit tests and tests of independence are used for quite different experimental designs and test different null hypotheses, so I treat the chi-square test of goodness-of-fit and the chi-square test of independence as two distinct statistical tests.

The chi-square test of independence is an alternative to the G-test of independence. Most of the information on this page is identical to that on the G-test page. You should read the section on "Chi-square vs. G-test", pick either chi-square or G-test, then stick with that choice for the rest of your life.

When to use it

The chi-squared test of independence is used when you have two nominal variables, each with two or more possible values. A data set like this is often called an "R×C table," where R is the number of rows and C is the number of columns. For example, if you surveyed the frequencies of three flower phenotypes (red, pink, white) in four geographic locations, you would have a 3×4 table. You could also consider it a 4×3 table; it doesn't matter which variable is the columns and which is the rows.

It is also possible to do a chi-squared test of independence with more than two nominal variables, but that experimental design doesn't occur very often and is rather complicated to analyze and interpret, so I won't cover it (except for the special case of repeated 2x2 tables, analyzed with the Cochran-Mantel-Haenszel test).

Null hypothesis

The null hypothesis is that the relative proportions of one variable are independent of the second variable; in other words, the proportions at one variable are the same for different values of the second variable. In the flower example, the null hypothesis is that the proportions of red, pink and white flowers are the same at the four geographic locations.

For some experiments, you can express the null hypothesis in two different ways, and either would make sense. For example, when an individual clasps their hands, there is one comfortable position; either the right thumb is on top, or the left thumb is on top. Downey (1926) collected data on the frequency of right-thumb vs. left-thumb clasping in right-handed and left-handed individuals. You could say that the null hypothesis is that the proportion of right-thumb-clasping is the same for right-handed and left-handed individuals, or you could say that the proportion of right-handedness is the same for right-thumb-clasping and left-thumb-clasping individuals.

For other experiments, it only makes sense to express the null hypothesis one way. In the flower example, it would make sense to say that the null hypothesis is that the proportions of red, pink and white flowers are the same at the four geographic locations; it wouldn't make sense to say that the proportions of locations are the same for red, pink, and white flowers.

How the test works

The math of the chi-square test of independence is the same as for the chi-square test of goodness-of-fit, only the method of calculating the expected frequencies is different. For the goodness-of-fit test, a theoretical relationship is used to calculate the expected frequencies. For the test of independence, only the observed frequencies are used to calculate the expected. For the hand-clasping example, Downey (1926) found 190 right-thumb and 149 left-thumb-claspers among right-handed women, and 42 right-thumb and 49 left-thumb-claspers among left-handed women. To calculate the estimated frequency of right-thumb-claspers among right-handed women, you would first calculate the overall proportion of right-thumb-claspers: (190+42)/(190+42+149+49)=0.5395. Then you would multiply this overall proportion times the total number of right-handed women, 0.5395×(190+149)=182.9. This is the expected number of right-handed right-thumb-claspers under the null hypothesis; the observed number is 190. Similar calculations would be done for each of the cells in this 2×2 table of numbers.

The degrees of freedom in a test of independence are equal to (number of rows)−1 × (number of columns)−1. Thus for a 2×2 table, there are (2−1)×(2−1)=1 degree of freedom; for a 4×3 table, there are (4−1)×(3−1)=6 degrees of freedom.

Examples

Gardemann et al. (1998) surveyed genotypes at an insertion/deletion polymorphism of the apolipoprotein B signal peptide in 2259 men. Of men without coronary artery disease, 268 had the ins/ins genotype, 199 had the ins/del genotype, and 42 had the del/del genotype. Of men with coronary artery disease, there were 807 ins/ins, 759 ins/del, and 184 del/del.

The two nominal variables are genotype (ins/ins, ins/del, or del/del) and disease (with or without). The biological null hypothesis is that the apolipoprotein polymorphism doesn't affect the likelihood of getting coronary artery disease. The statistical null hypothesis is that the proportions of men with coronary artery disease are the same for each of the three genotypes.

The result is chi^2=7.26, 2 d.f., P=0.027. This indicates that the null hypothesis can be rejected; the three genotypes have significantly different proportions of men with coronary artery disease.

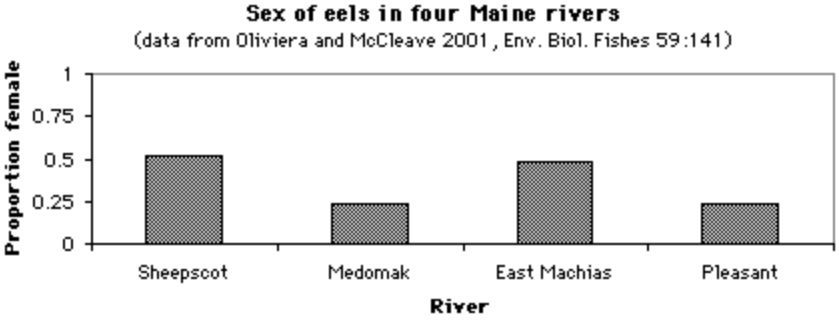
Young and Winn (2003) counted sightings of the spotted moray eel, *Gymnothorax moringa*, and the purplemouth moray eel, *G. vicinus*, in a 150-m by 250-m area of reef in Belize. They identified each eel they saw, and classified the locations of the sightings into three types: those in grass beds, those in sand and rubble, and those within one meter of the border between grass and sand/rubble. The number of sightings are shown in the table, with percentages in parentheses:

	G. moringa	G. vicinus
Grass	127 (25.9)	116 (33.7)
Sand	99 (20.2)	67 (19.5)
Border	264 (53.9)	161 (46.8)

The nominal variables are the species of eel (*G. moringa* or *G. vicinus*) and the habitat type (grass, sand, or border). The difference in habitat use between the species is significant (chi-square=6.26, 2 d.f., P=0.044).

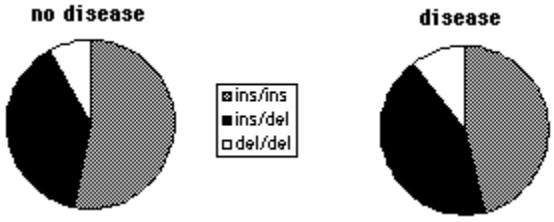
Graphing the results

The data used in a test of independence are usually displayed with a bar graph, with the values of one variable on the X-axis and the proportions of the other variable on the Y-axis. If the variable on the Y-axis only has two values, you only need to plot one of them:



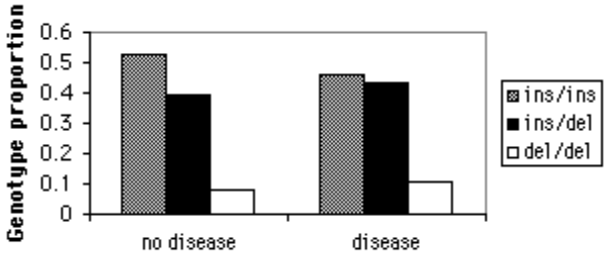
A bar graph for when the nominal variable has only two values.

If the variable on the Y-axis has more than two values, you should plot all of them. Sometimes pie charts are used for this:



A pie chart for when the nominal variable has more than two values.

But as much as I like pie, I think pie charts make it difficult to see small differences in the proportions, and difficult to show confidence intervals. In this situation, I prefer bar graphs:



A bar graph for when the nominal variable has more than two values.

Similar tests

There are several tests that use chi-square statistics. The one described here is formally known as Pearson's chi-square. It is by far the most common chi-square test, so it is usually just called the chi-square test.

If the expected numbers in some classes are small, the chi-squared test will give inaccurate results. In that case, you should try Fisher's exact test; if that doesn't work (because the total sample size is too big, or because there are too many values of one of the nominal variables), you can use the

randomization test of independence. See the web page on small sample sizes for further discussion.

If the samples are not independent, but instead are before-and-after observations on the same individuals, you should use McNemar's test.

Chi-square vs. G-test

The chi-square test gives approximately the same results as the G-test. Unlike the chi-square test, G-values are additive, which means they can be used for more elaborate statistical designs. G-tests are a subclass of likelihood ratio tests, a general category of tests that have many uses for testing the fit of data to mathematical models; the more elaborate versions of likelihood ratio tests don't have equivalent tests using the Pearson chi-square statistic. The G-test is therefore preferred by many, even for simpler designs. On the other hand, the chi-square test is more familiar to more people, and it's always a good idea to use statistics that your readers are familiar with when possible. You may want to look at the literature in your field and see which is more commonly used.

How to do the test

Spreadsheet

I have set up a spreadsheet that performs this test for up to 10 columns and 50 rows. It is largely self-explanatory; you just enter you observed numbers, and the spreadsheet calculates the chi-squared test statistic, the degrees of freedom, and the P-value.

Web page

There are many web pages that do chi-squared tests of independence, but most are limited to fairly small numbers of rows and columns. One page that will handle large data sets is here (http://department.obg.cuhk.edu.hk/researchsupport/RxC_contingency_table.asp) . Robert Huber has put together a web page that will do a chi-squared test of independence for up to a 10×10 table. Be sure to scroll to the bottom of the page and set the number of rows and columns.

SAS

Here is a SAS program that uses PROC FREQ for a chi-square test. It uses the handclasping data from above.

```
data handclasp;
   input thumb $ hand $ count;
   cards;
rightthumb righthand 190
leftthumb  righthand 149
rightthumb lefthand   42
leftthumb  lefthand   49
;
proc freq data=handclasp;
   weight count / zeros;
   tables thumb*hand / chisq;
run;
```

The output includes the following:

```
         Statistics for Table of thumb by hand

Statistic                          DF       Value      Prob
------------------------------------------------------------
Chi-Square                          1      2.8265    0.0927
Likelihood Ratio Chi-Square         1      2.8187    0.0932
Continuity Adj. Chi-Square          1      2.4423    0.1181
Cochran-Mantel-Haenszel Chi-Square  1      2.8199    0.0931
Phi Coefficient                                      0.0811
Contingency Coefficient                              0.0808
Cramer's V                                           0.0811
```

The "Chi-Square" on the first line is the P-value for the chi-square test; in this case, chi-square=2.8265, 1 d.f., P=0.0927.

Power analysis

For a 2×2 table, you can use the technique described for Fisher's exact test, even if the resulting sample size will be much too large to actually do Fisher's exact test.

For a test with more than 2 rows or columns, the G*Power (http://www.psycho.uni-duesseldorf.de/abteilungen/aap/gpower3/) program will calculate the sample size needed for a test of independence, but you need to calculate the effect size parameter, *w*, separately. The chi-squared test of independence spreadsheet can be used for this. Enter the data you hope to see; you can enter proportions, percentages, or raw numbers. Then go to G*Power and choose "Chi-squared tests" from the "Test family" menu and "Goodness-of-fit tests: Contingency tables" from the "Statistical test" menu. Copy the "Effect size w" from the spreadsheet to the G*Power form, then enter your alpha (usually 0.05), your power (often 0.8 or 0.9), and your degrees of freedom (for a test with R rows and C columns, remember that degrees of freedom is (R−1)×(C−1)). This analysis assumes that your total sample will be divided equally among the groups; if it isn't, you'll need a larger sample size than the one you estimate.

As an example, let's say you're looking for a relationship between bladder cancer and genotypes at a polymorphism in the catechol-O-methyltransferase gene in humans. In the population you're studying, you know that the genotype frequencies in people without bladder cancer are 0.36 GG, 0.48 GA, and 0.16 AA; you want to know how many people with bladder cancer you'll have to genotype to get a significant result if they have 6 percent more AA genotypes. Enter 0.36, 0.48, and 0.16 in the first column of the spreadsheet, and 0.33, 0.45, and 0.22 in the second column; the effect size (w) is 0.10838. Enter this in the G*Power page, enter 0.05 for alpha, 0.80 for power, and 2 for degrees of freedom. The result is a total sample size of 821, so you'll need 411 people with bladder cancer and 411 people without bladder cancer.

Further reading

Sokal and Rohlf, pp. 736-737.

Zar, pp. 486-500.

References

Downey, J.E. 1926. Further observations on the manner of clasping the hands. American Naturalist 60: 387-391.

Gardemann, A., D. Ohly, M. Fink, N. Katz, H. Tillmanns, F.W. Hehrlein, and W. Haberbosch. 1998. Association of the insertion/deletion gene polymorphism of the apolipoprotein B signal peptide with myocardial infarction. Atheroslerosis 141: 167-175.

Young, R.F., and H.E. Winn. 2003. Activity patterns, diet, and shelter site use for two species of moray eels, *Gymnothorax moringa* and *Gymnothorax vicinus*, in Belize. Copeia 2003: 44-55.

G-test of independence

The G-test may be used both as a test of goodness-of-fit (comparing frequencies of one nominal variable to theoretical expecations) and as a test of independence (comparing frequencies of one nominal variable for different values of a second nominal variable). The underlying arithmetic of the test is the same. Goodness-of-fit tests and tests of independence are used for quite different experimental designs and test different null hypotheses, so I treat the G-test of goodness-of-fit and the G-test of independence as two distinct statistical tests.

The G-test of independence is an alternative to the chi-square test of independence. Most of the information on this page is identical to that on the chi-square page. You should read the section on "Chi-square vs. G-test", pick either chi-square or G-test, then stick with that choice for the rest of your life.

When to use it

The G-test of independence is used when you have two nominal variables, each with two or more possible values. A data set like this is often called an "R×C table," where R is the number of rows and C is the number of columns. For example, if you surveyed the frequencies of three flower phenotypes (red, pink, white) in four geographic locations, you would have a 3×4 table. You could also consider it a 4×3 table; it doesn't matter which variable is the columns and which is the rows.

It is also possible to do a G-test of independence with more than two nominal variables, but that experimental design doesn't occur very often and is rather complicated to analyze and interpret, so I won't cover it.

Null hypothesis

The null hypothesis is that the relative proportions of one variable are independent of the second variable; in other words, the proportions at one variable are the same for different values of the second variable. In the flower example, you would probably say that the null hypothesis was that the proportions of red, pink and white were the same at the four locations.

For some experiments, you can express the null hypothesis in two different ways, and either would make sense. For example, when an individual clasps their hands, there is one comfortable position; either the right thumb is on top, or the left thumb is on top. Downey (1926) collected data on the frequency of right-thumb

vs. left-thumb clasping in right-handed and left-handed individuals. You could say that the null hypothesis is that the proportion of right-thumb-clasping is the same for right-handed and left-handed individuals, or you could say that the proportion of right-handedness is the same for right-thumb-clasping and left-thumb-clasping individuals.

For other experiments, it only makes sense to express the null hypothesis one way. In the flower example, it would make sense to say that the null hypothesis is that the proportions of red, pink and white flowers are the same at the four geographic locations; it wouldn't make sense to say that the proportion of flowers at each location is the same for red, pink, and white flowers.

How the test works

The math of the G-test of independence is the same as for the G-test of goodness-of-fit, only the method of calculating the expected frequencies is different. For the goodness-of-fit test, a theoretical relationship is used to calculate the expected frequencies. For the test of independence, only the observed frequencies are used to calculate the expected. For the hand-clasping example, Downey (1926) found 190 right-thumb and 149 left-thumb-claspers among right-handed women, and 42 right-thumb and 49 left-thumb-claspers among left-handed women. To calculate the estimated frequency of right-thumb-claspers among right-handed women, you would first calculate the overall proportion of right-thumb-claspers: (190+42)/(190+42+149+49)=0.5395. Then you would multiply this overall proportion times the total number of right-handed women, 0.5395×(190+149)=182.9. This is the expected number of right-handed right-thumb-claspers under the null hypothesis; the observed number is 190. Similar calculations would be done for each of the cells in this 2×2 table of numbers.

(In practice, the calculations for the G-test of independence use shortcuts that don't require calculating the expected frequencies; see Sokal and Rohlf, pp. 731-732.)

The degrees of freedom in a test of independence are equal to (number of rows)−1 × (number of columns)−1. Thus for a 2×2 table, there are (2−1)×(2−1)=1 degree of freedom; for a 4×3 table, there are (4−1)×(3−1)=6 degrees of freedom.

Examples

Gardemann et al. (1998) surveyed genotypes at an insertion/deletion polymorphism of the apolipoprotein B signal peptide in 2259 men. Of men without coronary artery disease, 268 had the ins/ins genotype, 199 had the ins/del genotype, and 42 had the del/del genotype. Of men with coronary artery disease, there were 807 ins/ins, 759 ins/del, and 184 del/del.

The biological null hypothesis is that the apolipoprotein polymorphism doesn't affect the likelihood of getting coronary artery disease. The statistical null

hypothesis is that the proportions of men with coronary artery disease are the same for each of the three genotypes.

The result is G=7.30, 2 d.f., P=0.026. This indicates that the null hypothesis can be rejected; the three genotypes have significantly different proportions of men with coronary artery disease.

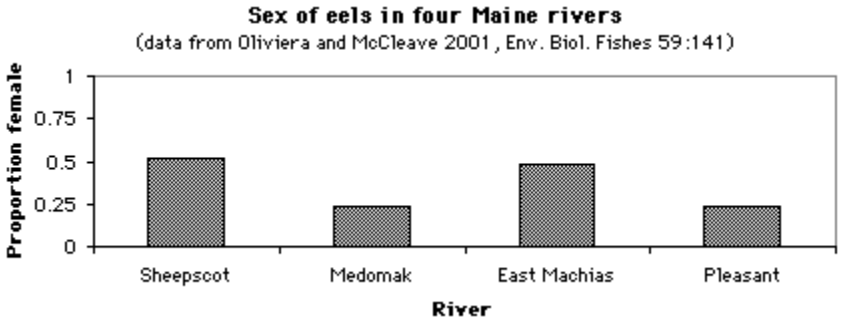
Young and Winn (2003) counted sightings of the spotted moray eel, *Gymnothorax moringa*, and the purplemouth moray eel, *G. vicinus*, in a 150-m by 250-m area of reef in Belize. They identified each eel they saw, and classified the locations of the sightings into three types: those in grass beds, those in sand and rubble, and those within one meter of the border between grass and sand/rubble. The number of sightings are shown in the table, with percentages in parentheses:

	G. moringa	*G. vicinus*
Grass	127 (25.9)	116 (33.7)
Sand	99 (20.2)	67 (19.5)
Border	264 (53.9)	161 (46.8)

The nominal variables are the species of eel (*G. moringa* or *G. vicinus* and the habitat type (grass, sand, or border). The difference in habitat use between the species is significant (G=6.23, 2 d.f., P=0.044).

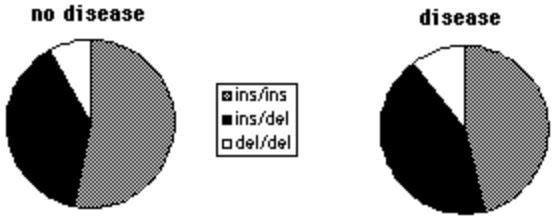
Graphing the results

The data used in a test of independence are usually displayed with a bar graph, with the values of one variable on the X-axis and the proportions of the other variable on the Y-axis. If the variable on the Y-axis only has two values, you only need to plot one of them:



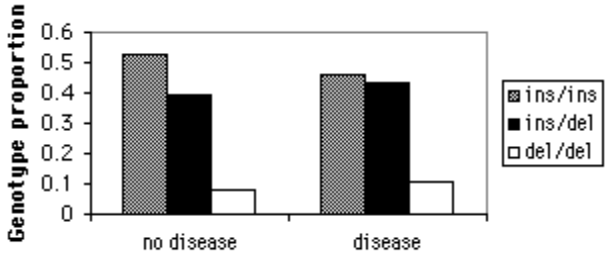
A bar graph for when the nominal variable has only two values.

If the variable on the Y-axis has more than two values, you should plot all of them. Sometimes pie charts are used for this:



A pie chart for when the nominal variable has more than two values.

But as much as I like pie, I think pie charts make it difficult to see small differences in the proportions, and difficult to show error bars. In this situation, I prefer bar graphs:



A bar graph for when the nominal variable has more than two values.

Similar tests

If the expected numbers in some classes are small, the G-test will give inaccurate results. In that case, you should try Fisher's exact test; if that doesn't work (because the total sample size is too big, or because there are too many values of one of the nominal variables), you can use the randomization test of independence. See the web page on small sample sizes for further discussion.

If the samples are not independent, but instead are before-and-after observations on the same individuals, you should use McNemar's test.

Chi-square vs. G-test

The chi-square test gives approximately the same results as the G-test. Unlike the chi-square test, G-values are additive, which means they can be used for more elaborate statistical designs. G-tests are a subclass of likelihood ratio tests, a general category of tests that have many uses for testing the fit of data to mathematical models; the more elaborate versions of likelihood ratio tests don't have equivalent tests using the Pearson chi-square statistic. The G-test is therefore preferred by many, even for simpler designs. On the other hand, the chi-square test is more familiar to more people, and it's always a good idea to use statistics

that your readers are familiar with when possible. You may want to look at the literature in your field and see which is more commonly used.

How to do the test

Spreadsheet

I have set up an Excel spreadsheet that performs this test for up to 10 columns and 50 rows. It is largely self-explanatory; you just enter you observed numbers, and the spreadsheet calculates the G-test statistic, the degrees of freedom, and the P-value.

Web pages

There is a web page that will do a G-test of independence (http://caspar.bgsu.edu/~software/Java/1Contingency.html) for up to a 10×10 table. Be sure to scroll to the bottom of the page and set the number of rows and columns.

SAS

Here is a SAS program that uses PROC FREQ for a G-test. It uses the handclasping data from above.

```
data handclasp;
   input thumb $ hand $ count;
   cards;
rightthumb righthand 190
leftthumb  righthand 149
rightthumb lefthand   42
leftthumb  lefthand   49
;
proc freq data=handclasp;
   weight count / zeros;
   tables thumb*hand / chisq;
run;
```

The output includes the following:

```
        Statistics for Table of thumb by hand
Statistic                          DF       Value      Prob
------------------------------------------------------------
Chi-Square                          1      2.8265    0.0927
Likelihood Ratio Chi-Square         1      2.8187    0.0932
Continuity Adj. Chi-Square          1      2.4423    0.1181
Cochran-Mantel-Haenszel Chi-Square  1      2.8199    0.0931
Phi Coefficient                                      0.0811
Contingency Coefficient                              0.0808
Cramer's V                                           0.0811
```

The "Likelihood Ratio Chi-Square" is the P-value for the G-test; in this case, G=2.8187, 1 d.f., P=0.0932.

Power analysis

If each nominal variable has just two values (a 2×2 table), use the power analysis for Fisher's exact test.

If either nominal variable has more than two values, use the power analysis for chi-squared tests of independence.

Further reading

Sokal and Rohlf, pp. 729-739.

Zar, pp. 505-506.

References

Downey, J.E. 1926. Further observations on the manner of clasping the hands. American Naturalist 60: 387-391.

Gardemann, A., D. Ohly, M. Fink, N. Katz, H. Tillmanns, F.W. Hehrlein, and W. Haberbosch. 1998. Association of the insertion/deletion gene polymorphism of the apolipoprotein B signal peptide with myocardial infarction. Atheroslerosis 141: 167-175.

Young, R.F., and H.E. Winn. 2003. Activity patterns, diet, and shelter site use for two species of moray eels, *Gymnothorax moringa* and *Gymnothorax vicinus*, in Belize. Copeia 2003: 44-55.

Fisher's exact test of independence

When to use it

Fisher's exact test is used when you have two nominal variables. A data set like this is often called an "R×C table," where R is the number of rows and C is the number of columns. Fisher's exact test is more accurate than the chi-squared test or G-test of independence when the expected numbers are small. See the web page on small sample sizes for further discussion.

The most common use of Fisher's exact test is for 2×2 tables, so that's mostly what I'll describe here. You can do Fisher's exact test for greater than two rows and columns.

Null hypothesis

The null hypothesis is that the relative proportions of one variable are independent of the second variable. For example, if you counted the number of male and female mice in two barns, the null hypothesis would be that the proportion of male mice is the same in the two barns.

How it works

The hypogeometric distribution is used to calculate the probability of getting the observed data, and all data sets with more extreme deviations, under the null hypothesis that the proportions are the same. For example, if one barn has 3 male and 7 female mice, and the other barn has 15 male and 5 female mice, the probability of getting 3 males in the first barn and 15 males in the second, or 2 and 16, or 1 and 17, or 0 and 18, is calculated. For the usual two-tailed test, the probability of getting deviations as extreme as the observed, but in the opposite direction, is also calculated. This is an exact calculation of the probability; unlike most statistical tests, there is no intermediate step of calculating a test statistic whose probability is approximately known.

When there are more than two rows or columns, you have to decide how you measure deviations from the null expectation, so you can tell what data sets would

be more extreme than the observed. The usual method is to calculate the chi-square statistic (formally, it's the Pearson chi-square statistic) for each possible set of numbers, and those with chi-square values equal to or greater than the observed data are considered as extreme as the observed data.

(Note—Fisher's exact test assumes that the row and column totals are fixed. An example would be putting 12 female hermit crabs and 9 male hermit crabs in an aquarium with 7 red snail shells and 14 blue snail shells, then counting how many crabs of each sex chose each color. The total number of female crabs is fixed at 12, and the total number of male crabs, red shells, and blue shells are also fixed. There are few biological experiments where this assumption is true. In the much more common design, the row totals and/or column totals are free to vary. For example, if you took a sample of mice from two barns and counted the number of males and females, you wouldn't know the total number of male mice before doing the experiment; it would be free to vary. In this case, the Fisher's exact test is not, strictly speaking, exact. It is still considered to be more accurate than the chi-square or G-test, and you should feel comfortable using it for any test of independence with small numbers.)

Examples

McDonald and Kreitman (1991) sequenced the alcohol dehydrogenase gene in several individuals of three species of *Drosophila.* Varying sites were classified as synonymous (the nucleotide variation does not change an amino acid) or amino acid replacements, and they were also classified as polymorphic (varying within a species) or fixed differences between species. The two nominal variables are thus synonymicity ("synonymous" or "replacement") and fixity ("polymorphic" or "fixed"). In the absence of natural selection, the ratio of synonymous to replacement sites should be the same for polymorphisms and fixed differences. There were 43 synonymous polymorphisms, 2 replacement polymorphisms, 17 synonymous fixed differences, and 7 replacement fixed differences.

	synonymous	**replacement**
polymorphisms	43	2
fixed	17	7

The result is P=0.0067, indicating that the null hypothesis can be rejected; there is a significant difference in synonymous/replacement ratio between polymorphisms and fixed differences.

The eastern chipmunk trills when pursued by a predator, possibly to warn other chipmunks. Burke da Silva et al. (2002) released chipmunks either 10 or 100 meters from their home burrow, then chased them (to simulate predator pursuit). Out of 24 female chipmunks released 10 m from their burrow, 16 trilled and 8 did not trill. When released 100 m from their burrow, only 3 female chipmunks trilled,

while 18 did not trill. Applying Fisher's exact test, the proportion of chipmunks trilling is signficantly higher (P=0.0007) when they are closer to their burrow.

Descamps et al. (2009) tagged 50 king penguins (*Aptenodytes patagonicus*) in each of three nesting areas (lower, middle, and upper) on Possession Island in the Crozet Archipelago, then counted the number that were still alive a year later. Seven penguins had died in the lower area, six had died in the middle area, and only one had died in the uppper area. Descamps et al. analyzed the data with a G-test of independence, yielding a significant (P=0.048) difference in survival among the areas; however, analyzing the data with Fisher's exact test yields a non-significant (P=0.090) result.

Custer and Galli (2002) flew a light plane to follow great blue herons (*Ardea herodias*) and great egrets (*Casmerodius albus*) from their resting site to their first feeding site at Peltier Lake, Minnesota, and recorded the type of substrate each bird landed on.

	Heron	Egret
Vegetation	15	8
Shoreline	20	5
Water	14	7
Structures	6	1

Fisher's exact test yields P=0.54, so there is no evidence that the two species of birds use the substrates in different proportions.

Graphing the results

You plot the results of Fisher's exact test the same way would any other test of independence.

Similar tests

The chi-squared test of independence or the G-test of independence may be used on the same kind of data as Fisher's exact test. When some of the expected values are small, Fisher's exact test is more accurate than the chi-squared or G-test of independence. If all of the expected values are very large, Fisher's exact test becomes computationally impractical; fortunately, the chi-squared or G-test will then give an accurate result. See the web page on small sample sizes for further discussion.

If the number of rows, number of columns, or total sample size become too large, the program you're using may not be able to perform the calculations for Fisher's exact test in a reasonable length of time, or it may fail entirely. If Fisher's doesn't work, you can use the randomization test of independence.

McNemar's test is used when the two samples are not independent, but instead are two sets of observations on the same individuals. For example, let's say you have 92 children who don't like broccoli and 77 children who like broccoli. You give then your new BroccoYum™ pills for a week, then observe that 14 of the children switched from not liking broccoli before taking the pills to liking broccoli after taking the pills. Three of the children switched in the opposite direction (from liking broccoli to not liking broccoli), and the remaining children stayed the same. The statistical null hypothesis is that the number of switchers in one direction is equal to the number of switchers in the opposite direction. McNemar's test compares the observed data to the null expectation using a goodness-of-fit test. The numbers are almost always small enough that you can make this comparison using the exact binomial test. For the example data of 14 switchers in one direction and 3 in the other direction, P=0.013.

How to do the test

Spreadsheet

I've written a spreadsheet to perform Fisher's exact test for 2×2 tables. It handles samples with the smaller column total less than 500. [An earlier version of this spreadsheet gave slightly inaccurate P-values for small sample sizes. I fixed it on July 4, 2009. Thanks to Patrick Spagon for pointing out the error.]

Web pages

Several people have created web pages that perform Fisher's exact test for 2×2 tables. I like Øyvind Langsrud's web page for Fisher's exact test (http://www.matforsk.no/ola/fisher.htm) . Just enter the numbers into the cells on the web page, hit the Compute button, and get your answer. You should almost always use the "2-tail p-value" given by the web page.

There is also a web page for Fisher's exact test for up to 6×6 tables (http://www.physics.csbsju.edu/stats/exact_NROW_NCOLUMN_form.html) . It will only take data with fewer than 100 observations in each cell.

SAS

Here is a SAS program that uses PROC FREQ for a Fisher's exact test. It uses the chipmunk data from above.

```
data chipmunk;
   input distance $ sound $ count;
   cards;
10m  trill   16
10m  notrill  8
100m trill    3
100m notrill 18
;
proc freq data=chipmunk;
```

```
   weight count / zeros;
   tables distance*sound / chisq;
run;
```

The output includes the following:

```
       Fisher's Exact Test
---------------------------------
Cell (1,1) Frequency (F)        18
Left-sided Pr <= F          1.0000
Right-sided Pr >= F      4.321E-04

Table Probability (P)    4.012E-04
Two-sided Pr <= P        6.862E-04
```

The "Two-sided Pr <= P" is the two-tailed P-value that you want.

SAS automatically does Fisher's exact test for 2×2 tables. For greater numbers of rows or columns, you add a line saying `exact chisq;`. Here is an example using the data on heron and egret substrate use from above:

```
data birds;
   input bird $ substrate $ count;
   cards;
heron vegetation 15
heron shoreline  20
heron water      14
heron structures  6
egret vegetation  8
egret shoreline   5
egret water       7
egret structures  1
;
proc freq data=birds;
   weight count / zeros;
   tables bird*substrate / chisq;
   exact chisq;
run;
```

The results of the exact test are labelled "Exact Pr >= ChiSq"; in this case, P=0.5357.

```
     Pearson Chi-Square Test
---------------------------------
Chi-Square                 2.2812
DF                              3
Asymptotic Pr >  ChiSq     0.5161
Exact      Pr >= ChiSq     0.5357
```

Power analysis

The G*Power program will calculate the sample size needed for Fisher's exact test. Choose "Exact" from the "Test family" menu and "Proportions: Inequality, two independent groups (Fisher's exact test)" from the "Statistical test" menu. Enter the proportions you hope to see, your alpha (usually 0.05) and your power (usually 0.80 or 0.90). If you plan to have more observations in one group than in the other, you can make the "Allocation ratio" different from 1.

As an example, let's say you're looking for a relationship between bladder cancer and genotypes at the catechol-O-methyltransferase gene in humans. Based on previous research, you're going to pool together the GG and GA genotypes and compare these "GG+GA" and AA genotypes. In the population you're studying, you know that the genotype frequencies in people without bladder cancer are 0.84 GG+GA and 0.16 AA; you want to know how many people with bladder cancer you'll have to genotype to get a significant result if they have 6 percent more AA genotypes. It's easier to find controls than people with bladder cancer, so you're planning to have twice as many people without bladder cancer. On the G*Power page, enter 0.16 for proportion p1, 0.22 for proportion p2, 0.05 for alpha, 0.80 for power, and 0.5 for allocation ratio. The result is a total sample size of 1523, so you'll need 508 people with cancer and 1016 people without cancer.

Note that the sample size will be different if your effect size is a 6 percent lower frequency of AA in bladder cancer patients, instead of 6 percent higher. If you don't have a strong idea about which direction of difference you're going to see, you should do the power analysis both ways and use the larger sample size.

Further reading

Sokal and Rohlf, pp. 734-736.

Zar, pp. 543-555.

References

Burke da Silva, K., C. Mahan, and J. da Silva. 2002. The trill of the chase: eastern chipmunks call to warn kin. J. Mammol. 83: 546-552.

Custer, C.M., and J. Galli. 2002. Feeding habitat selection by great blue herons and great egrets nesting in east central Minnesota. Waterbirds 25: 115-124.

Descamps, S., C. le Bohec, Y. le Maho, J.-P. Gendner, and M. Gauthier-Clerc. 2009. Relating demographic performance to breeding-site location in the king penguin. Condor 111: 81-87.

McDonald, J.H. and M. Kreitman. 1991. Adaptive protein evolution at the *Adh* locus in *Drosophila*. Nature 351: 652-654.

Randomization test of independence

When to use it

The randomization test of independence is used when you have two nominal variables. A data set like this is often called an "R×C table," where R is the number of rows and C is the number of columns. The randomization test is more accurate than the chi-squared test or G-test of independence when the expected numbers are small. See the web page on small sample sizes for further discussion.

Fisher's exact test would be just as good as a randomization test, but there may be situations where the computer program you're using can't handle the calculations required for the Fisher's test.

Null hypothesis

The null hypothesis is that the relative proportions of one variable are independent of the second variable. For example, if you counted the number of male and female mice in two barns, the null hypothesis would be that the proportion of male mice is the same in the two barns.

How it works

Fisher's exact test works by calculating the probabilities of all possible combinations of numbers in an R×C table, then adding the probabilities of those combinations that are as extreme or more extreme than the observed data. As R and C get larger, and as the total sample size gets larger, the number of possible combinations increases dramatically, to the point where a computer may have a hard time doing all the calculations in a reasonable period of time.

The randomization test works by generating random combinations of numbers in the R×C table, with the probability of generating a particular combination equal to its probability under the null hypothesis. For each combination, the Pearson's chi-square statistic is calculated. The proportion of these random combinations that have a chi-square statistic equal to or greater than the observed data is the P-value.

Because it is taking a random sample of all possible combinations, the randomization test will give slightly different estimates of the P-value every time you run it. The more replicates you run, the more accurate your estimate of the P-value will be. You might want to start with a small number of replicates, such as 1,000, to be sure everything is working properly, then change the number of replicates to 100,000 or even 1,000,000 for your final result.

Examples

Custer and Galli (2002) flew a light plane to follow great blue herons (*Ardea herodias*) and great egrets (*Casmerodius albus*) from their resting site to their first feeding site at Peltier Lake, Minnesota, and recorded the type of substrate each bird landed on.

	Heron	Egret
Vegetation	15	8
Shoreline	20	5
Water	14	7
Structures	6	1

A randomization test with 100,000 replicates yields P=0.54, so there is no evidence that the two species of birds use the substrates in different proportions.

Young and Winn (2003) counted prey items in the stomach of the spotted moray eel, *Gymnothorax moringa*, and the purplemouth moray eel, *G. vicinus*. They identified each eel they saw, and classified the locations of the sightings into three types: those in grass beds, those in sand and rubble, and those within one meter of the border between grass and sand/rubble. The number of prey items are shown in the table:

	G. moringa	G. vicinus
Slippery dick	10	6
Unidentified wrasses	3	7
Moray eels	1	1
Squirrelfish	1	1
Unidentified fish	6	3
Oval urn crab	31	10
Emerald crab	3	2
Portunus crab spp.	1	0
Arrow crab	1	0
Unidentified crabs	15	1
Spiny lobster	0	1
Octopus	3	2
Unidentified	4	1

The nominal variables are the species of eel (*G. moringa* or *G. vicinus*) and the prey type. The difference in stomach contents between the species is not significant (randomization test with 100,000 replicates, P=0.11).

There are a lot of small numbers in this data set. If you pool the data into fish (the first five species), crustaceans (crabs and lobster), and octopus+unidentified, the P-value from 100,000 randomizations is 0.029; *G. moringa* eat a higher proportion of crustaceans than *G. vicinus*. Of course, it would be best to decide to pool the data this way before collecting the data. If you decided to pool the numbers after seeing them, you'd have to make it clear that you did that, writing something like "After seeing that many of the numbers were very small when divided into individual species, we also analyzed the data after pooling into fish, crustaceans, and other/unidentified."

Graphing the results

You plot the results of a randomization test the same way would any other test of independence.

Similar tests

The chi-squared test of independence or the G-test of independence may be used on the same kind of data as a randomization test of independence. When some of the expected values are small, Fisher's exact test or the randomization test is more accurate than the chi-squared or G-test of independence. If all of the expected values are very large, Fisher's exact test and the randomization test become computationally impractical; fortunately, the chi-squared or G-test will then give an accurate result. See the web page on small sample sizes for further discussion.

If the number of rows, number of columns, or total sample size become too large, the program you're using may not be able to perform the calculations for Fisher's exact test in a reasonable length of time, or it may fail entirely. I'd try Fisher's test first, then do the randomization test if Fisher's doesn't work.

How to do the test

Spreadsheet

I haven't written a spreadsheet for this test.

Web pages

I don't know of a web page that will perform this test.

SAS

Here is a SAS program that uses PROC FREQ to do the randomization test of independence. The example uses the data on heron and egret substrate use from above. In the statement `exact chisq / mc n=100000`, "mc" tells SAS to do

randomization (also known as Monte Carlo simulation), and "n=100000" tells it how many replicates to run.

```
data birds;
   input bird $ substrate $ count;
   cards;
heron vegetation 15
heron shoreline  20
heron water      14
heron structures  6
egret vegetation  8
egret shoreline   5
egret water       7
egret structures  1
;
proc freq data=birds;
   weight count;
   tables bird*substrate / chisq;
   exact chisq / mc n=100000;
run;
```

The results of the randomization test are labelled "Pr >= ChiSq"; in this case, P=0.5392.

```
Monte Carlo Estimate for the Exact Test

Pr >= ChiSq                 0.5392
99% Lower Conf Limit        0.5351
99% Upper Conf Limit        0.5432

Number of Samples           100000
Initial Seed             952082114
```

Power analysis

Unless your numbers are very small, the power analysis described for the chi-square test of independence should work well enough.

References

Custer, C.M., and J. Galli. 2002. Feeding habitat selection by great blue herons and great egrets nesting in east central Minnesota. Waterbirds 25: 115-124.

Young, R.F., and H.E. Winn. 2003. Activity patterns, diet, and shelter site use for two species of moray eels, *Gymnothorax moringa* and *Gymnothorax vicinus*, in Belize. Copeia 2003: 44-55.

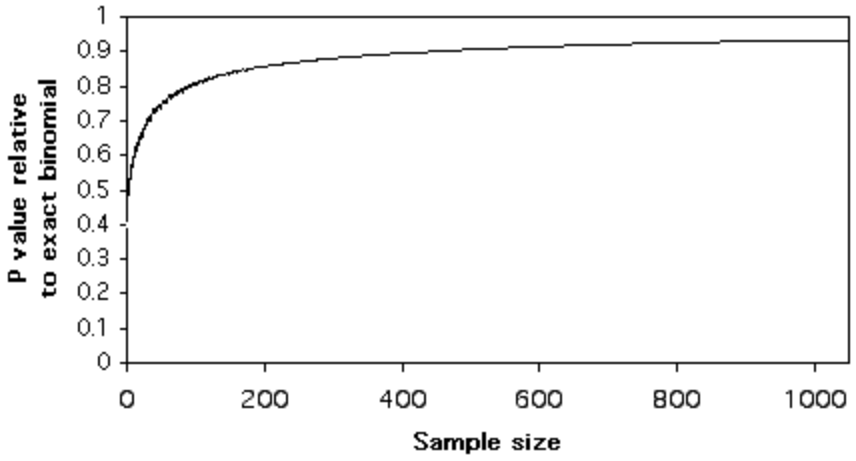
Small numbers in chi-square and G-tests

Chi-square and G-tests of goodness-of-fit or independence give inaccurate results when the expected numbers are small. For example, if you observe 11 people with torn anterior cruciate ligaments, and 9 have torn their right ACL and 2 have torn their left ACL, you would compare the observed ratio to an expected 1:1 ratio to see if there's evidence that people are more likely to tear one ACL than the other. The exact binomial test gives a P-value of 0.065, the chi-square test of goodness-of-fit gives a P-value of 0.035, and the G-test of goodness-of-fit gives a P-value of 0.028. If you analyzed the data using the chi-square or G-test, you would conclude that people tear their right ACL significantly more than their left ACL; if you used the exact binomial test, which is more accurate, the evidence would not be quite strong enough to reject the null hypothesis.

When the sample sizes are too small, alternatives to the chi-square test or G-test are recommended. However, how small is "too small"? The conventional rule of thumb is that if all of the expected numbers are greater than 5, it's acceptable to use the chi-square or G-test; if an expected number is less than 5, you should use an alternative, such as an exact test or a randomization test for goodness-of-fit, or a Fisher's exact test or randomization test of independence.

This rule of thumb is left over from the olden days, when the calculations necessary for an exact test were exceedingly tedious and error-prone, and a randomization test would have required flipping actual coins or rolling actual dice thousands of times. Now that we have these new-fangled gadgets called computers, it's time to retire the "expected less than 5" rule. But what new rule should you use?

Here is a graph of relative P-values versus sample size. For each sample size, a pair of numbers were found that would give a P-value for the exact binomial test (null hypothesis, 1:1 ratio) that was as close as possible to P=0.05 without going under it. For example, with a sample size of 11, the numbers 9 and 2 give a P-value of 0.065. The chi-square test was then done on these numbers, and the chi-square P-value was divided by the exact binomial P-value. For 9 and 2, the chi-square P-value is 0.035, so the ratio is 0.035/0.065 = 0.54. In other words, the chi-square test gives a P-value that is only 54 percent as large as the more accurate exact binomial test. The G-test gives almost the same results as the chi-square test.



P-values of chi-square tests, as a proportion of the P-value from the exact binomial test.

Plotting these relative P-values vs. sample size , it is clear that the chi-square and G-tests give P-values that are too low, even for sample sizes in the hundreds. This means that if you use a chi-square or G-test of goodness-of-fit and the P-value is just barely significant, you will reject the null hypothesis, even though the more accurate P-value of the exact binomial test would be above 0.05. The results are similar for 2×2 tests of independence; the chi-square and G-tests give P-values that are considerably lower than that of the more accurate Fisher's exact test.

Yates' and William's corrections

One solution to this problem is to use Yates' correction for continuity, sometimes just known as the continuity correction. To do this, you subtract 0.5 from each observed value that is greater than the expected, add 0.5 to each observed value that is less than the expected, then do the chi-square or G-test. This only applies to tests with one degree of freedom: goodness-of-fit tests with only two categories, and 2×2 tests of independence. It works quite well for goodness-of-fit, yielding P-values that are quite close to those of the exact binomial. For tests of independence, Yates' correction yields P-values that are too high.

Another correction that is sometimes used is Williams' correction. For a goodness-of-fit test, Williams' correction is found by dividing the chi-square or G values by the following:

$$q=1+(a^2-1)/6nv$$

where a is the number of categories, n is the total sample size, and v is the number of degrees of freedom. For a test of independence with R rows and C columns,

Williams' correction is found by dividing the chi-square or G value by the following:

$$q=1+(n\{[1/(\text{row 1 total})]+\ldots+[1/(\text{row R total})]\}-1)(n\{[1/(\text{column 1 total})]+\ldots[1/(\text{column C total})]\}-1)/\ 6n(R-1)(C-1)$$

Unlike Yates' correction, it can be applied to tests with more than one degree of freedom. For the numbers I've tried, it increases the P-value a little, but not enough to make it very much closer to the more accurate P-value provided by the exact binomial or Fisher's exact test.

Some software may apply the Yates' or Williams' correction automatically. When reporting your results, be sure to say whether or not you used one of these corrections.

Pooling

When a variable has more than two categories, and some of them have small numbers, it often makes sense to pool some of the categories together. For example, let's say you want to compare the proportions of different kinds of ankle injuries in basketball players vs. volleyball players, and your numbers look like this:

	basketball	volleyball
sprains	18	16
breaks	13	5
torn ligaments	9	7
cuts	3	5
puncture wounds	1	3
infections	2	0

The numbers for cuts, puncture wounds, and infections are pretty small, and this will cause the P-value for your test of independence to be inaccurate. Having a large number of categories with small numbers will also decrease the power of your test to detect a significant difference; adding categories with small numbers can't increase the chi-square value or G-value very much, but it does increase the degrees of freedom. It would therefore make sense to pool some categories:

	basketball	volleyball
sprains	18	16
breaks	13	5
torn ligaments	9	7
other injuries	6	8

Depending on the question you're interested in, it might make sense to pool the data further:

	basketball	volleyball
orthopedic injuries	40	28
non-orthopedic injuries	6	8

It is important to make decisions about pooling before analyzing the data. In this case, you might have known, based on previous studies, that cuts, puncture wounds, and infections would be relatively rare and should be pooled. You could have decided before the study to pool all injuries for which the total was 10 or fewer, or you could have decided to pool all non-orthopedic injuries because they're just not biomechanically interesting.

Recommendations

Goodness-of-fit with two categories: Use the exact binomial test for sample sizes of 1000 or less. Spreadsheets, web pages and SAS should have no problem doing the exact binomial test for sample sizes less than 1000, but they may not be able to handle the calculations for larger sample sizes. For sample sizes greater than 1000, use the chi-square or G-test of goodness-of-fit with Yates' correction (unless you are doing a replicated G-test of goodness-of-fit, in which case you must use the G-test without any continuity correction).

Goodness-of-fit with more than two categories: Use exact tests or randomization tests for sample sizes of 1000 or less. Try the exact tests first, but if the program you're using can't handle it, use randomization tests. Use the chi-square or G-test of goodness of fit for sample sizes greater than 1000. Don't use Williams' correction. If the total sample size is greater than 1000, but some expected numbers are small (less than 5), use randomization tests. Consider pooling rare categories.

2×2 test of independence: Use Fisher's exact test for sample sizes up to 1000. Use the chi-square or G-test of independence, with Yates' correction, for sample sizes greater than 1000.

Greater than 2×2 test of independence: Use either an exact test or randomization test for total sample sizes of 1000 or less. Try the exact test first, but if the program you're using can't handle it, use a randomization test. Use a chi-square or G-test of independence, without Williams' correction, for sample sizes greater than 1000.

Further reading

Sokal and Rohlf, pp. 698-703, 729-730.

Zar, pp. 470, 504-505.

Replicated G-tests of goodness-of-fit

When to use it

Sometimes you'll do a goodness-of-fit experiment more than once; for example, you might look at the fit to a 3:1 ratio of a genetic cross in more than one family, or fit to a 1:1 sex ratio in more than one population, or fit to a 1:1 ratio of broken right and left ankles on more than one sports team. One question then is, should you analyze each experiment separately, risking the chance that the small sample sizes will have insufficient power? Or should you pool all the data, risking the chance that the different experiments gave different results? This is when the additive property of the G-test of goodness-of-fit becomes important, because you can do a replicated G-test of goodness-of-fit.

You use the replicated G-test of goodness of fit when you have two nominal variables with two or more values (such as red vs. pink vs. white flowers for one variable), one of the nominal variables represents different replicates of the same experiment (different days, different locations, different pairs of parents), and the observed data are compared with an extrinsic theoretical expectation (such as an expected 1: 2: 1 ratio in a genetic cross). I do not know if this analysis would be appropriate with an intrinsic hypothesis, such as the p^2: 2pq: q^2 Hardy-Weinberg proportions of population genetics.

Null hypotheses

This technique tests four null hypotheses. The first statistical null hypothesis is that the number of observations in each category is equal to that predicted by a biological theory, and the alternative hypothesis is that the observed numbers are different from the expected. This is the same null hypothesis as for a regular G-test of goodness-of-fit. This is tested for each individual experiment. The second null hypothesis is that overall, the data from the individual experiments fit the expectations. This null hypothesis is a bit difficult to grasp, but being able to test it is the main value of doing a replicated G-test of goodness-of-fit. The third null hypothesis is that the relative proportions are the same across the different

experiments; this is the same as the null hypothesis for a G-test of independence. The fourth null hypothesis is that the pooled data set fits the expected proportions.

How to do the test

First, do a G-test of goodness-of-fit for each individual data set. The resulting G-values are the "individual G-values." Also record the number of degrees of freedom for each individual data set; these are the "individual degrees of freedom." Even if nothing else is significant, it is interesting if one or more of these tests are significant.

(Note: Some programs use "continuity corrections," such as the Yates correction or the Williams correction, in an attempt to make G-tests more accurate for small sample sizes. Do not use any continuity corrections when doing a replicated G-test, or the G-values will not add up properly.

Next, add up all the individual G-values to get the "total G-value", and add up the individual degrees of freedom to get the "total degrees of freedom." Use the CHIDIST function in a spreadsheet or online chi-square calculator to find the P value for the total G-value with the total degrees of freedom. For example, if your total G-value is 12.33 and your total degrees of freedom is 6, enter "=CHIDIST(13.43, 6)". The result will be the P-value for the total G; in this case, 0.0367. If it is significant, you can reject one null hypothesis, that all of the data from the different experiments fit the expected ratio, but you cannot tell yet in what way the data are inconsistent with the expected ratio.

Next, add up the number of observations in each class. For the genetic cross example, you would add up the number of red flowers in all the crosses, all the pink flowers, and all the white flowers. Do a G-test of goodness-of-fit on this pooled data set. This gives you the "pooled G-value." The degrees of freedom is the "pooled degrees of freedom," and it is just the number of classes minus one (the same as for a regular goodness-of-fit test). Find the P-value using the CHIDIST function. The P-value for this test tells you whether the pooled data set deviates significantly from the expected ratio.

Finally, subtract the pooled G-value from the total G-value, and subtract the pooled degrees of freedom from the total degrees of freedom. This gives you the "heterogeneity G-value" and "heterogeneity degrees of freedom." Find the P-value using the CHIDIST function. If the heterogeneity G-value is significant, the individual data sets have significantly different ratios from each other. The heterogeneity G is the same as a G-test of independence comparing the different ratios.

Interpretation

If the heterogeneity G-value is not significant, you can accept one null hypothesis (that the replicates have the same ratios), pool the data and treat them

as if they came from one big experiment. Then you can use the pooled G-value to test the null hypothesis that the data fit the expected ratio.

However, if the heterogeneity G-value is significant, you reject the null hypothesis that the replicates have the same ratios. This means that you cannot pool the data and use the pooled G-value to test anything; you shouldn't pool data sets that are significantly different from each other. In this case, you would investigate the individual data sets more thoroughly, starting by looking at the significance of the individual G-values and then using more sophisticated methods that aren't described here (see Sokal and Rohlf 1995, pp. 722-724).

It won't happen very often, but it's possible that neither the heterogeneity G-value nor the pooled G-value will be significant, but the total G-value will. This rather frustrating result would mean that you could reject the hypothesis that the data all fit the theoretical expectation, but you wouldn't know whether it was due to heterogeneity among the data sets, an overall deviation, or some combination of the two.

Examples

The imaginary data set shown below is the result of eight crosses of heterozygous red/white flowers. The expected proportions of offspring are 0.25 red, 0.50 pink, and 0.25 white. The two nominal variables are color (red, pink, or white) and which cross it is.

Cross	Red	Pink	White		G-value	d.f.	P-value
A	28	56	27		0.03	2	0.986
B	29	56	15		5.98	2	0.050
C	23	53	17		2.73	2	0.256
D	30	60	12		11.16	2	0.004
E	29	49	37		3.49	2	0.174
F	27	46	19		1.40	2	0.497
G	32	52	33		1.46	2	0.481
H	32	58	16		6.38	2	0.041
				total G	32.63	16	0.008
pooled	230	430	176	pooled G	7.89	2	0.019
				heterogeneity G	24.74	14	0.037

The total G-value (32.63), found by summing the eight individual G-values, is significant (P=0.008). This means that the data do not fit the expected 1:2:1 proportions in some way. The pooled G-value (7.89), found by doing a G-test of goodness-of-fit on the pooled data (230 red, 430 pink, 176 white), is significant (P=0.019), which might suggest that there is an overall deviation from the expected proportions. However, the heterogeneity G-value (24.74) is also significant

(P=0.037). This means that the eight crosses were significantly different from each other in their ratios of red to pink to white, so it would be incorrect to use the pooled G-value for any hypothesis test.

Connallon and Jakubowski (2009) performed mating competitions among male *Drosophila melanogaster*. They took the "unpreferred" males that had lost three competitions in a row and mated them with females, then looked at the sex ratio of the offspring. They did this for three separate sets of flies.

	Males	Females		G-value	d.f.	P-value
Trial 1	296	366		7.42	1	0.006
Trial 2	78	72		0.24	1	0.624
Trial 3	417	467		2.83	1	0.093
			total G	10.49	3	0.015
pooled	791	905	pooled G	7.67	1	0.006
			heterogeneity G	2.82	2	0.24

The total G-value is significant, so we can reject the null hypotheses that all three trials have the same 1:1 sex ratio. The heterogeneity G-value is not significant; although the results of the second trial may look quite different from the results of the first and third trials, the three trials are not significantly different. We can therefore look at the pooled G-value. It is significant; the unpreferred males have significantly more daughters than sons.

Similar tests

For replicated goodness-of-fit tests, you must use the G-test, not the chi-squared test. Chi-squared values, although they are similar to G-values, do not add up the same way; the heterogeneity chi-square plus the pooled chi-square does not equal the total of the individual chi-squares. You could do a chi-squared test of independence among the replicates, then if that is not significant, pool the data and do a chi-squared goodness-of-fit test. However, you would not be able to detect the kind of overall deviation that the total G-value tests for.

Further reading

Sokal and Rohlf, pp. 715-724.

Reference

Connallon, T., and E. Jakubowski. 2009. Association between sex ratio distortion and sexually antagonistic fitness consequences of female choice. Evolution 63: 2179-2183.

Cochran–Mantel–Haenszel test for repeated tests of independence

When to use it

You use the Cochran–Mantel–Haenszel test (which is sometimes called the Mantel–Haenszel test) for repeated tests of independence. There are three nominal variables; you want to know whether two of the variables are independent of each other, and the third variable identifies the repeats. The most common situation is that you have multiple 2×2 tables of independence, so that's what I'll talk about here. There are versions of the Cochran–Mantel–Haenszel test for any number of rows and columns in the individual tests of independence, but I won't cover them.

For example, let's say you've found several hundred pink knit polyester legwarmers that have been hidden in a warehouse since they went out of style in 1984. You decide to see whether they reduce the pain of ankle osteoarthritis by keeping the ankles warm. In the winter, you recruit 36 volunteers with ankle arthritis, randomly assign 20 to wear the legwarmers under their clothes at all times while the other 16 don't wear the legwarmers, then after a month you ask them whether their ankles are pain-free or not. With just the one set of people, you'd have two nominal variables (legwarmers vs. control, pain-free vs. pain), each with two values, so you'd analyze the data with Fisher's exact test.

However, let's say you repeat the experiment in the spring, with 50 new volunteers. Then in the summer you repeat the experiment again, with 28 new volunteers. You could just add all the data together and do Fisher's exact test on the 114 total people, but it would be better to keep each of the three experiments separate. Maybe the first time you did the experiment there was an overall higher level of ankle pain than the second time, because of the different time of year or the different set of volunteers. You want to see whether there's an overall effect of legwarmers on ankle pain, but you want to control for possibility of different levels of ankle pain at the different times of year.

Null hypothesis

The null hypothesis is that the two nominal variables that are tested within each repetition are independent of each other; having one value of one variable does not mean that it's more likely that you'll have one value of the second variable. For your imaginary legwarmers experiment, the null hypothesis would be that the proportion of people feeling pain was the same for legwarmer-wearers and non-legwarmer wearers, after controlling for the time of year. The alternative hypothesis is that the proportion of people feeling pain was different for legwarmer and non-legwarmer wearers.

Technically, the null hypothesis of the Cochran–Mantel–Haenszel test is that the odds ratios within each repetition are equal to 1. The odds ratio is equal to 1 when the proportions are the same, and the odds ratio is different from 1 when the proportions are different from each other. I think proportions are easier to grasp than odds ratios, so I'll put everything in terms of proportions.

How it works

If the four numbers in a 2×2 test of independence are labelled like this:

a	b
c	d

and (a+b+c+d)=n, the equation for the Cochran–Mantel–Haenszel test statistic can be written like this:

$$\chi^2_{MH} = \frac{\{|\sum[a-(a+b)(a+c)/n]|-0.5\}^2}{\sum(a+b)(a+c)(b+d)(c+d)/(n^3-n^2)}$$

The numerator contains the absolute value of the difference between the observed value in one cell (*a*) and the expected value under the null hypothesis, (a+b)(a+c)/n, so the numerator is the squared sum of deviations between the observed and expected values. It doesn't matter how you arrange the 2×2 tables, any of the four values can be used as *a*. The 0.5 is subtracted as a continuity correction. The denominator contains an estimate of the variance of the squared differences.

The test statistic, χ^2_{MH}, gets bigger as the differences between the observed and expected values get larger, or as the variance gets smaller (primarily due to the sample size getting bigger). It is chi-square distributed with one degree of freedom.

Different sources present the formula for the Cochran–Mantel–Haenszel test in different forms, but they are all algebraically equivalent. The formula I've shown

here includes the continuity correction (subtracting 0.5 in the numerator); sometimes the Cochran–Mantel–Haenszel test is done without the continuity correction, so you should be sure to specify whether you used it when reporting your results.

Some statisticians recommend that you test the homogeneity of the odds ratios in the different repeats, and if different repeats show significantly different odds ratios, you shouldn't do the Cochran–Mantel–Haenszel test. In our arthritis-legwarmers example, they would say that if legwarmers have a significantly different effect on pain in the different seasons, you should analyze each experiment separately, rather than all together as the Cochran–Mantel–Haenszel test does. The most common way to test the homogeneity of odds ratios is with the Breslow–Day test, which I won't cover here.

Other statisticians will tell you that it's perfectly okay to use the Cochran–Mantel–Haenszel test when the odds ratios are significantly heterogeneous. The different recommendations depend on what your goal is. If your main goal is hypothesis testing—you want to know whether legwarmers reduce pain, in our example—then the Cochran–Mantel–Haenszel test is perfectly appropriate. A significant result will tell you that yes, the proportion of people feeling ankle pain does depend on whether or not they're wearing legwarmers. If your main goal is estimation—you want to estimate how well legwarmers work and come up with a number like "people with ankle arthritis are 50% less likely to feel pain if they wear fluorescent pink polyester knit legwarmers"—then it would be inappropriate to combine the data using the Cochran–Mantel–Haenszel test. If legwarmers reduce pain by 70% in the winter, 50% in the spring, and 30% in the summer, it would be misleading to say that they reduce pain by 50%; instead, it would be better to say that they reduce pain, but the amount of pain reduction depends on the time of year.

Examples

McDonald and Siebenaller (1989) surveyed allele frequencies at the *Lap* locus in the mussel *Mytilus trossulus* on the Oregon coast. At four estuaries, samples were taken from inside the estuary and from a marine habitat outside the estuary. There were three common alleles and a couple of rare alleles; based on previous results, the biologically interesting question was whether the Lap^{94} allele was less common inside estuaries, so all the other alleles were pooled into a "non-*94*" class.

There are three nominal variables: allele (94 or non-94), habitat (marine or estuarine), and area (Tillamook, Yaquina, Alsea, or Umpqua). The null hypothesis is that at each area, there is no difference in the proportion of Lap^{94} alleles between the marine and estuarine habitats, after controlling for area.

This table shows the number of *94* and non-*94* alleles at each location. There is a smaller proportion of *94* alleles in the estuarine location of each estuary when

compared with the marine location; we wanted to know whether this difference is significant.

Location	Allele	Marine	Estuarine
Tillamook	94	56	69
	non-94	40	77
Yaquina	94	61	257
	non-94	57	301
Alsea	94	73	65
	non-94	71	79
Umpqua	94	71	48
	non-94	55	48

Applying the formula given above, the numerator is 355.84, the denominator is 70.47, so the result is $\chi^2_{MH}=5.05$, 1 d.f., P=0.025. You can reject the null hypothesis that the proportion of *Lap*^{94} alleles is the same in the marine and estuarine locations.

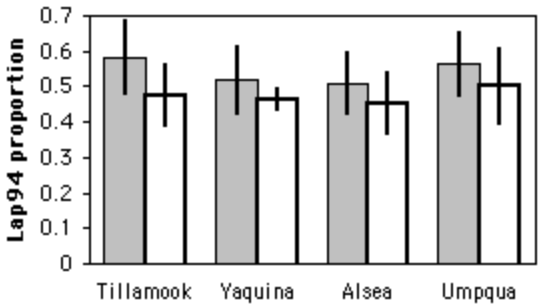
Gagnon et al. (2007) studied elk use of wildlife underpasses on a highway in Arizona. Using video surveillance cameras, they recorded each elk that started to cross under the highway. When a car or truck passed over while the elk was in the underpass, they recorded whether the elk continued through the underpass ("crossing") or turned around and left ("retreat"). The overall traffic volume was divided into low (fewer than 4 vehicles per minute) and high. There are three nominal variables: vehicle type (truck or car), traffic volume (low or high), and elk behavior (crossing or retreat). The question is whether trucks or cars are more likely to scare elk out of underpasses.

		Crossing	Retreat
Low traffic	Car	287	57
	Truck	40	42
High traffic	Car	237	52
	Truck	57	12

The result of the test is $\chi^2_{MH}=24.39$, 1 d.f., $P=7.9\times10^{-7}$. More elk are scared out of the underpasses by trucks than by cars.

Graphing the results

To graph the results of a Cochran–Mantel–Haenszel test, pick one of the two values of the nominal variable that you're observing and plot its proportions on a bar graph, using bars of two different patterns.



Lap^{94} allele proportions in the mussel *Mytilus trosulus* at four bays in Oregon. Gray bars are marine samples and empty bars are estuarine samples. Error bars are 95% confidence intervals.

Similar tests

Sometimes the Cochran–Mantel–Haenszel test is just called the Mantel–Haenszel test. This is confusing, as there is also a test for homogeneity of odds ratios called the Mantel–Haenszel test, and a Mantel–Haenszel test of independence for one 2×2 table. Mantel and Haenszel (1959) came up with a fairly minor modification of the basic idea of Cochran (1954), so it seems appropriate (and somewhat less confusing) to give Cochran credit in the name of this test.

If you have at least six 2×2 tables, and you're only interested in the *direction* of the differences in proportions, not the size of the differences, you could do a sign test. See the sign test web page for an example of an experiment with a very similar design to the *Lap* in *Mytilus trossulus* experiment described above, where because of the different biology of the organism, a sign test was more appropriate.

The Cochran–Mantel–Haenszel test for nominal variables is analogous to a two-way anova or paired t-test for a measurement variable, or a Wilcoxon signed-rank test for rank data. In the arthritis-legwarmers example, if you measured ankle pain on a 10-point scale (a measurement variable) instead of categorizing it as pain/no pain, you'd analyze the data with a two-way anova.

How to do the test

Spreadsheet

I've written a spreadsheet to perform the Cochran–Mantel–Haenszel test. It handles up to 50 2×2 tables (and you should be able to modify it to handle more, if necessary).

Web pages

I'm not aware of any web pages that will perform the Cochran–Mantel–Haenszel test.

SAS

Here is a SAS program that uses PROC FREQ for a Cochran–Mantel–Haenszel test. It uses the mussel data from above. In the TABLES statement, the variable that labels the repeats is listed first; in this case it is LOCATION.

```
data lap;
   input location $ habitat $ allele $ count;
   cards;
Tillamook marine         94     56
Tillamook estuarine      94     69
Tillamook marine     non-94     40
Tillamook estuarine  non-94     77
Yaquina   marine         94     61
Yaquina   estuarine      94    257
Yaquina   marine     non-94     57
Yaquina   estuarine  non-94    301
Alsea     marine         94     73
Alsea     estuarine      94     65
Alsea     marine     non-94     71
Alsea     estuarine  non-94     79
Umpqua    marine         94     71
Umpqua    estuarine      94     48
Umpqua    marine     non-94     55
Umpqua    estuarine  non-94     48
;
proc freq data=lap;
   weight count / zeros;
   tables location*habitat*allele / cmh;
   run;
```

There is a lot of output, but the important part looks like this:

```
  Cochran-Mantel-Haenszel Statistics (Based on Table Scores)

Statistic    Alternative Hypothesis    DF       Value      Prob
------------------------------------------------------------------
    1        Nonzero Correlation        1      5.3209    0.0211
    2        Row Mean Scores Differ     1      5.3209    0.0211
    3        General Association        1      5.3209    0.0211
```

For repeated 2x2 tables, the three statistics are identical; they are the Cochran–Mantel–Haenszel chi-square statistic, *without* the continuity correction. For repeated tables with more than two rows or columns, the "general association" statistic is used when the values of the different nominal variables do not have an order (you cannot arrange them from smallest to largest); you should use it unless you have a good reason to use one of the other statistics.

Further reading

Sokal and Rohlf, pp. 764-766.

References

Cochran, W.G. 1954. Some methods for strengthening the common χ^2 tests. Biometrics 10: 417-451.

Gagnon, J.W., T.C. Theimer, N.L. Dodd, A.L. Manzon, and R.E. Schweinsburg. 2007. Effects of traffic on elk use of wildlife underpasses in Arizona. J. Wildl. Manage. 71: 2324-2328.

Mantel, N., and W. Haenszel. 1959. Statistical aspects of the analysis of data from retrospective studies of disease. J. Natl. Cancer Inst. 22: 719-748.

McDonald, J.H. and J.F. Siebenaller. 1989. Similar geographic variation at the *Lap* locus in the mussels *Mytilus trossulus* and *M. edulis.* Evolution 43: 228-231.

Statistics of central tendency

All of the tests in the first part of this handbook have analyzed nominal variables. Data from a nominal variable is summarized as a percentage or a proportion. For example, 76.1 percent (or 0.761) of the peas in one of Mendel's genetic crosses were smooth, and 23.9 percent were wrinkled. If you have the percentage and the sample size (556, for Mendel's peas), you have all the information you need about the variable.

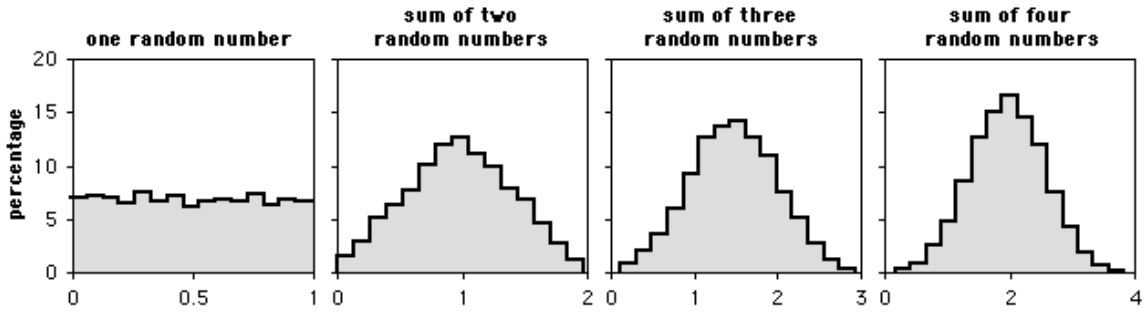
The rest of the tests in this handbook analyze measurement variables. Summarizing data from a measurement variable is more complicated, and requires a number that represents the "middle" of a set of numbers (known as a "statistic of central tendency" or "statistic of location"), along with a measure of the "spread" of the numbers (known as a "statistic of dispersion"). The arithmetic mean is the most common statistic of central tendency, while the variance or standard deviation are usually used to describe the dispersion.

The statistical tests for measurement variables assume that the probability distribution of the observations fits the normal (bell-shaped) curve. If this is true, the distribution can be accurately described by two parameters, the arithmetic mean and the variance. Because they assume that the distribution of the variables can be described by these two parameters, tests for measurement variables are called "parametric tests." If the distribution of a variable doesn't fit the normal curve, it can't be accurately desribed by just these two parameters, and the results of a parametric test may be inaccurate. In that case, the data are usually converted to ranks and analyzed using a non-parametric test, which is less sensitive to deviations from normality.

The normal distribution

Many measurement variables in biology fit the normal distribution fairly well. According to the central limit theorem, if you have several different variables that each have some distribution of values and add them together, the sum follows the normal distribution fairly well. It doesn't matter what the shape of the distribution of the individual variables is, the sum will still be normal. The distribution of the sum fits the normal distribution more closely as the number of variables increases. The graphs below are frequency histograms of 5,000 numbers. The first graph shows the distribution of a single number with a uniform distribution between 0

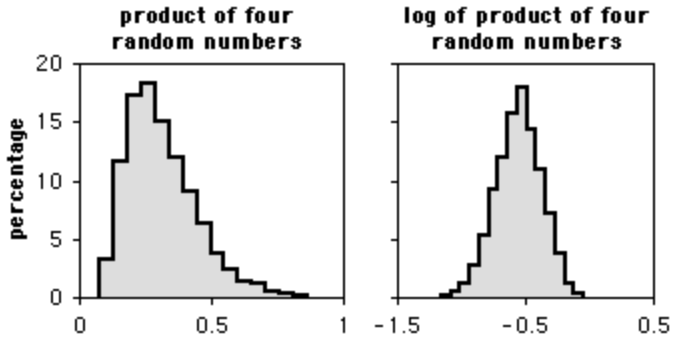
and 1. The other graphs show the distributions of the sums of two, three, or four random numbers.



Histograms of sums of random numbers.

As you can see, as more random numbers are added together, the frequency distribution of the sum quickly approaches a bell-shaped curve. This is analogous to a biological variable that is the result of several different factors. For example, let's say that you've captured 100 lizards and measured their maximum running speed. The running speed of an individual lizard would be a function of its genotype at many genes; its nutrition as it was growing up; the diseases it's had; how full its stomach is now; how much water it's drunk; and how motivated it is to run fast on a lizard racetrack. Each of these variables might not be normally distributed; the effect of disease might be to either subtract 10 cm/sec if it has had lizard-slowing disease, or add 20 cm/sec if it has not; the effect of gene A might be to add 25 cm/sec for genotype AA, 20 cm/sec for genotype Aa, or 15 cm/sec for genotype aa. Even though the individual variables might not have normally distributed effects, the running speed that is the sum of all the effects would be normally distributed.

If the different factors interact in a multiplicative, not additive, way, the distribution will be log-normal. An example would be if the effect of lizard-slowing disease is not to subtract 10 cm/sec from the total speed, but instead to reduce the speed by 10% (in other words, multiply the total speed by 0.9). The distribution of a log-normal variable will look like a bell curve that has been pushed to the left, with a long tail going to the right. Taking the log of such a variable will produce a normal distribution. This is why the log transformation is used so often.



Histograms of the product of four random numbers, without or with log transformation.

The figure above shows the frequency distribution for the product of four numbers, with each number having a uniform random distribution between 0.5 and 1. The graph on the left shows the untransformed product; the graph on the right is the distribution of the log-transformed products.

Different measures of central tendency

While the arithmetic mean is by far the most commonly used statistic of central tendency, you should be aware of a few others.

Arithmetic mean: The arithmetic mean is the sum of the observations divided by the number of observations. It is the most common statistic of central tendency, and when someone says simply "the mean" or "the average," this is what they mean. It is often symbolized by putting a bar over a letter; the mean of Y_1, Y_2, Y_3,… is $\overline{Y}$.

The arithmetic mean works well for values that fit the normal distribution. It is sensitive to extreme values, which makes it not work well for data that are highly skewed. For example, imagine that you are measuring the heights of fir trees in an area where 99 percent of trees are young trees, about 1 meter tall, that grew after a fire, and 1 percent of the trees are 50-meter-tall trees that survived the fire. If a sample of 20 trees happened to include one of the giants, the arithmetic mean height would be 3.45 meters; a sample that didn't include a big tree would have a mean height of about 1 meter. The mean of a sample would vary a lot, depending on whether or not it happened to include a big tree.

In a spreadsheet, the arithmetic mean is given by the function AVERAGE(Ys), where Ys represents a listing of cells (A2, B7, B9) or a range of cells (A2:A20) or both (A2, B7, B9:B21). Note that spreadsheets only count those cells that have numbers in them; you could enter AVERAGE(A1:A100), put numbers in cells A1 to A9, and the spreadsheet would correctly compute the arithmetic mean of those 9 numbers. This is true for other functions that operate on a range of cells.

Geometric mean: The geometric mean is the Nth root of the product of N values of Y; for example, the geometric mean of 5 values of Y would be the 5th root of $Y_1 \times Y_2 \times Y_3 \times Y_4 \times Y_5$. It is given by the spreadsheet function GEOMEAN(Ys). The

geometric mean is used for variables whose effect is multiplicative. For example, if a tree increases its height by 60 percent one year, 8 percent the next year, and 4 percent the third year, its final height would be the initial height multiplied by 1.60×1.08 ×1.04=1.80. Taking the geometric mean of these numbers (1.216) and multiplying that by itself three times also gives the correct final height (1.80), while taking the arithmetic mean (1.24) times itself three times does not give the correct final height. The geometric mean is slightly smaller than the arithmetic mean; unless the data are highly skewed, the difference between the arithmetic and geometric means is small. If any of your values are zero or negative, the geometric mean will be undefined.

The geometric mean has some useful applications in economics, but it is rarely used in biology. You should be aware that it exists, but I see no point in memorizing the definition.

Harmonic mean: The harmonic mean is the reciprocal of the arithmetic mean of reciprocals of the values; for example, the harmonic mean of 5 values of Y would be $5/(1/Y_1 + 1/Y_2 + 1/Y_3 + 1/Y_4 + 1/Y_5)$. It is given by the spreadsheet function HARMEAN(Ys). The harmonic mean is less sensitive to a few large values than are the arithmetic or geometric mean, so it is sometimes used for highly skewed variables such as dispersal distance. For example, if six birds set up their first nest 1.0, 1.4, 1.7, 2.1, 2.8, and 47 km from the nest they were born in, the arithmetic mean dispersal distance would be 9.33 km, the geometric mean would be 2.95 km, and the harmonic mean would be 1.90 km.

If any of your values are zero, the harmonic mean will be undefined.

I think the harmonic mean has some useful applications in engineering, but it is rarely used in biology. You should be aware that it exists, but I see no point in memorizing the definition.

Median: When the Ys are sorted from lowest to highest, this is the value of Y that is in the middle. For an odd number of Ys, the median is the single value of Y in the middle of the sorted list; for an even number, it is the arithmetic mean of the two values of Y in the middle. Thus for a sorted list of 5 Ys, the median would be Y_3; for a sorted list of 6 Ys, the median would be the arithmetic mean of Y_3 and Y_4. The median is given by the spreadsheet function MEDIAN(Ys).

The median is useful when dealing with highly skewed distributions. For example, if you were studying acorn dispersal, you might find that the vast majority of acorns fall within 5 meters of the tree, while a small number are carried 500 meters away by birds. The arithmetic mean of the dispersal distances would be greatly inflated by the small number of long-distance acorns. It would depend on the biological question you were interested in, but for some purposes a median dispersal distance of 3.5 meters might be a more useful statistic than a mean dispersal distance of 50 meters.

The second situation where the median is useful is when it is impractical to measure all of the values, such as when you are measuring the time until something happens. Survival time is a good example of this; in order to determine

the mean survival time, you have to wait until every individual is dead, while determining the median survival time only requires waiting until half the individuals are dead.

There are statistical tests for medians, such as Mood's median test, but they are rarely used due to their lack of power and won't be discussed in this handbook. If you are working with survival times of long-lived organisms (such as people), you'll need to learn about the specialized statistics for that; Bewick et al. (2004) is one place to start.

Mode: This is the most common value in a data set. It requires that a continuous variable be grouped into a relatively small number of classes, either by making imprecise measurments or by grouping the data into classes. For example, if the heights of 25 people were measured to the nearest millimeter, there would likely be 25 different values and thus no mode. If the heights were measured to the nearest 5 centimeters, or if the original precise measurements were grouped into 5-centimeter classes, there would probably be one height that several people shared, and that would be the mode.

It is rarely useful to determine the mode of a set of observations, but it is useful to distinguish between unimodal, bimodal, etc. distributions, where it appears that the parametric frequency distribution underlying a set of observations has one peak, two peaks, etc. The mode is given by the spreadsheet function MODE(Ys).

Example

The Maryland Biological Stream Survey used electrofishing to count the number of individuals of each fish species in randomly selected 75-m long segments of streams in Maryland. Here are the numbers of blacknose dace, *Rhinichthys atratulus,* in streams of the Rock Creek watershed:

Mill_Creek_1	76
Mill_Creek_2	102
North_Branch_Rock_Creek_1	12
North_Branch_Rock_Creek_2	39
Rock_Creek_1	55
Rock_Creek_2	93
Rock_Creek_3	98
Rock_Creek_4	53
Turkey_Branch	102

Here are the statistics of central tendency. In reality, you would rarely have any reason to report more than one of these:

Arithmetic mean	70.0
Geometric mean	59.8
Harmonic mean	45.1
Median	76
Mode	102

How to calculate the statistics

Spreadsheet

I have made a spreadsheet that calculates the arithmetic, geometric and harmonic means, the median, and the mode, for up to 1000 observations.

Web pages

This web page (http://home.ubalt.edu/ntsbarsh/Business-stat/otherapplets/Descriptive.htm) calculates arithmetic mean, median, and mode for up to 80 observations. It also includes range, variance, standard deviation, coefficient of variation, and standard error of the mean.

This web page (http://graphpad.com/quickcalcs/CImean1.cfm) calculates arithmetic mean and median for up to 10,000 observations. It also calculates standard deviation, standard error of the mean, and confidence intervals.

This web page (http://www.ruf.rice.edu/~lane/stat_analysis/descriptive.html) calculates arithmetic mean and median, along with range, variance, standard deviation, and standard error of the mean. I don't know the maximum number of observations it can handle.

SAS

There are three SAS procedures that do descriptive statistics, PROC MEANS, PROC SUMMARY, and PROC UNIVARIATE. I don't know why there are three. PROC UNIVARIATE will calculate a longer list of statistics, so you might as well use it. Here is an example, using the fish data from above.

```
data fish;
   input location $ dacenumber;
   cards;
Mill_Creek_1               76
Mill_Creek_2              102
North_Branch_Rock_Creek_1  12
North_Branch_Rock_Creek_2  39
Rock_Creek_1               55
Rock_Creek_2               93
Rock_Creek_3               98
Rock_Creek_4               53
Turkey_Branch             102
;
proc univariate data=fish;
run;
```

There's a lot of output from PROC UNIVARIATE, including the arithmetic mean, median, and mode:

```
          Basic Statistical Measures

      Location                    Variability
```

```
Mean       70.0000     Std Deviation           32.08582
Median     76.0000     Variance                    1030
Mode      102.0000     Range                   90.00000
                       Interquartile Range     45.00000
```

You can specify which variables you want the mean, median and mode of, using a VAR statement. You can also get the statistics for just those values of the measurement variable that have a particular value of a nominal variable, using a CLASS statement. This example calculates the statistics for the length of mussels, separately for each of two species, *Mytilus edulis* and *M. trossulus*.

```
data mussel;
   input species $ length width;
   cards;
edulis 49.0 11.0
tross  51.2  9.1
tross  45.9  9.4
edulis 56.2 13.2
edulis 52.7 10.7
edulis 48.4 10.4
tross  47.6  9.5
tross  46.2  8.9
tross  37.2  7.1
;
proc univariate;
   var length;
   class species;
run;
```

Surprisingly, none of the SAS procedures calculate harmonic or geometric mean. There are functions called HARMEAN and GEOMEAN, but they only calculate the means for a list of variables, not all the values of a single variable.

Further reading

Sokal and Rohlf, pp. 39-47.

Zar, pp. 20-28.

References

Blacknose dace data from Maryland Biological Stream Survey (http://www.dnr.state.md.us/streams/data/index.html) .

Bewick, V., L. Cheek, and J. Ball. 2004. Statistics review 12: Survival analysis. Crit. Care 8: 389-394.

Statistics of dispersion

Summarizing data from a measurement variable requires a number that represents the "middle" of a set of numbers (known as a "statistic of central tendency" or "statistic of location"), along with a measure of the "spread" of the numbers (known as a "statistic of dispersion"). Statistics of dispersion are used to give a single number that describes how compact or spread out a distribution of observations is. Although statistics of dispersion are usually not very interesting by themselves, they form the basis of most statistical tests used on measurement variables.

Range: This is simply the difference between the largest and smallest observations. This is the statistic of dispersion that people use in everyday conversation, but it is not very informative for statistical purposes. The range depends only on the largest and smallest values, so that two sets of data with very different distributions could have the same range. In addition, the range is expected to increase as the sample size increases; the more samples you take, the greater the chance that you'll sample a very large or very small value.

There is no range function in spreadsheets; the range can be found by using =MAX(Ys)−MIN(Ys), where Ys represents a set of cells. When you have a large data set, it's a good idea to look at the minimum and maximum values; if they're ridiculously small or large, they might be errors of some kind, such as a misplaced decimal point.

Sum of squares: This is not really a statistic of dispersion by itself, but it is mentioned here because it forms the basis of the variance and standard deviation. Subtract the sample mean from an observation and square this "deviate". Squaring the deviates makes all of the squared deviates positive and has other statistical advantages. Do this for each observation, then sum these squared deviates. This sum of the squared deviates from the mean is known as the sum of squares. It is given by the spreadsheet function DEVSQ(Ys) (*not* by the function SUMSQ).

Parametric variance: If you take the sum of squares and divide it by the number of observations (n), you are computing the average squared deviation from the mean. As observations get more and more spread out, they get farther from the mean, and the average squared deviate gets larger. This average squared deviate, or sum of squares divided by n, is the variance. You can only calculate the variance of a population this way if you have observations for every member of a population, which is almost never the case. I can't think of a good biological

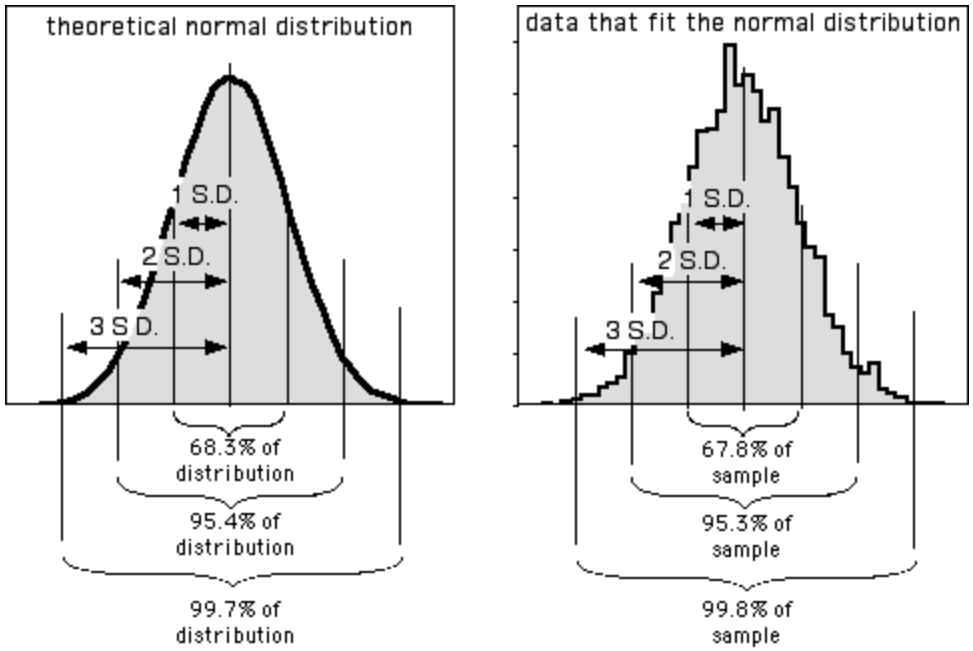
example where using the parametric variance would be appropriate. The parametric variance is given by the spreadsheet function VARP(Ys).

Sample variance: You almost always have a sample of observations that you are using to estimate a population parameter. To get an unbiased estimate of the population variance, divide the sum of squares by $n-1$, not by n. This sample variance, which is the one you will almost always use, is given by the spreadsheet function VAR(Ys). From here on, when you see "variance," it means the sample variance.

You might think that if you set up an experiment where you gave 10 guinea pigs little argyle sweaters, and you measured the body temperature of all 10 of them, that you should use the parametric variance and not the sample variance. You would, after all, have the body temperature of the entire population of guinea pigs wearing argyle sweaters in the world. However, for statistical purposes you should consider your sweater-wearing guinea pigs to be a sample of all the guinea pigs in the world who *could* have worn an argyle sweater, so it would be best to use the sample variance. Even if you go to Española Island and measure the length of every single tortoise (*Geochelone nigra hoodensis*) in the population of tortoises living there, it would be best to consider them a sample of all the tortoises that could have been living there.

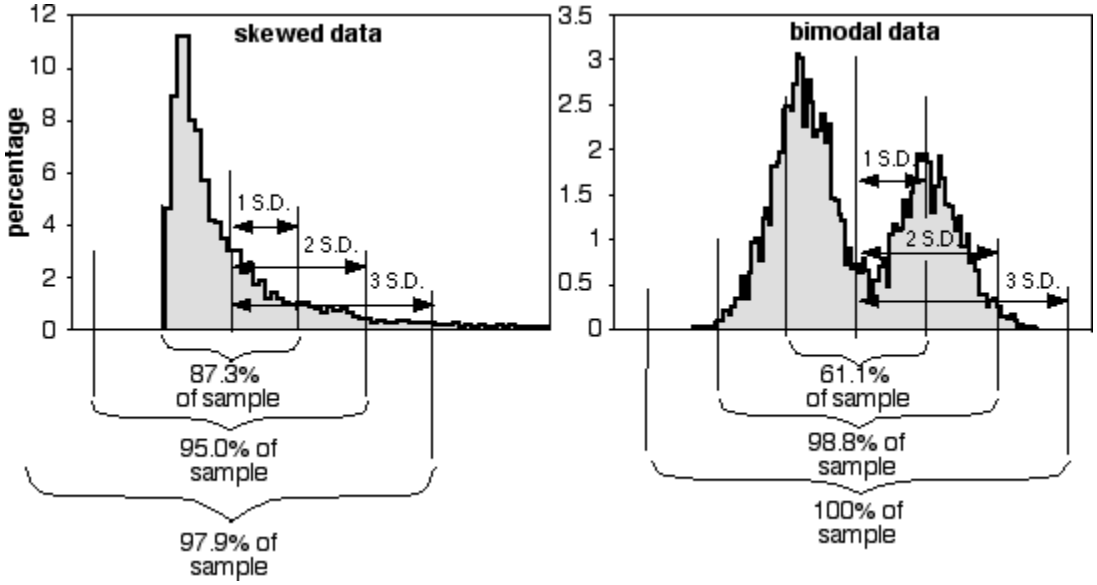
Standard deviation: Variance, while it has useful statistical properties that make it the basis of many statistical tests, is in squared units. A set of lengths measured in centimeters would have a variance expressed in square centimeters, which is just weird. Taking the square root of the variance gives a measure of dispersion that is in the original units. The square root of the parametric variance is the parametric standard deviation, which you will almost never use; is given by the spreadsheet function STDEVP(Ys). The sample standard deviation requires a rather complicated correction factor and is given by the spreadsheet function STDEV(Ys). You will almost always use the sample standard deviation; from here on, when you see "standard deviation," it means the sample standard deviation.

In addition to being more understandable than the variance as a measure of the amount of variation in the data, the standard deviation summarizes how close observations are to the mean in a very nice way. Many variables in biology fit the normal probability distribution fairly well. If a variable fits the normal distribution, 68.3 percent (or roughly two-thirds) of the values are within one standard deviation of the mean, 95.4 percent are within two standard deviations of the mean, and 99.7 (or almost all) are within 3 standard deviations of the mean. Here's a histogram that illustrates this:



Left: The theoretical normal distribution. Right: Frequencies of 5,000 numbers randomly generated to fit the normal distribution. The proportions of this data within 1, 2, or 3 standard deviations of the mean fit quite nicely to that expected from the theoretical normal distribution.

The proportions of the data that are within 1, 2, or 3 standard deviations of the mean are different if the data do not fit the normal distribution, as shown for these two very non-normal data sets:



Left: Frequencies of 5,000 numbers randomly generated to fit a distribution skewed to the right. Right: Frequencies of 5,000 numbers randomly generated to fit a bimodal distribution.

Coefficient of variation. Coefficient of variation is the standard deviation divided by the mean; it summarizes the amount of variation as a percentage or proportion of the total. It is useful when comparing the amount of variation among groups with different means. For example, let's say you wanted to know which had more variation, pinky finger length or little toe length; you want to know whether stabilizing selection is stronger on fingers than toes, since we use our fingers for more precise activities than our toes. Pinky fingers would almost certainly have a higher standard deviation than little toes, because fingers are several times longer than toes. However, the coefficient of variation might show that the standard deviation, as a percentage of the mean, was greater for toes.

Example

Here are the statistics of dispersion for the blacknose dace data from the central tendency web page. In reality, you would rarely have any reason to report all of these:

```
Range                      90
Variance                 1029.5
Standard deviation         32.09
Coefficient of variation   45.8%
```

How to calculate the statistics

Spreadsheet

I have made a spreadsheet that calculates the range, sample variance, sample standard deviation, and coefficient of variation, for up to 1000 observations.

Web pages

This web page (http://home.ubalt.edu/ntsbarsh/Business-stat/otherapplets/Descriptive.htm) calculates range, variance, standard deviation, and coefficient of variation for up to 80 observations.

This web page (http://www.ruf.rice.edu/~lane/stat_analysis/descriptive.html) calculates range, variance, and standard deviation. I don't know the maximum number of observations it can handle.

SAS

PROC UNIVARIATE will calculate the range, variance, standard deviation, and coefficient of variation. It calculates the sample variance and sample standard deviation. For examples, see the central tendency web page.

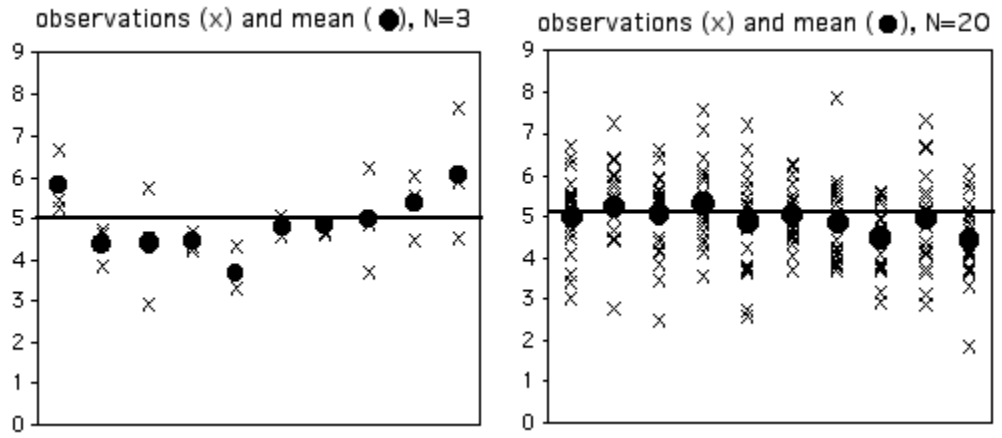
Further reading

Sokal and Rohlf, pp. 48-53, 57-59, 98-105.

Zar, pp. 32-40.

Standard error of the mean

When you take a sample of observations from a population, the mean of the sample is an estimate of the parametric mean, or mean of all of the observations in the population. If your sample size is small, your estimate of the mean won't be as good as an estimate based on a larger sample size. Here are 10 random samples from a simulated data set with a true (parametric) mean of 5. The X's represent the individual observations, the circles are the sample means, and the horizontal line is the parametric mean.

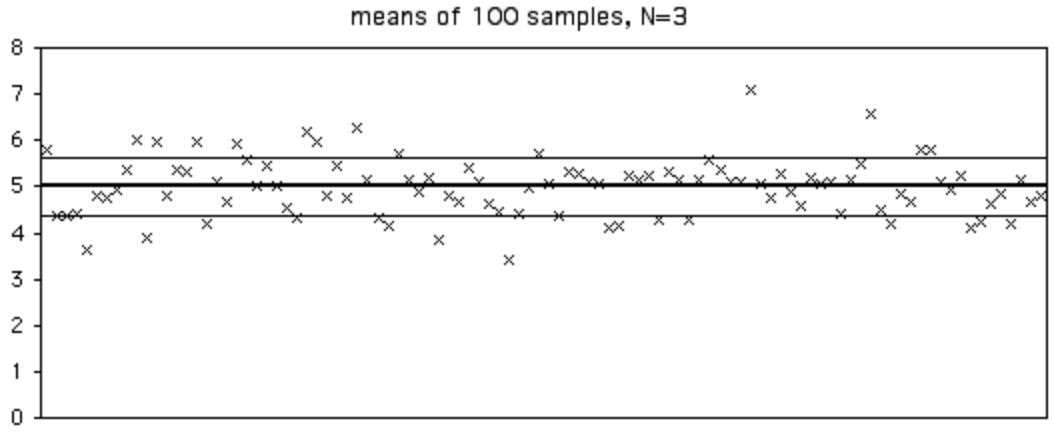


Individual observations (X's) and means (circles) for random samples from a population with a parametric mean of 5 (horizontal line).

As you can see, with a sample size of only 3, some of the sample means aren't very close to the parametric mean. The first sample happened to be three observations that were all greater than 5, so the sample mean is too high. The second sample has three observations that were less than 5, so the sample mean is too low. With 20 observations per sample, the sample means are generally closer to the parametric mean.

You'd often like to give some indication of how close your sample mean is likely to be to the parametric mean. One way to do this is with the standard error of the mean. If you take many random samples from a population, the standard error of the mean is the standard deviation of the different sample means. About

two-thirds (68.3%) of the sample means would be within one standard error of the parametric mean, 95.4% would be within two standard errors, and almost all (99.7%) would be within three standard errors.

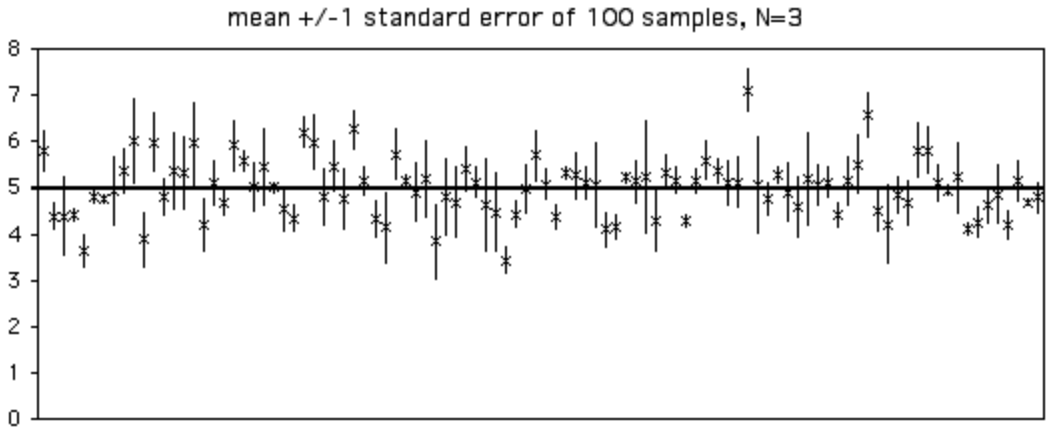


Means of 100 random samples (N=3) from a population with a parametric mean of 5 (horizontal line).

Here's a figure illustrating this. I took 100 samples of 3 from a population with a parametric mean of 5 (shown by the horizontal line). The standard deviation of the 100 means was 0.63. Of the 100 sample means, 70 are between 4.37 and 5.63 (the parametric mean ±one standard error).

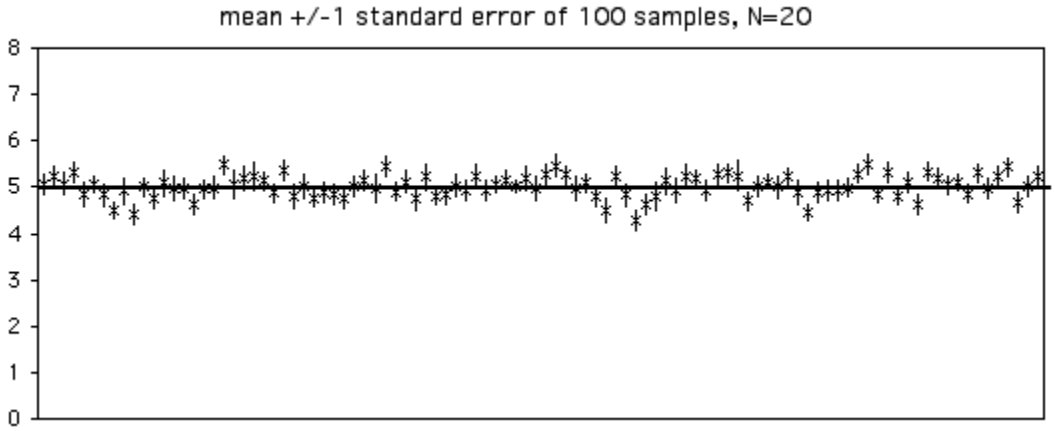
Usually you won't have multiple samples to use in making multiple estimates of the mean. Fortunately, it is possible to estimate the standard error of the mean using the sample size and standard deviation of a single sample of observations. The standard error of the mean is estimated by the standard deviation of the observations divided by the square root of the sample size. For some reason, there's no spreadsheet function for standard error, so you can use STDEV(Ys)/SQRT(COUNT(Ys)), where Ys is the range of cells containing your data.

This figure is the same as the one above, only this time I've added error bars indicating ±1 standard error. Because the estimate of the standard error is based on only three observations, it varies a lot from sample to sample.



Means ±1 standard error of 100 random samples (*n*=3) from a population with a parametric mean of 5 (horizontal line).

With a sample size of 20, each estimate of the standard error is more accurate. Of the 100 samples in the graph below, 68 include the parametric mean within ±1 standard error of the sample mean.



Means ±1 standard error of 100 random samples (N=20) from a population with a parametric mean of 5 (horizontal line).

As you increase your sample size, sample standard deviation will fluctuate, but it will not consistently increase or decrease. It will become a more accurate estimate of the parametric standard deviation of the population. In contrast, the standard error of the means will become smaller as the sample size increases. With bigger sample sizes, the sample mean becomes a more accurate estimate of the parametric mean, so the standard error of the mean becomes smaller.

"Standard error of the mean" and "standard deviation of the mean" are equivalent terms. "Standard error of the mean" is generally used to avoid confusion with the standard deviation of observations. Sometimes "standard error" is used by itself; this almost certainly indicates the standard error of the mean, but because there are also statistics for standard error of the variance, standard error of the median, etc., you should specify standard error of the mean.

Similar statistics

Confidence intervals and standard error of the mean serve the same purpose, to express the reliability of an estimate of the mean. In some publications, vertical error bars on data points represent the standard error of the mean, while in other publications they represent 95% confidence intervals. I prefer 95% confidence intervals. When I see a graph with a bunch of points and vertical bars representing means and confidence intervals, I know that most (95%) of the vertical bars include the parametric means. When the vertical bars are standard errors of the mean, only about two-thirds of the bars are expected to include the parametric means; I have to mentally double the bars to get the approximate size of the 95% confidence interval. In addition, for very small sample sizes, the 95% confidence interval is larger than twice the standard error, and the correction factor is even more difficult to do in your head. Whichever statistic you decide to use, be sure to make it clear what the error bars on your graphs represent. I have seen lots of graphs in scientific journals that gave no clue about what the error bars represent, which makes them pretty useless.

Standard deviation and coefficient of variation are used to show how much variation there is among individual observations, while standard error or confidence intervals are used to show how good your estimate of the mean is. The only time you would report standard deviation or coefficient of variation would be if you're actually interested in the amount of variation. For example, if you grew a bunch of soybean plants with two different kinds of fertilizer, your main interest would probably be whether the yield of soybeans was different, so you'd report the mean yield ± either standard error or confidence intervals. If you were going to do artificial selection on the soybeans to breed for better yield, you might be interested in which treatment had the greatest variation (making it easier to pick the fastest-growing soybeans), so then you'd report the standard deviation or coefficient of variation.

There's no point in reporting both standard error of the mean and standard deviation. As long as you report one of them, plus the sample size (N), anyone who needs to can calculate the other one.

Example

The standard error of the mean for the blacknose dace data from the central tendency web page is 10.70.

How to calculate the standard error

Spreadsheet

The descriptive statistics spreadsheet calculates the standard error of the mean for up to 1000 observations, using the function =STDEV(Ys)/SQRT(COUNT(Ys)).

Web pages

Web pages that will calculate standard error of the mean are here (http://home.ubalt.edu/ntsbarsh/Business-stat/otherapplets/Descriptive.htm) , here (http://graphpad.com/quickcalcs/CImean1.cfm) , and here (http://www.ruf.rice.edu/~lane/stat_analysis/descriptive.html) .

SAS

PROC UNIVARIATE will calculate the standard error of the mean. For examples, see the central tendency web page.

Further reading

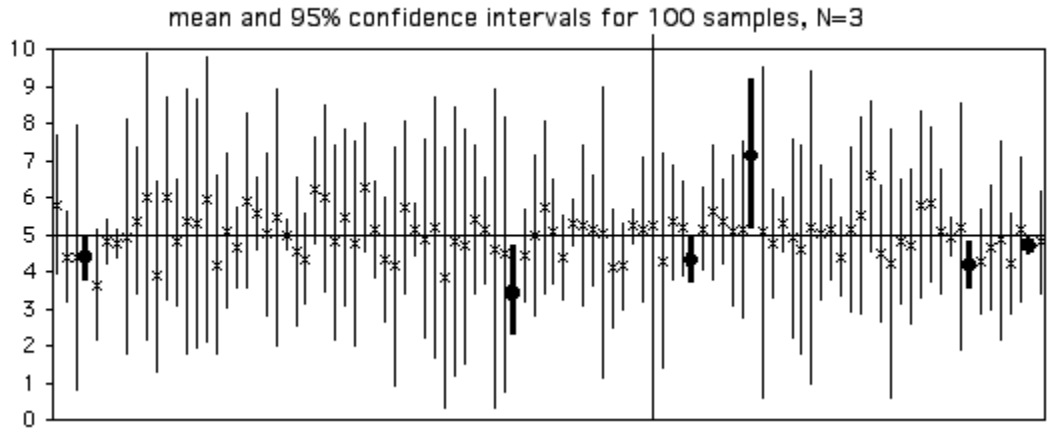
Sokal and Rohlf, pp. 127-136.

Zar, pp. 76-79.

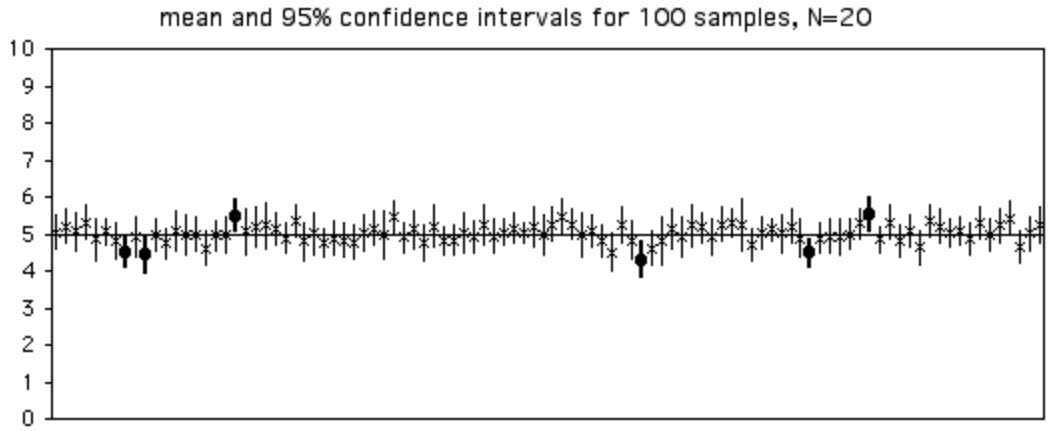
Confidence limits

After you've calculated the mean of a set of observations, you'd often like to give some indication of how close your estimate is likely to be to the parametric mean. One way to do this is with confidence limits, numbers at the upper and lower end of a confidence interval. Usually, 95% confidence limits are used, although you could use other values. Setting 95% confidence limits means that if you took repeated random samples from a population and calculated the mean and confidence limits for each sample, the confidence interval for 95% of your samples would include the parametric mean.

To illustrate this, here are the means and confidence intervals for 100 samples of 3 observations from a population with a parametric mean of 5. Of the 100 samples, 94 (shown with X for the mean and a thin line for the confidence interval) have the parametric mean within their 95% confidence interval, and 6 (shown with circles and thick lines) have the parametric mean outside the confidence interval.



With larger sample sizes, the 95% confidence intervals get smaller:



When you calculate the confidence limits for a single sample, it is tempting to say that "there is a 95% probability that the confidence interval includes the parametric mean." This is technically incorrect, because it implies that if you collected samples with the same confidence interval, sometimes they would include the parametric mean and sometimes they wouldn't. For example, the first sample in the figure above has confidence limits of 4.59 and 5.51. It would be incorrect to say that 95% of the time, the parametric mean for this population would lie between 4.59 and 5.51. If you took repeated samples from this same population and repeatedly got confidence limits of 4.59 and 5.51, the parametric mean (which is 5, remember) would be in this interval 100% of the time. Some statisticians don't care about the details of the definition, but others are very picky about this, so it's good to know.

Confidence limits for measurement variables

To calculate the confidence limits for a measurement variable, multiply the standard error of the mean times the appropriate t-value. The t-value is determined by the probability (0.05 for a 95% confidence interval) and the degrees of freedom (n−1). In a spreadsheet, you could use =(STDEV(Ys)/ SQRT(COUNT(Ys)))*TINV(0.05, COUNT(Ys)-1), where Ys is the range of cells containing your data. This value is added to and subtracted from the mean to get the confidence limits. Thus if the mean is 87 and the t-value times the standard error is 10.3, the confidence limits would be 76.7 to 97.3. You could also report this as "87 ±10.3 (95% confidence limits)." Both confidence limits and standard errors are reported as the "mean ± something," so always be sure to specify which you're talking about.

All of the above applies only to normally distributed measurement variables. For measurement data from a highly non-normal distribution, bootstrap techniques, which I won't talk about here, might yield better estimates of the confidence limits.

Confidence limits for nominal variables

There is a different, more complicated formula, based on the binomial distribution, for calculating confidence limits of proportions (nominal data). Importantly, it yields confidence limits that are not symmetrical around the proportion, especially for proportions near zero or one. John Pezzullo has an easy-to-use web page for confidence intervals of a proportion. To see how it works, let's say that you've taken a sample of 20 men and found 2 colorblind and 18 non-colorblind. Go to the web page and enter 2 in the "Numerator" box and 20 in the "Denominator" box," then hit "Compute." The results for this example would be a lower confidence limit of 0.0124 and an upper confidence limit of 0.3170. You can't report the proportion of colorblind men as "0.10 ± something," instead you'd have to say "0.10, 95% confidence limits of 0.0124, 0.3170," or maybe "0.10 +0.2170/ -0.0876 (95% confidence limits)."

An alternative technique for estimating the confidence limits of a proportion assumes that the sample proportions are normally distributed. This approximate technique yields symmetrical confidence limits, which for proportions near zero or one are obviously incorrect. For example, the confidence limits calculated with the normal approximation on 0.10 with a sample size of 20 are -0.03 to 0.23, which is ridiculous (you couldn't have less than zero percent of men being color-blind). It would also be incorrect to say that the confidence limits were 0 and 0.23, because you know the proportion of colorblind men in your population is greater than 0 (your sample had two colorblind men, so you know the population has at least two colorblind men). I consider confidence limits for proportions that are based on the normal approximation to be obsolete for most purposes; you should use the confidence interval based on the binomial distribution, unless the sample size is so large that it is computationally impractical. Unfortunately, you will see the confidence limits based on the normal approximation used more often than the correct, binomial confidence limits.

The formula for the 95% confidence interval using the normal approximation is $p \pm 1.96\sqrt{[p(1-p)/n]}$, where p is the proportion and n is the sample size. Thus, for p=0.20 and n=100, the confidence interval would be $\pm 1.96\sqrt{[0.20(1-0.20)/100]}$, or 0.20±0.078. A common rule of thumb says that it is okay to use this approximation as long as npq is greater than 5; my rule of thumb is to only use the normal approximation when the sample size is so large that calculating the exact binomial confidence interval makes smoke come out of your computer.

Similar statistics

Confidence limits and standard error of the mean serve the same purpose, to express the reliability of an estimate of the mean. In some publications, vertical error bars on data points represent the standard error of the mean, while in other publications they represent 95% confidence intervals. I prefer 95% confidence intervals. When I see a graph with a bunch of points and vertical bars representing

means and confidence intervals, I know that most (95%) of the vertical bars include the parametric means. When the vertical bars are standard errors of the mean, only about two-thirds of the bars are expected to include the parametric means; I have to mentally double the bars to get the approximate size of the 95% confidence interval (because t(0.05) is approximately 2 for all but very small values of n). Whichever statistic you decide to use, be sure to make it clear what the error bars on your graphs represent.

Examples

Measurement data: The blacknose dace data from the central tendency web page has an arithmetic mean of 70.0, with a 95% confidence interval of 24.7. The lower confidence limit is 70.0−24.7=45.3, and the upper confidence limit is 70+24.7=94.7.

Nominal data: If you work with a lot of proportions, it's good to have a rough idea of confidence limits for different sample sizes, so you have an idea of how much data you'll need for a particular comparison. For proportions near 50%, the confidence intervals are roughly ±30%, 10%, 3%, and 1% for n=10, 100, 1000, and 10,000, respectively. Of course, this rough idea is no substitute for an actual power analysis.

n	proportion=0.10	proportion=0.50
10	0.0025, 0.4450	0.1871, 0.8129
100	0.0490, 0.1762	0.3983, 0.6017
1000	0.0821, 0.1203	0.4685, 0.5315
10,000	0.0942, 0.1060	0.4902, 0.5098

How to calculate confidence limits

Spreadsheets

The descriptive statistics spreadsheet calculates 95% confidence limits of the mean for up to 1000 measurements. The confidence intervals for a binomial proportion spreadsheet calculates 95% confidence limits for nominal variables, using both the exact binomial and the normal approximation. (A corrected version of this spreadsheet was posted on Dec. 20, 2007; if you have the older version, discard it.)

Web pages

This web page (http://graphpad.com/quickcalcs/CImean1.cfm) calculates confidence intervals of the mean for up to 10,000 measurement observations. The web page for confidence intervals of a proportion (http://statpages.org/confint.html) handles nominal variables.

SAS

To get confidence limits for a measurement variable, add CIBASIC to the PROC UNIVARIATE statement, like this:

```
data fish;
   input location $ dacenumber;
   cards;
Mill_Creek_1               76
Mill_Creek_2              102
North_Branch_Rock_Creek_1  12
North_Branch_Rock_Creek_2  39
Rock_Creek_1               55
Rock_Creek_2               93
Rock_Creek_3               98
Rock_Creek_4               53
Turkey_Branch             102
;
proc univariate data=fish cibasic;
run;
```

The output will include the 95% confidence limits for the mean (and for the standard deviation and variance, which you would hardly ever need):

```
        Basic Confidence Limits Assuming Normality

   Parameter          Estimate     95% Confidence Limits

   Mean               70.00000      45.33665    94.66335
   Std Deviation      32.08582      21.67259    61.46908
   Variance               1030     469.70135        3778
```

This shows that the blacknose dace data have a mean of 70, with confidence limits of 45.3 to 94.7.

You can get the confidence limits for a binomial proportion using PROC FREQ. Here's the sample program from the exact binomial page:

```
data gus;
   input paw $;
   cards;
right
left
right
right
right
right
left
right
right
right
;
```

```
proc freq data=gus;
   tables paw / binomial(p=0.5);
exact binomial;
run;
```

And here is part of the output:

```
         Binomial Proportion
            for paw = left
   --------------------------------
   Proportion                0.2000
   ASE                       0.1265
   95% Lower Conf Limit      0.0000
   95% Upper Conf Limit      0.4479

   Exact Conf Limits
   95% Lower Conf Limit      0.0252
   95% Upper Conf Limit      0.5561
```

The first pair of confidence limits shown is based on the normal approximation; the second pair is the better one, based on the exact binomial calculation. Note that if you have more than two values of the nominal variable, the confidence limits will only be calculated for the value whose name is first alphabetically. For example, if the Gus data set included "left," "right," and "both" as values, SAS would only calculate the confidence limits on the proportion of "both." One clumsy way to solve this would be to run the program three times, changing the name of "left" to "aleft," then changing the name of "right" to "aright," to make each one first in one run.

Further reading

Sokal and Rohlf, pp. 139-151 (means).

Zar, pp. 98-100 (means), 527-530 (proportions).

Student's t-test

Any statistical test that uses the t-distribution can be called a t-test. One of the most common is Student's t-test, named after "Student," the pseudonym that William Gosset used to hide his employment by the Guinness brewery in the early 1900s (they didn't want their competitors to know that they were making better beer with statistics). Student's t-test is used to compare the means of two samples. Other t-tests include tests to compare a single observation to a sample, or to compare a sample mean to a theoretical mean (I won't cover either of these, as they are not used very often in biology), and the paired t-test.

When to use it

Use Student's t-test when you have one nominal variable and one measurement variable, and you want to compare the mean values of the measurement variable. The nominal variable must have only two values, such as "male" and "female" or "treated" and "untreated."

Null hypothesis

The statistical null hypothesis is that the means of the measurement variable are equal for the two categories.

How the test works

The test statistic, t_s, is calculated using a formula that has the difference between the means in the numerator; this makes t_s get larger as the means get further apart. The denominator is the standard error of the difference in the means, which gets smaller as the sample variances decrease or the sample sizes increase. Thus t_s gets larger as the means get farther apart, the variances get smaller, or the sample sizes increase.

The probability of getting the observed t_s value under the null hypothesis is calculated using the t-distribution. The shape of the t-distribution, and thus the probability of getting a particular t_s value, depends on the number of degrees of freedom. The degrees of freedom for a t-test is the total number of observations in the groups minus 2, or n_1+n_2-2.

Assumptions

The t-test assumes that the observations within each group are normally distributed and the variances are equal in the two groups. It is not particularly sensitive to deviations from these assumptions, but if the data are very non-normal, the Mann-Whitney U-test can be used. Welch's t-test can be used if the variances are unequal.

Example

In fall 2004, students in the 2 p.m. section of my Biological Data Analysis class had an average height of 66.6 inches, while the average height in the 5 p.m. section was 64.6 inches. Are the average heights of the two sections significantly different? Here are the data:

2 p.m.	5 p.m.
69	68
70	62
66	67
63	68
68	69
70	67
69	61
67	59
62	62
63	61
76	69
59	66
62	62
62	62
75	61
62	70
72	
63	

There is one measurement variable, height, and one nominal variable, class section. The null hypothesis is that the mean heights in the two sections are the same. The results of the t-test (t=1.29, 32 d.f., P=0.21) do not reject the null hypothesis.

Graphing the results

Because it's just comparing two numbers, you'd rarely put the results of a t-test in a graph for publication. For a presentation, you could draw a bar graph like the one for a one-way anova.

Similar tests

Student's t-test is mathematically identical to a one-way anova done on data with two categories. The t-test is easier to do and is familiar to more people, but it is limited to just two categories of data. The anova can be done on two or more categories. I recommend that if your research always involves comparing just two means, you should use the t-test, because it is more familiar to more people. If you write a paper that includes some comparisons of two means and some comparisons of more than two means, you may want to call all the tests one-way anovas, rather than switching back and forth between two different names (t-test and one-way anova) for what is essentially the same thing.

If the data are not normally distributed, and they can't be made normal using data transformations, it may be better to compare the ranks using a Mann-Whitney U-test. Student's t-test is not very sensitive to deviations from the normal distribution, so unless the non-normality is really dramatically obvious, you can use the t-test.

If the variances are far from equal, you can use Welch's t-test for unequal variances; you can do it in a spreadsheet using "=TTEST(array1, array2, tails, type)" by entering "3" for "type" instead of "2". You can also do Welch's t-test using this web page. The spreadsheet described on the homoscedasticity page can help you decide whether the difference in variances is be enough that you should use Welch's t-test.

The paired t-test is used when the measurement observations come in pairs, such as comparing the strengths of the right arm with the strength of the left arm on a set of people.

How to do the test

Spreadsheets

The easiest way to do the test is with the TTEST function. This takes the form "=TTEST(array1, array2, tails, type)". "Array1" is the set of cells with the measurement variables from your first class of observations, and "array2" is the set of cells with your second class of observations. "Tails" is either 1 (for a one-tailed test) or 2 (for a two-tailed test). You'll almost always want to do a two-tailed test. To do a regular t-test, enter "2" for the "type." The function returns the P-value of the test.

For the above height data, enter the first column of numbers in cells A2 through A23, and the second column of numbers in cells B2 through B18. In an empty cell, enter "=TTEST(A2:A23, B2:B18, 2, 2)". The result is P=0.207, so the difference in means is not significant.

Web pages

There are web pages to do the t-test here (http://graphpad.com/quickcalcs/ttest1.cfm) , here (http://www.physics.csbsju.edu/stats/t-test_NROW_form.html) , here (http://www.fon.hum.uva.nl/Service/Statistics/2Sample_Student_t_Test.html) , and here (http://faculty.vassar.edu/lowry/t_ind_stats.html) .

SAS

You can use PROC TTEST for Student's t-test; the CLASS parameter is the nominal variable, and the VAR parameter is the measurement variable. Here is an example program for the height data above.

```
data sectionheights;
 input section $ height;
 cards;
2pm 69
```

====See the web page for the full data set====

```
5pm 70
proc ttest;
 class section;
 var height;
 run;
```

The output includes a lot of information; the P-value for the Student's t-test is under "Pr > |t| on the line labelled "Pooled". For these data, the P-value is 0.2067.

Variable	Method	Variances	DF	t Value	Pr > |t|
height	Pooled	Equal	32	1.29	0.2067
height	Satterthwaite	Unequal	31.2	1.31	0.1995

Power analysis

To estimate the sample sizes needed to detect a significant difference between two means, you need the following:

- the effect size, or the difference in means you hope to detect;
- the standard deviation. Usually you'll use the same value for each group, but if you know ahead of time that one group will have a larger standard deviation than the other, you can use different numbers;
- alpha, or the significance level (usually 0.05);
- beta, the probability of accepting the null hypothesis when it is false (0.50, 0.80 and 0.90 are common values);
- the ratio of one sample size to the other. The most powerful design is to have equal numbers in each group (N_1/N_2=1.0), but sometimes it's easier to get large numbers of one of the groups. For example, if you're

comparing the bone strength in mice that have been reared in zero gravity aboard the International Space Station vs. control mice reared on earth, you might decide ahead of time to use three control mice for every one space mouse ($N_1/N_2=3.0$)

As an example, let's say you're planning a clinical trial of Niaspan in people with low levels of HDL (the "good cholesterol"). You're going to take a bunch of people with low HDL, give half of them Niaspan, and give the rest of them a placebo. The average HDL level before the trial is 32 mg/dl, and you decide you want to detect a difference of 10 percent (3.2 mg/dl), at the $P<0.05$ level, with a probability of detecting a difference this large, if it exists, of 80 percent ($1-\text{beta}=0.80$). Based on prior research, you estimate the standard deviation as 4.3 mg/dl in each group.

On the form on the web page, enter 3.2 for "Difference in means", 4.3 for both of the "Standard deviation" values, 0.05 for the alpha, 0.80 for the power, and 1.0 for N_1/N_2. The result is 29, so you'll need a minimum of 29 people in your placebo group and 29 in the Niaspan group.

Further reading

Sokal and Rohlf, pp. 223-227.

Zar, pp. 122-129.

One-way anova: Introduction

When to use it

Analysis of variance (anova) is the most commonly used technique for comparing the means of groups of measurement data. There are lots of different experimental designs that can be analyzed with different kinds of anova; in this handbook, I describe only one-way anova, nested anova and two-way anova.

In a one-way anova (also known as a single-classification anova), there is one measurement variable and one nominal variable. Multiple observations of the measurement variable are made for each value of the nominal variable. For example, you could measure the amount of transcript of a particular gene for multiple samples taken from arm muscle, heart muscle, brain, liver, and lung. The transcript amount would be the measurement variable, and the tissue type (arm muscle, brain, etc.) would be the nominal variable.

Null hypothesis

The statistical null hypothesis is that the means of the measurement variable are the same for the different categories of data; the alternative hypothesis is that they are not all the same.

How the test works

The basic idea is to calculate the mean of the observations within each group, then compare the variance among these means to the average variance within each group. Under the null hypothesis that the observations in the different groups all have the same mean, the weighted among-group variance will be the same as the within-group variance. As the means get further apart, the variance among the means increases. The test statistic is thus the ratio of the variance among means divided by the average variance within groups, or F_s. This statistic has a known distribution under the null hypothesis, so the probability of obtaining the observed F_s under the null hypothesis can be calculated.

The shape of the F-distribution depends on two degrees of freedom, the degrees of freedom of the numerator (among-group variance) and degrees of freedom of the denominator (within-group variance). The among-group degrees of

freedom is the number of groups minus one. The within-groups degrees of freedom is the total number of observations, minus the number of groups. Thus if there are n observations in a groups, numerator degrees of freedom is a-1 and denominator degrees of freedom is n-a.

Steps in performing a one-way anova

1. Decide whether you are going to do a Model I or Model II anova.
2. If you are going to do a Model I anova, decide whether you will do planned comparisons of means or unplanned comparisons of means. A planned comparison is where you compare the means of certain subsets of the groups that you have chosen in advance. In the arm muscle, heart muscle, brain, liver, lung example, an obvious planned comparison might be muscle (arm and heart) vs. non-muscle (brain, liver, lung) tissue. An unplanned comparison is done when you look at the data and then notice that something looks interesting and compare it. If you looked at the data and then noticed that the lung had the highest expression and the brain had the lowest expression, and you then compared just lung vs. brain, that would be an unplanned comparison. The important point is that planned comparisons must be planned before analyzing the data (or even collecting them, to be strict about it).
3. If you are going to do planned comparsions, decide which comparisons you will do. If you are going to do unplanned comparisons, decide which technique you will use.
4. Collect your data.
5. Make sure the data do not violate the assumptions of the anova (normality and homoscedasticity) too severely. If the data do not fit the assumptions well enough, try to find a data transformation that makes them fit. If this doesn't work, do a Welch's anova or a Kruskal–Wallis test instead of a one-way anova.
6. If the data do fit the assumptions of an anova, test the heterogeneity of the means.
7. If you are doing a Model I anova, do your planned or unplanned comparisons among means.
8. If the means are significantly heterogeneous, and you are doing a Model II anova, estimate the variance components (the proportion of variation that is among groups and the proportion that is within groups).

Similar tests

If you have only two groups, you can do a Student's t-test. This is mathematically equivalent to an anova, so if all you'll ever do is comparisons of two groups, you might as well use t-tests. If you're going to do some comparisons

of two groups, and some with more than two groups, it will probably be less confusing if you call all of your tests one-way anovas.

If there are two or more nominal variables, you should use a two-way anova, a nested anova, or something more complicated that I won't cover here. If you're tempted to do a very complicated anova, you may want to break your experiment down into a set of simpler experiments for the sake of comprehensibility.

If the data severely violate the assumptions of the anova, you can use Welch's anova if the variances are heterogeneous or use the Kruskal-Wallace test if the distributions are non-normal.

Power analysis

To do a power analysis for a one-way anova is kind of tricky, because you need to decide what kind of effect size you're looking for. If you're mainly interested in the overall significance test, the sample size needed is a function of the standard deviation of the group means. Your estimate of the standard deviation of means that you're looking for may be based on a pilot experiment or published literature on similar experiments.

If you're mainly interested in the planned or unplanned comparisions of means, there are other ways of expressing the effect size. Your effect could be a difference between the smallest and largest means, for example, that you would want to be significant by a Tukey-Kramer test. There are ways of doing a power analysis with this kind of effect size, but I don't know much about them and won't go over them here.

To do a power analysis for a one-way anova using the free program G*Power (http://www.psycho.uni-duesseldorf.de/abteilungen/aap/gpower3/) , choose "F tests" from the "Test family" menu and "ANOVA: Fixed effects, omnibus, one-way" from the "Statistical test" menu. To determine the effect size, click on the Determine button and enter the number of groups, the standard deviation within the groups (the program assumes they're all equal), and the mean you want to see in each group. Usually you'll leave the sample sizes the same for all groups (a balanced design), but if you're planning an unbalanced anova with bigger samples in some groups than in others, you can enter different relative sample sizes. Then click on the "Calculate and transfer to main window" button; it calculates the effect size and enters it into the main window. Enter your alpha (usually 0.05) and power (typically 0.80 or 0.90) and hit the Calculate button. The result is the total sample size in the whole experiment; you'll have to do a little math to figure out the sample size for each group.

As an example, let's say you're studying transcript amount of some gene in arm muscle, heart muscle, brain, liver, and lung. Based on previous research, you decide that you'd like the anova to be significant if the means were 10 units in arm muscle, 10 units in heart muscle, 15 units in brain, 15 units in liver, and 15 units in lung. The standard deviation of transcript amount within a tissue type that you've seen in previous research is 12 units. Entering these numbers in G*Power, along

with an alpha of 0.05 and a power of 0.80, the result is a total sample size of 295. Since there are five groups, you'd need 59 observations per group to have an 80 percent chance of having a significant ($P<0.05$) one-way anova.

Further reading

Sokal and Rohlf, pp. 206-217.

Zar, pp. 177-195.

Model I vs. Model II anova

One of the first steps in performing a one-way anova is deciding whether to do a Model I or Model II anova. The test of homogeneity of means is the same for both models, but the choice of models determines what you do if the means are significantly heterogeneous.

Model I anova

In a model I anova (also known as a fixed-effects model anova), the groups are identified by some characteristic that is repeatable and interesting. If there is a difference among the group means and you repeat the experiment, you would expect to see the same pattern of differences among the means, because you could classify the observations into the same groups. The group labels are meaningful (such as "seawater, glucose solution, mannose solution"). You are interested in the relationship between the way the data are grouped (the "treatments") and the group means. Examples of data for which a model I anova would be appropriate are:

- Time of death of amphipod crustaceans being suffocated in plain seawater, a glucose solution, or a mannose solution. The three different solutions are the treatments, and the question is whether amphipods die more quickly in one solution than another. If you find that they die the fastest in the mannose solution, you would expect them to die the fastest in mannose if you repeated the experiment.
- Amounts of a particular transcript in tissue samples from arm muscle, heart muscle, brain, liver and lung, with multiple samples from each tissue. The tissue type is the treatment, and the question you are interested in is which tissue has the highest amount of transcript. Note that "treatment" is used in a rather broad sense. You didn't "treat" a bunch of cells and turn them into brain cells; you just sampled some brain cells.
- The tastiness of peaches from 10 different peach trees, if you want to use cuttings from the tree or trees with the tastiest peaches to start an orchard.

If you have significant heterogeneity among the means in a model I anova, the next step (if there are more than two groups) is usually to try to determine which means are significantly different from other means. In the amphipod example, if

there were significant heterogeneity in time of death among the treatments, the next question would be "Is that because mannose kills amphipods, while glucose has similar effects to plain seawater? Or does either sugar kill amphipods, compared with plain seawater? Or is it glucose that is deadly?" To answer questions like these, you will do either planned comparisons of means (if you decided, *before looking at the data,* on a limited number of comparisons) or unplanned comparisons of means (if you just looked at the data and picked out interesting comparisons to do).

Model II anova

In a model II anova (also known as a random-effects model anova), the groups are identified by some characteristic that is not interesting; they are just groups chosen from a larger number of possible groups. If there is heterogeneity among the group means and you repeat the experiment, you would expect to see heterogeneity again, but you would not expect to see the same pattern of differences. The group labels are generally arbitrary (such as "family A, family B, family C"). You are interested in the amount of variation among the means, compared with the amount of variation within groups. Examples of data for which a model II anova would be appropriate are:

- Repeated measurements of glycogen levels in each of several pieces of a rat gastrocnemius muscle. If variance among pieces is a relatively small proportion of the total variance, it would suggest that a single piece of muscle would give an adequate measure of glycogen level. If variance among pieces is relatively high, it would suggest that either the sample preparation method needs to be better standardized, or there is heterogeneity in glycogen level among different parts of the muscle.
- Sizes of treehoppers from different sibships, all raised on a single host plant. If the variance among sibships is high relative to the variance within sibships (some families have large treehopppers, some families have small treehoppers), it would indicate that heredity (or maternal effects) play a large role in determining size.
- The tastiness of peaches from 10 different peach trees, if you want to estimate how much of the variation in peach tastiness is due to the tree, and how much is due to variation among peaches within each tree. If most of the variation in tastiness is among peaches within each tree, then using cuttings from the best tree won't do much to improve the tastiness of the peaches.

If you have significant heterogeneity among the means in a model II anova, the next step is to partition the variance into the proportion due to the treatment effects and the proportion within treatments.

How to tell the difference

If you are going to follow up a significant result with planned or unplanned comparisons of means, it's model I; if you are going to follow up by partitioning the variance, it's model II. Sometimes it's not obvious which model to use; I've seen many examples of researchers partitioning the variance after a model I anova, or doing comparisons of means after a model II anova, just because their software outputs both sets of numbers. I find it helpful to imagine that you've written all the observations for the measurement variable on cards, with one card for each group. At the top of each card you've written the name of the group. For example, imagine you've written the tastiness measurements for 10 peaches from tree A on one card, 10 peaches from tree B on a second card, etc. Then imagine that your scientific arch-enemy sneaks into your lab and erases the tree identification letter (A, B, C, ...) from each card. Now if one of the trees has significantly better tastiness measures than the other trees, you don't know which tree it was. If your experiment is completely ruined, and you have to wait a year until you can go back to the same peach trees and get new tastiness measures, and therefore your scientific arch-enemy is cackling with glee, that's a model I anova. But if your experiment isn't ruined—if your goal was to see how much tree-to-tree variation there was, and you can just write new arbitrary tree names on each card and still answer the question, and your arch enemy is going "Curses! Foiled again!"—that's a model II anova.

Further reading

Sokal and Rohlf, pp. 201-205.

Zar, pp. 184-185.

One-way anova: Testing the homogeneity of means

Once you have chosen between a model I and model II anova, the next step is to test the homogeneity of means. The null hypothesis is that the all the groups have the same mean, and the alternate hypothesis is that at least one of the means is different from the others.

To test the null hypothesis, the variance of the population is estimated in two different ways. I'll explain this in a way that is strictly correct only for a "balanced" one-way anova, one in which the sample size for each group is the same, but the basic concept is the same for unbalanced anovas.

If the null hypothesis is true, all the groups are samples from populations with the same mean. One of the assumptions of the anova is that the populations have the same variance, too. One way to estimate this variance starts by calculating the variance within each sample—take the difference between each observation and its group's mean, square it, then sum these squared deviates and divide by the number of observations in the group minus one. Once you've estimated the variance within each group, you can take the average of these variances. This is called the "within-group mean square," or MS_{within}.

For another way to estimate the variance within groups, remember that if you take repeated samples of a population, you expect the means you get from the multiple samples to have a standard deviation that equals the standard deviation within groups divided by the square root of *n*; this is the definition of standard error of the mean, or

$$E(SD_{means}) = SD_{within}/\sqrt{n}$$

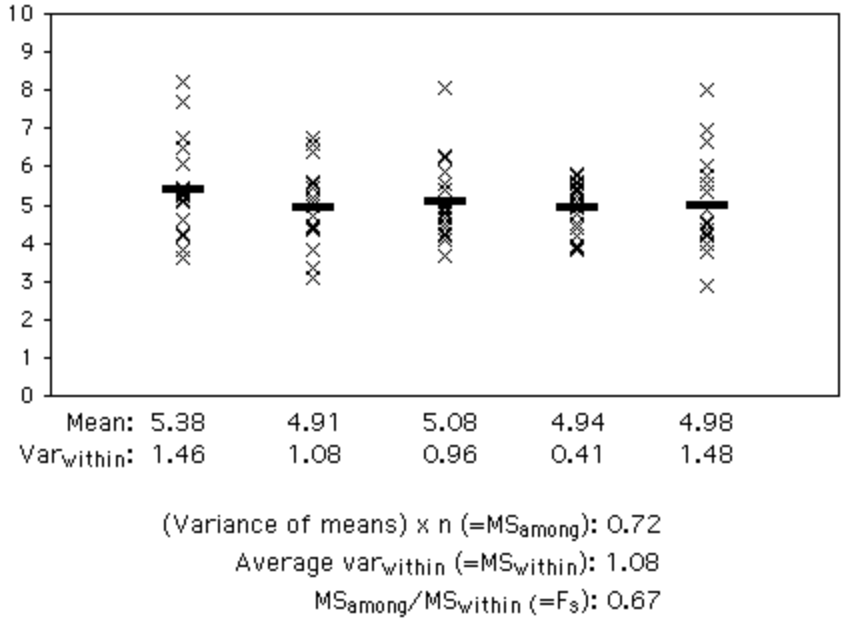
Remember that the standard deviation is just the square root of the variance, so squaring both sides of this gives:

$$E(Var_{means}) = Var_{within}/n$$

so the second way of estimating the variance within groups is $n \times Var_{means}$, the sample size within a group times the variance of the group means. This quantity is known as the among-group mean square, abbreviated MS_{among} or MS_{group}.

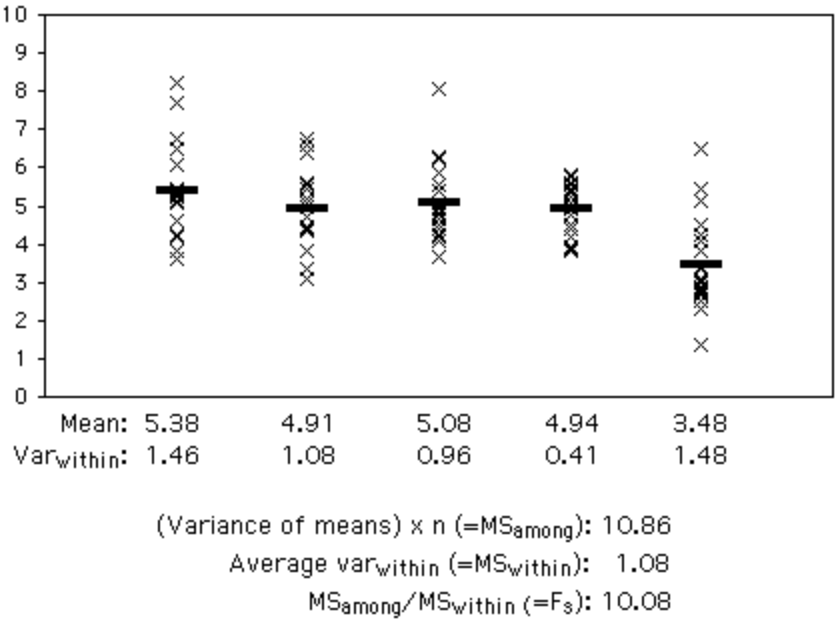
If the null hypothesis is true and the groups are all samples from populations with the same mean, the two estimates of within-group variance, MS_{within} and MS_{among}, should be about the same; they're just different ways of estimating the same quantity. Dividing MS_{among} by MS_{within} should therefore be around 1. This quantity, MS_{among}/MS_{within}, is known as F_s, and it is the test statistic for the anova.

If the null hypothesis is *not* true, and the groups are samples of populations with different means, then MS_{among} will be bigger than MS_{within}, and F_s will be greater than 1. To illustrate this, here are two sets of five samples (n=20) taken from normally distributed populations. The first set of five samples are from populations with a mean of 5; the null hypothesis, that the populations all have the same mean, is true.



Five samples (n=20) from populations with parametric means of 5. Thick horizontal lines indicate sample means.

The variance among the five group means is quite small; multiplying it by the sample size (20) yields 0.72, about the same as the average variance within groups (1.08). These are both about the same as the parametric variance for these populations, which I set to 1.0.



Four samples (n=20) from populations with parametric means of 5; the last sample is from a population with a parametric mean of 3.5. Thick horizontal lines indicate sample means.

The second graph is the same as the first, except that I have subtracted 1.5 from each value in the last sample. The average variance within groups (MS_{within}) is exactly the same, because each value was reduced by the same amount; the size of the variation among values within a group doesn't change. The variance among groups does get bigger, because the mean for the last group is now quite a bit different from the other means. MS_{among} is therefore quite a bit bigger than MS_{within}, so the ratio of the two (F_s) is much larger than 1.

The theoretical distribution of F_s under the null hypothesis is given by the F-distribution. It depends on the degrees of freedom for both the numerator (among-groups) and denominator (within-groups). The probability associated with an F-statistic is given by the spreadsheet function FDIST(x, df1, df2), where x is the observed value of the F-statistic, df1 is the degrees of freedom in the numerator (the number of groups minus one, for a one-way anova) and df2 is the degrees of freedom in the denominator (total n minus the number of groups, for a one-way anova).

Example

Here are some data on a shell measurement (the length of the anterior adductor muscle scar, standardized by dividing by length) in the mussel *Mytilus trossulus* from five locations: Tillamook, Oregon; Newport, Oregon; Petersburg, Alaska;

Magadan, Russia; and Tvarminne, Finland, taken from a much larger data set used in McDonald et al. (1991).

Tillamook	Newport	Petersburg	Magadan	Tvarminne
0.0571	0.0873	0.0974	0.1033	0.0703
0.0813	0.0662	0.1352	0.0915	0.1026
0.0831	0.0672	0.0817	0.0781	0.0956
0.0976	0.0819	0.1016	0.0685	0.0973
0.0817	0.0749	0.0968	0.0677	0.1039
0.0859	0.0649	0.1064	0.0697	0.1045
0.0735	0.0835	0.1050	0.0764	
0.0659	0.0725		0.0689	
0.0923				
0.0836				

The conventional way of reporting the complete results of an anova is with a table (the "sum of squares" column is often omitted). Here are the results of a one-way anova on the mussel data:

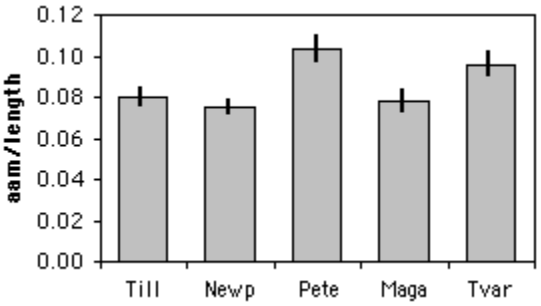
	sum of squares	**d.f.**	**mean square**	$\mathbf{F_s}$	**P**
among groups	0.00452	4	0.001113	7.12	2.8×10^{-4}
within groups	0.00539	34	0.000159		
total	0.00991	38			

If you're not going to use the mean squares for anything, you could just report this as "The means were significantly heterogeneous (one-way anova, $F_{4, 34}=7.12$, $P=2.8\times10^{-4}$)." The degrees of freedom are given as a subscript to F.

Note that statisticians often call the within-group mean square the "error" mean square. I think this can be confusing to non-statisticians, as it implies that the variation is due to experimental error or measurement error. In biology, the within-group variation is often largely the result of real, biological variation among individuals, not the kind of mistakes implied by the word "error."

Graphing the results

The usual way to graph the results of a one-way anova is with a bar graph. The heights of the bars indicate the means, and there's usually some kind of error bar: 95% confidence intervals, standard errors, or comparison intervals. Be sure to say in the figure caption what the error bars represent.



Length of the anterior adductor muscle scar divided by total length in *Mytilus trossulus*. Means ±one standard error are shown for five locations.

How to do the test

Spreadsheet

I have put together a spreadsheet to do one-way anova on up to 50 groups and 1000 observations per group. It calculates the P-value, does unplanned comparisons of means (appropriate for a model I anova) using Gabriel comparison intervals and the Tukey–Kramer test, and partitions the variance (appropriate for a model II anova) into among- and within-groups components.

Some versions of Excel include an "Analysis Toolpak," which includes an "Anova: Single Factor" function that will do a one-way anova. You can use it if you want, but I can't help you with it. It does not include any techniques for unplanned comparisons of means, and it does not partition the variance.

Web pages

Several people have put together web pages that will perform a one-way anova; one good one is here. (http://www.physics.csbsju.edu/stats/anova.html) It is easy to use, and will handle three to 26 groups and 3 to 1024 observations per group. It does not calculate statistics used for unplanned comparisons, and it does not partition the variance. Another good web page for anova is Rweb (http://rweb.stat.umn.edu/cgi-bin/Rweb/buildModules.cgi) .

SAS

There are several SAS procedures that will perform a one-way anova. The two most commonly used are PROC ANOVA and PROC GLM. Either would be fine for a one-way anova, but PROC GLM (which stands for "General Linear Models")

can be used for a much greater variety of more complicated analyses, so you might as well use it for everything.

Here is a SAS program to do a one-way anova on the mussel data from above.

```
data musselshells;
   input location $ aam;
   cards;
Tillamook 0.0571
```
====See the web page for the full data set====
```
Tvarminne 0.1045
proc glm data=musselshells;
  class location;
  model aam = location;
run;
```

The output includes the traditional anova table; the P-value is given under "Pr > F".

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	0.00451967	0.00112992	7.12	0.0003
Error	34	0.00539491	0.00015867		
Corrected Total	38	0.00991458			

Welch's anova

If the data show a lot of heteroscedasticity (different groups have different variances), the one-way anova can yield an inaccurate P-value; the probability of a false positive may be much higher than 5 percent. In that case, the most common alternative is Welch's anova. This can be done in SAS by adding a MEANS statement, the name of the nominal variable, and the word WELCH following a slash. Here is the example SAS program from above, modified to do Welch's anova:

```
proc glm data=musselshells;
  class location;
  model aam = location;
  means location / welch;
run;
```

Here is the output:

```
            Welch's ANOVA for aam

Source             DF    F Value    Pr > F

location       4.0000       5.66    0.0051
Error         15.6955
```

Further reading

Sokal and Rohlf, pp. 207-217.

Zar, pp. 183.

Reference

McDonald, J.H., R. Seed and R.K. Koehn. 1991. Allozymes and morphometric characters of three species of *Mytilus* in the Northern and Southern Hemispheres. Mar. Biol. 111:323-333.

One-way anova: Planned comparisons of means

With a Model I anova, in addition to testing the overall heterogeneity among the means of more than two groups, it is often desirable to perform additional comparisons of subsets of the means. For example, let's say you have measured the height of the arch of the foot in athletes from nine women's teams: soccer, basketball, rugby, swimming, softball, volleyball, lacrosse, crew and cross-country. You might want to compare the mean of sports that involve a lot of jumping (basketball and volleyball) vs. all other sports. Or you might want to compare swimming vs. all other sports. Or you might want to compare soccer vs. basketball, since they involve similar amounts of running but different amounts of kicking. There are thousands of ways of dividing up nine groups into subsets, and if you do unplanned comparisons, you have to adjust your P-value to a much smaller number to take all the possible tests into account. It is better to plan a small number of interesting comparisons before you collect the data, because then it is not necessary to adjust the P-value for all the tests you didn't plan to do.

Orthogonal comparisons

It is best if your planned comparisons are orthogonal, because then you do not need to adjust the P-value at all. Orthogonal comparisons are those in which all of the comparisons are independent; you do not compare the same means to each other twice. Doing one comparison of soccer vs. basketball and one of swimming vs. cross-country would be orthogonal, as would soccer vs. basketball and soccer vs. swimming. Jumping sports (basketball and volleyball) vs. non-jumping sports (all others) and rugby vs. lacrosse and softball vs. crew would be three orthogonal comparisons. Jumping sports vs. non-jumping sports and volleyball vs. swimming would not be orthogonal, because the volleyball vs. swimming comparison is included in the jumping vs. non-jumping comparison. Non-ball sports (swimming, crew, cross-country) vs. ball sports and jumping vs. non-jumping would not be orthogonal, because swimming vs. volleyball, among several other pairs, would be included in both comparisons.

The degrees of freedom for each planned comparison is equal to the number of groups, after pooling, minus one. Thus the jumping vs. non-jumping comparison

would have one degree of freedom, and non-jumping vs. basketball vs. volleyball would have two degrees of freedom. The maximum total number of degrees of freedom for a set of orthogonal comparisons is the numerator degrees of freedom for the original anova (the original number of groups minus one). You do not need to do a full set of orthogonal comparisons; in this example, you might want to do jumping vs. non-jumping, then stop. Here is an example of a full set of orthogonal comparisons for the sport example; note that the degrees of freedom add up to eight, the number of original groups minus one.

- Jumping (basketball and volleyball) vs. non-jumping sports; 1 d.f.
- Basketball vs. volleyball; 1 d.f.
- Soccer+rugby+lacrosse+softball (ball sports) vs. swimming vs. crew vs. cross-country; 3 d.f.
- Rugby+lacrosse+softball (non-kicking sports) vs. soccer; 1 d.f.
- Rugby vs. lacrosse vs. softball; 2 d.f.

To perform a planned comparison, you simply perform an anova on the pooled data. If you have a spreadsheet with the foot arch data from nine sports teams in nine columns, and you want to do jumping sports vs. non-jumping sports as a planned comparsion, simply copy the volleyball data from one column and paste it at the bottom of the basketball column. Then combine all of the data from the other sports into a single column. The resulting P-value is the correct value to use for the planned comparison.

Non-orthogonal comparisons

Sometimes the hypotheses you are interested in make it necessary to do non-orthogonal planned comparisons. For example, you might want to do jumping vs. non-jumping sports, ball sports vs. non-ball sports, swimming vs. crew, and soccer vs. all other sports. In this case, it is necessary to adjust the P-values downward to account for the multiple tests you are doing.

To understand why this is necessary, imagine that you did 100 planned comparisons on the sports data set. Under the null hypothesis that the means were homogeneous, you would expect about 5 of the comparisons to be "significant" at the $p<0.05$ level. This is what $p<0.05$ means, after all: 5% of the time you get a "significant" result even though the null hypothesis is true. Clearly it would be a mistake to consider those 5 comparisons that happen to have $P<0.05$ to be significant rejections of the particular null hypotheses that each comparison tests. Instead you want to use a lower alpha, so the overall probability is less than 0.05 that the set of planned comparisons includes one with a P-value less than the adjusted alpha.

The sequential Dunn–Sidák method is one good way to adjust alpha levels for planned, non-orthogonal comparisons. First, the P-values from the different comparisons are put in order from smallest to largest. If there are k comparisons,

the smallest P-value must be less than $1-(1-\text{alpha})^{1/k}$ to be significant at the alpha level. Thus if there are four comparisons, the smallest P-value must be less than $1-(1-0.05)^{1/4}=0.0127$ to be significant at the 0.05 level. If it is not significant, the analysis stops. If the smallest P-value is significant, the next smallest P-value must be less than $1-(1-\text{alpha})^{1/(k-1)}$, which in this case would be 0.0170. If it is significant, the next P-value must be less than $1-(1-\text{alpha})^{1/(k-2)}$, and so on until one of the P-values is not significant.

Other techniques for adjusting the alpha are less powerful than the sequential method, but you will often see them in the literature and should therefore be aware of them. The Bonferroni method uses alpha/k as the adjusted alpha level, while the Dunn–Sidák method uses $1-(1-\text{alpha})^{1/k}$. The difference between the Bonferroni and Dunn–Sidák adjusted alphas is quite small, so it usually wouldn't matter which you used. For example, the Bonferroni alpha for four comparisons is 0.0125, while the Dunn–Sidák is 0.0127. These are not sequential methods; the same adjusted alpha is used, no matter how many of the comparisons are significant.

Really important note about planned comparisons

Planned comparisons *must* be planned before you look at the data. If you look at some data, pick out an interesting comparison, then analyze it as if it were a planned comparison, you will be committing scientific fraud. For example, if you look at the mean arch heights for the nine sports, see that cross-country has the lowest mean and swimming has the highest mean, then compare just those two means, your P-value will be much too low. This is because there are 36 possible pairwise comparisons in a set of 9 means. You expect 5 percent, or 1 out of 20, tests to be "significant" at the $P<0.05$ level, even if all the data really fit the null hypothesis, so there's a good chance that the most extreme comparison in a set of 36 will have a P-value less than 0.05.

It would be acceptable to run a pilot experiment and plan your planned comparisons based on the results of the pilot experiment. However, if you do this you could not include the data from the pilot experiment in the analysis; you would have to limit your anova to the new data.

How to do the tests

Spreadsheet

To do a planned comparison using the one-way anova spreadsheet, just combine and delete data from the original data set to make a new data set with the comparison you want, then paste it into the anova spreadsheet. If you're moving data around within the anova spreadsheet, use the "copy" and "paste" commands to copy the data to the new destination, followed by "clear" to clear it from the original location; if you use the "cut" and "paste" commands, it will change the

references in some of the formulas and mess things up. You might be better off doing your rearranging in a separate spreadsheet, then copying and pasting from there into the anova spreadsheet.

For example, look at the mussel shell data from the previous page. If one of your planned contrasts was "Oregon vs. North Pacific", you'd put the data from Newport, Oregon and Tillamook, Oregon into one column labelled "Oregon," put the data from Petersburg, Alaska and Magadan, Russia in a second column labelled "North Pacific," and delete the Tvarminne data. Putting these two columns of data into the anova spreadsheet gives $F_{1, 31} = 5.31$, $P = 0.029$, so you would conclude that there was a significant difference between the Oregon and North Pacific mussels.

To do non-orthogonal planned comparisons with the sequential Dunn–Sidák method, do each comparison and collect the P-values into a separate spreadsheet. Sort the P-values from smallest to largest, then see which ones meet the sequential Dunn–Sidák criteria described above.

Web pages

To do a planned comparison using a web page, just clear the web page (there's usually a button marked "Clear" or "Reset") and enter the data for whichever comparison you want to do. You may want to rearrange your data in a spreadsheet, then paste it into the web page.

SAS

To do planned comparisons in SAS, the simplest way would be to make a new data set in which you delete the lines you don't want, and give new group names to the lines you want to group together. For the mussel shell data from the previous page, if one of your planned contrasts was "Oregon vs. North Pacific", you could change "Newport" and "Tillamook" to "Oregon," change "Petersburg" and "Magadan" to "North_Pacific," and delete the Tvarminne data, then run PROC GLM on the modified data set. If you're experienced with SAS, you can figure out easier ways to do this, but this will work.

Further reading

Sokal and Rohlf, pp. 229-242.

One-way anova: Unplanned comparisons of means

In a Model I anova, it is often desirable to perform additional comparisons of subsets of the means. If you didn't decide on some planned comparisons before doing the anova, you will be doing unplanned comparisons. Because these are unplanned, you can't just do the comparison as an anova and use the resulting P-value. Instead you have to use a test that takes into account the large number of possible comparisons you *could* have done. For example, if you did an anova with five groups (A, B, C, D, and E), then noticed that A had the highest mean and D had the lowest, you couldn't do an anova on just A and D. There are 10 possible pairs you could have compared (A with B, A with C, etc.) and the probability under the null hypothesis that one of those 10 pairs is "significant" at the $p<0.05$ level is much greater than 0.05. It gets much worse if you consider all of the possible ways of dividing the groups into two sets (A vs. B, A vs. B+C, A vs. B+C+D, A+B vs. C+D, etc.) or more than two sets (A vs. B. vs C, A vs. B vs. C+D, etc.).

There is a bewildering array of tests that have been proposed for unplanned comparisons; some of the more popular include the Student–Neuman–Keuls (SNK) test, Duncan's multiple range test, the Tukey–Kramer method, the REGWQ method, and Fisher's Least Significant Difference (LSD). For this handbook, I am only covering two techniques, Gabriel comparison intervals and the Tukey–Kramer method, that apply only to unplanned comparisons of pairs of group means.

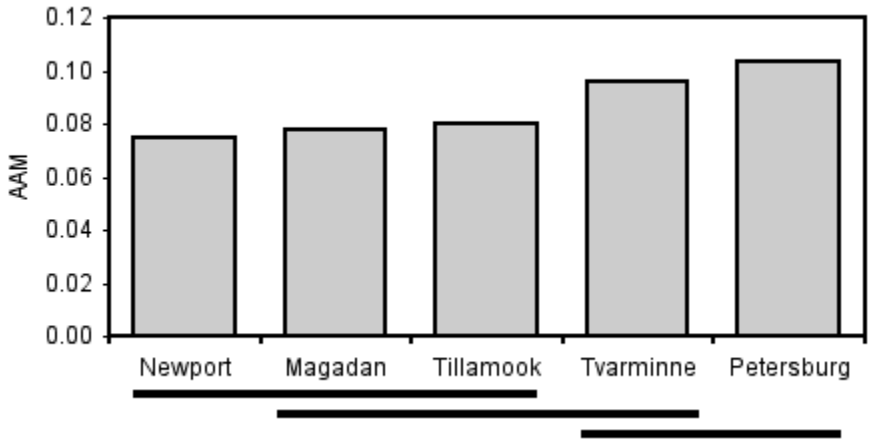
I will not consider tests that apply to unplanned comparisons of more than two means, or unplanned comparisons of subsets of groups. There are techniques available for this (the Scheffé test is probably the most common), but with a moderate number of groups, the number of possible comparisons becomes so large that the P-values required for significance become ridiculously small.

Gabriel comparison interval

To compute the Gabriel comparison interval (Gabriel 1978), the standard error of the mean for a group is multiplied by the studentized maximum modulus times the square root of one-half. The standard error of the mean is estimated by

dividing the MS_{within} from the entire anova by the number of observations in the group, then taking the square root of that quantity. The studentized maximum modulus is a statistic that depends on the number of groups, the total sample size in the anova, and the desired probability level (alpha).

Once the Gabriel comparison interval is calculated, the lower comparison limit is found by subtracting the interval from the mean, and the upper comparison limit is found by adding the interval to the mean. This is done for each group in an anova. Any pair of groups whose comparison intervals do not overlap is significantly different at the P<alpha level. For example, on the graph of the mussel data shown below, there is a significant difference in AAM between mussels from Newport and mussels from Petersburg. Tillamook and Newport do not have significantly different AAM, because their Gabriel comparison intervals overlap.



Mean AAM (anterior adductor muscle scar standardized by total shell length) for *Mytilus trossulus* from five locations. Means are shown with Gabriel comparison intervals (Gabriel 1978); pairs of means whose comparison intervals do not overlap are significantly different (P<0.05). Data from the one-way anova page.

I like Gabriel comparison intervals; the results are about the same as with other techniques for unplanned comparisons of pairs of means, but you can present them in a more easily understood form. However, Gabriel comparison intervals are not that commonly used. If you are using them, it is very important to emphasize that the vertical bars represent comparison intervals and not the more common (but less useful) standard errors of the mean or 95% confidence intervals. You must also explain that means whose comparison intervals do not overlap are significantly different from each other.

Tukey–Kramer method

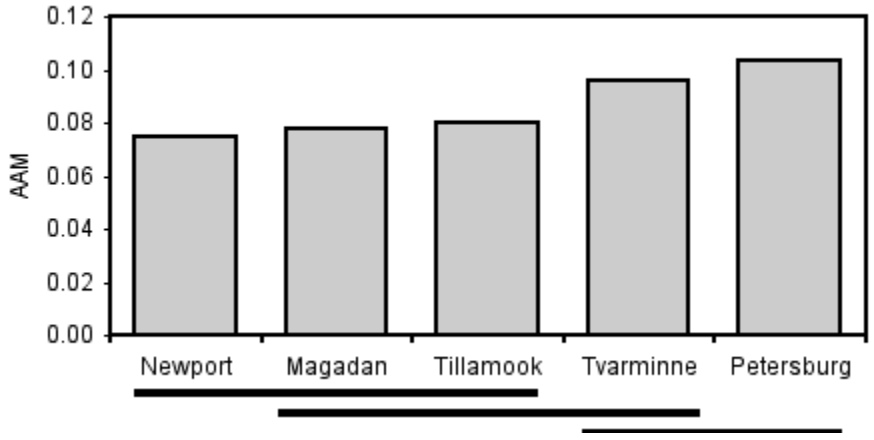
In the Tukey–Kramer method, the minimum significant difference (MSD) is calculated for each pair of means. If the observed difference between a pair of means is greater than the MSD, the pair of means is significantly different.

The Tukey–Kramer method is much more popular than Gabriel comparison intervals. It is not as easy to display the results of the Tukey–Kramer method, however. One technique is to find all the sets of groups whose means do not differ significantly from each other, then indicate each set with a different symbol, like this:

Location	mean AAM	Tukey–Kramer
Newport	0.0748	a
Magadan	0.0780	a, b
Tillamook	0.0802	a, b
Tvarminne	0.0957	b, c
Petersburg	0.103	c

Then you explain that "Means with the same letter are not significantly different from each other (Tukey–Kramer test, P<0.05)."

Another way that is used to illustrate the results of the Tukey–Kamer method is with lines connecting means that are not significantly different from each other. This is easiest when the means are sorted from smallest to largest:



Mean AAM (anterior adductor muscle scar standardized by total shell length) for *Mytilus trossulus* from five locations. Pairs of means grouped by a horizontal line are not significantly different from each other (Tukey–Kramer method, P>0.05).

How to do the tests

Spreadsheet

The one-way anova spreadsheet, described on the anova significance page, calculates Gabriel comparison intervals. The interval it reports is the number that is added to or subtracted from the mean to give the Gabriel comparison limits. The spreadsheet also does the Tukey–Kramer test at the alpha=0.05 level, if you have 20 or fewer groups. The results of the Tukey–Kramer test are shown on the second sheet of the workbook.

Web pages

I am not aware of any web pages that will calculate either Gabriel comparison intervals or do the Tukey–Kramer test.

SAS

To calculate Gabriel comparison limits using SAS, add a MEANS statement to PROC GLM. The first parameter after MEANS is the nominal variable, followed by a forward slash, then CLM and GABRIEL. CLM tells SAS to report the results of the Gabriel method as comparison limits. Here's the SAS program from the one-way anova web page, modified to present Gabriel comparison intervals:

```
proc glm data=musselshells;
  class location;
  model aam = location;
  means location / clm gabriel;
run;
```

The results are comparison limits. If you are graphing using a spreadsheet, you'll need to calculate the comparison interval, the difference between one of the comparison limits and the mean. For example, the comparison interval for Petersburg is 0.113475–0.103443=0.010032. This is what you put next to the mean on your spreadsheet, and you select it when you tell the spreadsheet what to add and subtract from the mean for the "error bars".

location	N	Mean	95% Comparison Limits	
Petersbu	7	0.103443	0.093411	0.113475
Tvarminn	6	0.095700	0.084864	0.106536
Tillamoo	10	0.080200	0.071806	0.088594
Magadan	8	0.078013	0.068628	0.087397
Newport	8	0.074800	0.065416	0.084184

For the Tukey–Kramer technique using SAS, add a MEANS statement to PROC GLM. The first parameter after MEANS is the nominal variable, followed by a forward slash, then LINES and TUKEY. LINES tells SAS to report the results of the Tukey–Kramer method by giving means that are not significantly different the

same letter. Here's the SAS program from the one-way anova web page, modified to do the Tukey–Kramer technique:

```
proc glm data=musselshells;
  class location;
  model aam = location;
  means location / lines tukey;
run;
```

Here's the output:

```
Means with the same letter are not significantly different.


      Tukey Grouping          Mean      N    location

                   A      0.103443      7    Petersbu
                   A
               B   A      0.095700      6    Tvarminn
               B
               B   C      0.080200     10    Tillamoo
               B   C
               B   C      0.078013      8    Magadan
                   C
                   C      0.074800      8    Newport
```

Further reading

Sokal and Rohlf, pp. 240-260 (unplanned comparisons in general), 247-249 (Gabriel comparison intervals).

Zar, pp. 208-222.

Reference

Gabriel, K.R. 1978. A simple method of multiple comparison of means. J. Amer. Stat. Assoc. 73: 724-729.

One-way anova: Partitioning variance components

In a Model II anova with significant heterogeneity among the means, the next step is to partition the variance into among-group and within-group components. Under the null hypothesis of homogeneity of means, the among-group mean square and within-group mean square are both estimates of the within-group parametric variance. If the means are heterogeneous, the within-group mean square is still an estimate of the within-group variance, but the among-group mean square estimates the sum of the within-group variance plus the group sample size times the added variance among groups. Therefore subtracting the within-group mean square from the among-group mean square, and dividing this difference by the average group sample size, gives an estimate of the added variance component among groups. The equation is:

$$\text{among-group variance} = (MS_{among} - MS_{within})/n_o$$

where n_o is a number that is close to, but usually slightly less than, the arithmetic mean of the sample size (n_i) of each of the *a* groups:

$$n_o = (1/(a-1)) * (\text{sum}(n_i) - (\text{sum}(n_i^2)/\text{sum}(n_i)).$$

Each component of the variance is often expressed as a percentage of the total variance components. Thus an anova table for a one-way anova would indicate the among-group variance component and the within-group variance component, and these numbers would add to 100%. In a more complicated anova, such as a nested anova, there would be several variance components, but they would still add up to 100%.

Here's an explanation that is not strictly correct (it obscures some of the mathematical details) but gives an intuitive idea of what it means to partition the variance. Here is a very simple anova, with just three observations in two categories:

A	B
10	4
12	5
8	3

First, calculate the mean for all six observations, which is 7. Subtracting 7 from each observation and squaring the difference gives you the squared deviates:

A	B
$(10-7)^2 = 9$	$(4-7)^2 = 9$
$(12-7)^2 = 25$	$(5-7)^2 = 4$
$(8-7)^2 = 1$	$(3-7)^2 = 16$

The sum of these squared deviates is the **total sum of squares.** In this case, the squared deviates add up to 64. This number is a measure of how far the individual observations are from the overall mean.

Next, calculate the mean for A, which is 10, and the mean for B, which is 4. Subtract each group's mean from the observations in that group and square the differences:

A	B
$(10-10)^2 = 0$	$(4-4)^2 = 0$
$(12-10)^2 = 4$	$(5-4)^2 = 1$
$(8-10)^2 = 4$	$(3-4)^2 = 1$

Notice that these squared deviates from the group means are, in general, smaller than the squared deviates from the overall mean. This makes sense; a member of a group is likely to be closer to its group's mean than it is to the mean of that group plus some other, different groups. Adding these squared deviates together gives us the **within-group sum of squares,** which in this case is 10. This is a measure of how far the individual observations are from their group means.

The difference between the total sum of squares and the within-group sum of squares is the **among-group sum of squares.** It is a measure of how much smaller the sum of squares gets when you use group means instead of the overall mean. When the group means are not very different from each other, they will all be close to the overall mean. In that case, the squared deviates from the group means will not be much smaller than the squared deviates from the overall mean, and the among-group sum of squares will be small. When the group means *are* very different from each other, the group means will be very different from the overall mean, the squared deviates from the group means will be a lot smaller, and the among-group sum of squares will be large.

The among-group sum of squares in this example is 64 minus 10, or 54, while the within-group sum of squares is 10. Expressed as a percentage of the total, the

among-group variation represents 54/64 = 84.4% of the total; another way of saying this is that the groups "explain" 84.4% of the variation. The remaining 15.6% of the variation is within groups.

Because the sums of squares are estimates of population parameters, converting them to estimates of the variance components is considerably more complicated; the actual estimate of the among-group component for this example is 87.3% of the total. But the basic idea, that a larger among-groups component indicates larger differences among the group means relative to the within-group variation, remains the same.

Although statisticians say that each level of an anova "explains" a proportion of the variation, this statistical jargon does not mean that you've found a biological cause-and-effect explanation. If you measure the number of ears of corn per stalk in 10 random locations in a field, analyze the data with a one-way anova, and say that the location "explains" 74.3% of the variation, you haven't really explained anything; you don't know whether some areas have higher yield because of different water content in the soil, different amounts of insect damage, different amounts of nutrients in the soil, or random attacks by a band of marauding corn bandits.

Partitioning the variance components is particularly useful in quantitative genetics, where the within-family component might reflect environmental variation while the among-family component reflects genetic variation. Of course, estimating heritability involves more than just doing a simple anova, but the basic concept is similar.

Another area where partitioning variance components is useful is in designing experiments. For example, let's say you're planning a big experiment to test the effect of different drugs on calcium uptake in rat kidney cells. You want to know how many rats to use, and how many measurements to make on each rat, so you do a pilot experiment in which you measure calcium uptake on 6 rats, with 4 measurements per rat. You analyze the data with a one-way anova and look at the variance components. If a high percentage of the variation is among rats, that would tell you that there's a lot of variation from one rat to the next, but the measurements within one rat are pretty uniform. You could then design your big experiment to include a lot of rats for each drug treatment, but not very many measurements on each rat. Or you could do some more pilot experiments to try to figure out why there's so much rat-to-rat variation (maybe the rats are different ages, or some have eaten more recently than others, or some have exercised more) and try to control it. On the other hand, if the among-rat portion of the variance was low, that would tell you that the mean values for different rats were all about the same, while there was a lot of variation among the measurements on each rat. You could design your big experiment with fewer rats and more observations per rat, or you could try to figure out why there's so much variation among measurements and control it better.

Partitioning the variance applies only to a model II one-way anova. It doesn't really tell you anything useful about a model I one-way anova, although sometimes people like to report it (because they're proud of how much of the variance their groups "explain," I guess).

Performing the analysis

Spreadsheet

The one-way anova spreadsheet calculates the within- and among-group components of variance and displays them as a percentage of the total.

Web pages

I don't know of any web pages that will calculate the variance components.

SAS

PROC GLM doesn't calculate the variance components for an anova. Instead, you use PROC VARCOMP. You set it up just like PROC GLM, with the addition of METHOD=TYPE1. The procedure has four different methods for estimating the variance components, and TYPE1 seems to be the same technique as the one I've described above. Here's how to do the one-way anova, including estimating the variance components, for the mussel shell example from the one-way anova page.

```
proc glm data=musselshells;
  class location;
  model aam = location;
proc varcomp data=musselshells method=type1;
  class location;
  model aam = location;
run;
```

The results include the following:

```
            Type 1 Estimates
Variance Component        Estimate

Var(location)            0.0001254
Var(Error)               0.0001587
```

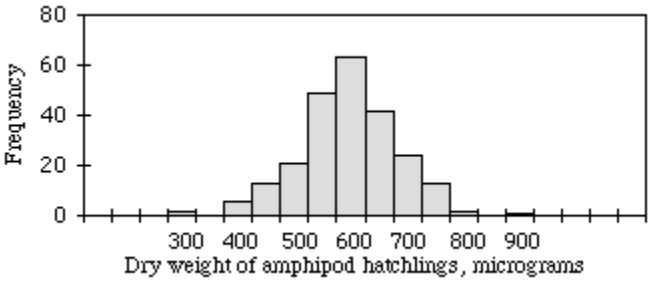
The output is not given as a percentage of the total, so you'll have to calculate that. For these results, the among-group component is 0.0001254/(0.0001254+0.0001586)=0.4415, or 44.15%; the within-group component is 0.0001587/(0.0001254+0.0001586)=0.5585, or 55.85%.

Further reading

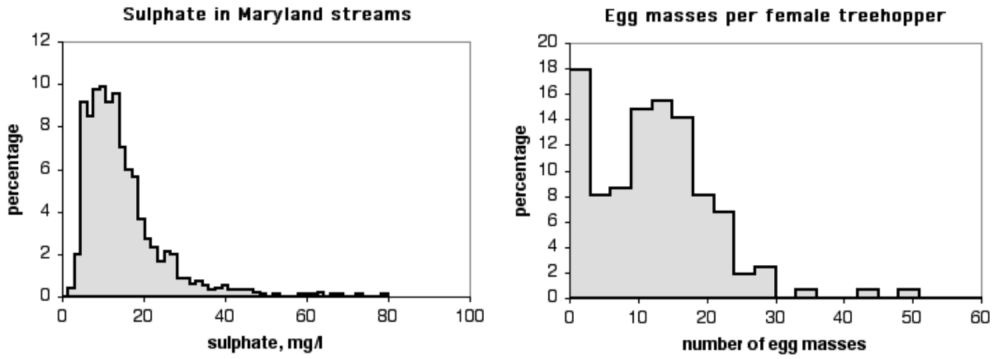
Sokal and Rohlf, pp. 194-197, 214.

Normality

One of the assumptions of an anova and other parametric tests is that the data are normally distributed. When you plot a frequency histogram, the frequencies should approximate the familiar bell-shaped normal curve. For example, the figure shown at the right is a histogram of dry weights of newly hatched amphipods (*Platorchestia platensis*). It fits the normal curve pretty well.



Histogram of dry weights of the amphipod crustacean *Platorchestia platensis*.



Two non-normal histograms.

Other data sets don't fit the normal curve very well. The histogram on the top is the level of sulphate in Maryland streams (data from the Maryland Biological Stream Survey). It doesn't fit the normal curve very well, because there are a small number of streams with very high levels of sulphate. The histogram on the bottom is the number of egg masses laid by indivuduals of the *lentago* host race of the treehopper *Enchenopa* (unpublished data courtesy of Michael Cast). The curve is bimodal, with one peak at around 14 egg masses and the other at zero.

Like other parametric tests, the analysis of variance assumes that the data fit the normal distribution. If your measurement variable is not normally distributed, you may be increasing your chance of a false positive result if you analyze the data with an anova or other test that assumes normality. Fortunately, an anova is not very sensitive to moderate deviations from normality; simulation studies, using a variety of non-normal distributions, have shown that the false positive rate is not affected very much by this violation of the assumption (Glass et al. 1972, Harwell et al. 1992, Lix et al. 1996). This is because when you take a large number of random samples from a population, the means of those samples are approximately normally distributed even when the population is not normal.

It is possible to test the goodness-of-fit of a data set to the normal distribution. I do not suggest that you do this, because many data sets that are significantly non-normal would be perfectly appropriate for an anova.

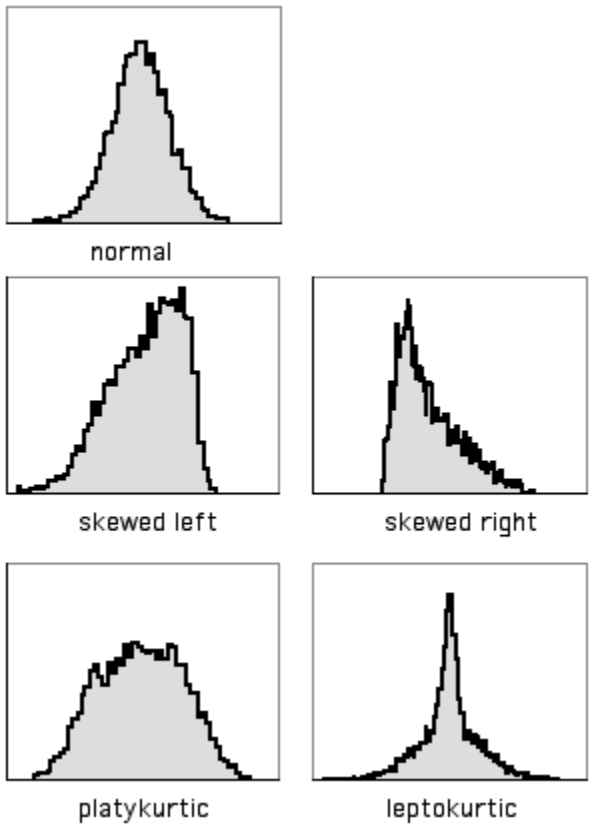
Instead, if you have a large enough data set, I suggest you just look at the frequency histogram. If it looks more-or-less normal, go ahead and perform an anova. If it looks like a normal distribution that has been pushed to one side, like the sulphate data above, you should try different data transformations and see if any of them make the histogram look more normal. If that doesn't work, and the data still look severely non-normal, it's probably still okay to analyze the data using an anova. However, you may want to analyze it using a non-parametric test. Just about every parametric statistical test has a non-parametric substitute, such as the Kruskal–Wallis test instead of a one-way anova, Wilcoxon signed-rank test instead of a paired t-test, and Spearman rank correlation instead of linear regression. These non-parametric tests do not assume that the data fit the normal distribution. They do assume that the data in different groups have the same distribution as each other, however; if different groups have different shaped distributions (for example, one is skewed to the left, another is skewed to the right), a non-parametric test may not be any better than a parametric one.

If you want to check the normality of data for an anova, but you don't have enough observations in each group, you can look at the residuals of all of the observations. Subtract the mean of each group from each observation in that group, then look at the histogram of the combined set of all residuals. This won't tell you whether there are differences in the shape of the distribution among groups, but it is better than nothing.

Skewness and kurtosis

A histogram with a long tail on the right side, such as the sulphate data above, is said to be skewed to the right; a histogram with a long tail on the left side is said to be skewed to the left. There is a statistic to describe skewness, g_1, but I don't know of any reason to calculate it; there is no rule of thumb that you shouldn't do an anova if g_1 is greater than some cutoff value.

Another way in which data can deviate from the normal distribution is kurtosis. A histogram that has a high peak in the middle and long tails on either side is leptokurtic; a histogram with a broad, flat middle and short tails is platykurtic. The statistic to describe kurtosis is g_2, but I can't think of any reason why you'd want to calculate it, either.



Graphs illustrating skewness and kurtosis.

How to look at normality

Spreadsheet

I've written a spreadsheet that will plot a frequency histogram for untransformed, log-transformed and square-root transformed data. It will handle up to 1000 observations.

If there are not enough observations in each group to check normality, you may want to examine the residuals (each observation minus the mean of its group). To do this, open a separate spreadsheet and put the numbers from each group in a separate column. Then create columns with the mean of each group subtracted from each observation in its group, as shown below. Copy these numbers into the histogram spreadsheet.

=B2-B$13

	A	B	C	D	E	F
1	**original data**	Tillamook	Newport	Petersburg	Magadan	Tvarminne
2		0.0571	0.0873	0.0974	0.1033	0.0703
3		0.0813	0.0662	0.1352	0.0915	0.1026
4		0.0831	0.0672	0.0817	0.0781	0.0956
5		0.0976	0.0819	0.1016	0.0685	0.0973
6		0.0817	0.0749	0.0968	0.0677	0.1039
7		0.0859	0.0649	0.1064	0.0697	0.1045
8		0.0735	0.0835	0.1050	0.0764	
9		0.0659	0.0725		0.0689	
10		0.0923				
11		0.0836				
12						
13	**group means**	0.0802	0.0748	0.1034	0.0780	0.0957
14						
15	**residuals**	-0.0231	0.0125	-0.0060	0.0253	-0.0254
16		0.0011	-0.0086	0.0318	0.0135	0.0069
17		0.0029	-0.0076	-0.0217	0.0001	-0.0001
18		0.0174	0.0071	-0.0018	-0.0095	0.0016
19		0.0015	0.0001	-0.0066	-0.0103	0.0082
20		0.0057	-0.0099	0.0030	-0.0083	0.0088
21		-0.0067	0.0087	0.0016	-0.0016	
22		-0.0143	-0.0023		-0.0091	
23		0.0121				
24		0.0034				

A spreadsheet showing the calculation of residuals.

Web pages

There are several web pages that will produce histograms, but most of them aren't very good. The interactive histogram (http://www.ruf.rice.edu/~lane/stat_analysis/histogram.html) web page is pretty cool. You enter your numbers (separated by spaces only, no tabs or line returns), and when you get a histogram, you can change the "binwidth" (the size of each interval on the histogram) by sliding a bar.

SAS

You can use the PLOTS option in PROC UNIVARIATE to get a stem-and-leaf display, which is a kind of very crude histogram. You can also use the HISTOGRAM option to get an actual histogram, but only if you know how to send the output to a graphics device driver.

Further reading

Sokal and Rohlf, pp. 698-703, 406-407.

Zar, pp. 185-188.

References

Glass, G.V., P.D. Peckham, and J.R. Sanders. 1972. Consequences of failure to meet assumptions underlying fixed effects analyses of variance and covariance. Rev. Educ. Res. 42: 237-288.

Harwell, M.R., E.N. Rubinstein, W.S. Hayes, and C.C. Olds. 1992. Summarizing Monte Carlo results in methodological research: the one- and two-factor fixed effects ANOVA cases. J. Educ. Stat. 17: 315-339.

Lix, L.M., J.C. Keselman, and H.J. Keselman. 1996. Consequences of assumption violations revisited: A quantitative review of alternatives to the one-way analysis of variance F test. Rev. Educ. Res. 66: 579-619.

Homoscedasticity and heteroscedasticity

One of the assumptions of an anova and other parametric tests is that the within-group variances of the groups are all the same (exhibit homoscedasticity). If the variances are different from each other (exhibit heteroscedasticity), the probability of obtaining a "significant" result even though the null hypothesis is true may be greater than the desired alpha level.

To illustrate this problem, I've done simulations of samples from three populations, all with the same population mean. I simulated taking samples of 10 observations from population A, 7 from population B, and 3 from population C, and repeated this process thousands of times. When the three populations are homoscedastic (have the same standard deviation), the one-way anova on the simulated data sets are significant ($P<0.05$) about 5 percent of the time, as they should be. However, when I make the standard deviations different (1.0 for population A, 2.0 for population B, and 3.0 for population C), I get a P value less than 0.05 in about 18 percent of the simulations. In other words, even though the population means are really all the same, my chance of getting a false positive result is 18 percent, not the desired 5 percent.

There have been a number of simulation studies that have tried to determine when heteroscedasticity is a big enough problem that other tests should be used. Early results suggested that heterscedasticity was not a problem if all the sample sizes were equal (Glass et al. 1972), but later results found that large amounts of heteroscedasticity can inflate the false positive rate, even when the sample sizes are equal (Harwell et al. 1992). The problem is made worse when the sample sizes are unequal and the smaller samples are from populations with larger variances; but when the smaller samples are from populations with smaller variances, the false positive rate can actually be much less than 0.05, meaning the power of the test is reduced (Glass et al. 1972).

Despite all of the simulation studies that have been done, there does not seem to be a consensus about when heteroscedasticity is a big enough problem that alternatives to anova should be used. I have written a spreadsheet to simulate one-way anova with heteroscedasticity that may help you decide this for a one-way anova or Student's t-test. To use it, enter the sample sizes for your observed data, plus the standard deviation within each group. Then hit the option, command, and

r keys simultaneously. This will run a macro that creates simulated data sets 1000 times. For each simulation, a set of observations is drawn at random for each group. Each group is normally distributed and has the population standard deviation that you've specified, and all the groups have the same population mean. A one-way anova is done on the simulated data (if you have only two groups, this is the same as Student's t-test), and the number of simulations that have a P-value less than 0.05 is recorded.

If the false positive rate in the simulations is near 0.05, then heteroscedasticity is probably not a problem for your data set. If the false positive rate is too high (defined as above 0.075, by Bradley's [1978] liberal criterion for robustness), you can try a data transformation; if that doesn't reduce the heteroscedasticity enough, you should use an alternative test that is less sensitive to heteroscedasticity.

Note that there is a problem with this simulation approach. Even if all the population standard deviations are the same, your sample standard deviations will be different from each other; with small sample sizes (fewer than 10), there can be quite a bit of variation among the sample standard deviations. Using your sample standard deviations as estimates of the population standard deviations in the simulations will therefore exaggerate the effects of heteroscedasticity. I don't know how to correct for this.

If the variances of your groups are very heterogeneous no matter what transformation you apply, there are a large number of alternative tests to choose from (Lix et al. 1996). The most commonly used is probably Welch's test, sometimes called Welch's t-test when there are two groups. Non-parametric tests, such as the Kruskal–Wallis test instead of a one-way anova, do not assume normality, but they do assume that the shapes of the distributions in different groups are the same, so they are not a good solution to the problem of heteroscedasticity.

All of the discussion above has been about one-way anovas. Homoscedasticity is also an assumption of other anovas, such as nested and two-way anovas, and regression and correlation. Much less work has been done on the effects of heteroscedasticity on these tests; all I can recommend is that you inspect the data for heteroscedasticity and hope that you don't find it.

Bartlett's test

The usual test for homogeneity of variances is Bartlett's test. This test is used when you have one measurement variable, one nominal variable, and you want to test the null hypothesis that the variances of the measurement variable are the same for the different groups. The basic idea is that the natural log of the variance is calculated for each group, then these are averaged. The variances are also averaged across groups. The average of the natural logs of the individual variances is subtracted from the natural log of the average variance. Under the null

hypothesis of homogeneity of variances, this statistic is chi-square distributed with d.f. equal to the number of groups minus one.

Bartlett's test is not a particularly good one, because it is sensitive to departures from normality as well as heteroscedasticity. It may be more helpful to use Bartlett's test to see what effect different transformations have on the heteroscedasticity, choosing the transformation with the highest (least significant) P-value, rather than take the P values too seriously; you shouldn't panic just because you have a significant Bartlett's test.

An alternative to Bartlett's test that I won't cover here is Levene's test (http://www.itl.nist.gov/div898/handbook/eda/section3/eda35a.htm) . It is less sensitive to departures from normality, but if the data are approximately normal, it is less powerful than Bartlett's test.

For many measurement variables in biology, the coefficient of variation is the same for different groups. When this occurs, groups with larger means will also have larger variances. For example, if the petal lengths of red roses and pink roses both have a coefficient of variation of 20%, and red roses have a mean petal length 10% longer than pink roses, then the standard deviation of red petal length is 10% larger than for pink petal length. This means the variance is 21% larger for red roses. This kind of heteroscedasticity, in which the variance increases as the mean increases, can usually be greatly reduced with the right data transformation.

While Bartlett's test is usually used when examining data to see if it's appropriate for a parametric test, there are times when testing the homogeneity of variances is the primary goal of an experiment. For example, let's say you want to know whether variation in stride length among runners is related to their level of experience—maybe as people run more, those who started with unusually long or short strides gradually converge on some ideal stride length. You could measure the stride length of non-runners, beginning runners, experienced amateur runners, and professional runners, with several individuals in each group, then use Bartlett's test to see whether there was significant heterogeneity in the variances.

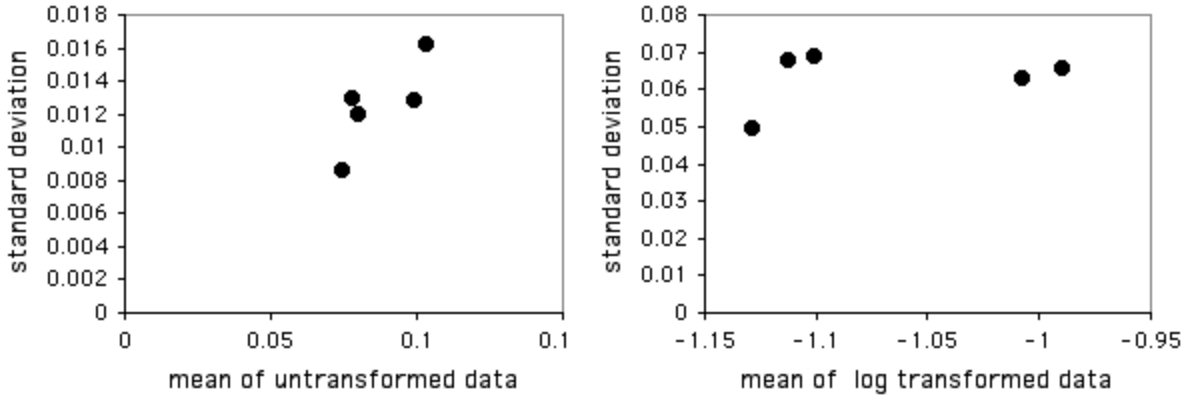
How to do Bartlett's test

Spreadsheet

I have put together a spreadsheet that performs Bartlett's test for homogeneity of variances for up to 1000 observations in each of up to 50 groups. It allows you to see what the log or square-root transformation will do. It also shows a graph of the standard deviations plotted vs. the means. This gives you a quick visual display of the difference in amount of variation among the groups, and it also shows whether the mean and standard deviation are correlated.

Entering the mussel shell data from the one-way anova web page into the spreadsheet, the P-values are 0.655 for untransformed data, 0.856 for square-root transformed, and 0.929 for log-transformed data. None of these is close to

significance, so there's no real need to worry. The graph of the untransformed data hints at a correlation between the mean and the standard deviation, so it might be a good idea to log-transform the data:



Standard deviation vs. mean AAM for untransformed and log-transformed data.

Web page

There is web page for Bartlett's test (http://home.ubalt.edu/ntsbarsh/Business-stat/otherapplets/BartletTest.htm) that will handle up to 14 groups. You have to enter the variances and sample sizes, not the raw data.

SAS

You can use the HOVTEST=BARTLETT option in the MEANS statement of PROC GLM to perform Bartlett's test. This modification of the program from the one-way anova page does Bartlett's test.

```
proc glm data=musselshells;
  class location;
  model aam = location;
  means location / hovtest=bartlett;
run;
```

Further reading

Sokal and Rohlf, pp. 398-399.

Zar, pp. 185, 202-204.

References

Bradley, J.V. 1978. Robustness? Brit. J. Math. Statis. Psychol. 31: 144-155.

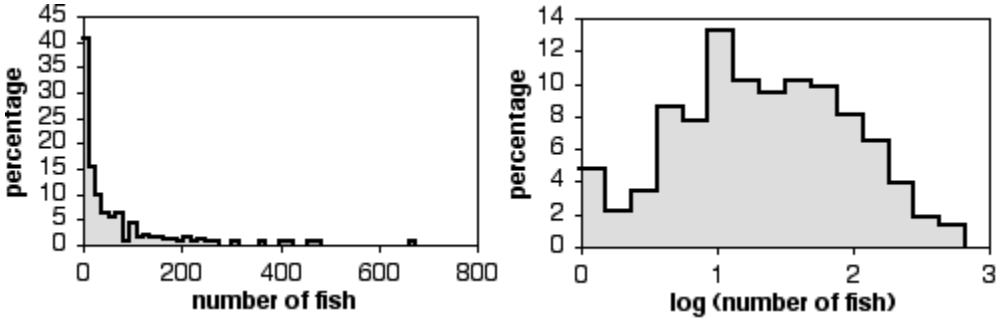
Glass, G.V., P.D. Peckham, and J.R. Sanders. 1972. Consequences of failure to meet assumptions underlying fixed effects analyses of variance and covariance. Rev. Educ. Res. 42: 237-288.

Harwell, M.R., E.N. Rubinstein, W.S. Hayes, and C.C. Olds. 1992. Summarizing Monte Carlo results in methodological research: the one- and two-factor fixed effects ANOVA cases. J. Educ. Stat. 17: 315-339.

Lix, L.M., J.C. Keselman, and H.J. Keselman. 1996. Consequences of assumption violations revisited: A quantitative review of alternatives to the one-way analysis of variance F test. Rev. Educ. Res. 66: 579-619.

Data transformations

Many biological variables do not meet the assumptions of parametric statistical tests: they are not normally distributed, the variances are not homogeneous, or both. Using a parametric statistical test (such as an anova or linear regression) on such data may give a misleading result. In some cases, transforming the data will make it fit the assumptions better.



Histograms of number of Eastern mudminnows per 75 m section of stream (samples with 0 mudminnows excluded). Untransformed data on left, log-transformed data on right.

To transform data, you perform a mathematical operation on each observation, then use these transformed numbers in your statistical test. For example, as shown in the first graph above, the abundance of the fish species *Umbra pygmaea* (Eastern mudminnow) in Maryland streams is non-normally distributed; there are a lot of streams with a small density of mudminnows, and a few streams with lots of them. Applying the log transformation makes the data more normal, as shown in the second graph.

To transform your data, apply a mathematical function to each observation, then use these numbers in your statistical test. Here are 12 numbers from the mudminnow data set; the first column is the untransformed data, the second column is the square root of the number in the first column, and the third column is the base-10 logarithm of the number in the first column.

Untransformed	Square-root transformed	Log transformed
38	6.164	1.580
1	1.000	0.000
13	3.606	1.114
2	1.414	0.301
13	3.606	1.114
20	4.472	1.301
50	7.071	1.699
9	3.000	0.954
28	5.292	1.447
6	2.449	0.778
4	2.000	0.602
43	6.557	1.633

You do the statistics on the transformed numbers. For example, the mean of the untransformed data is 18.9; the mean of the square-root transformed data is 3.89; the mean of the log transformed data is 1.044.

Back transformation

Even though you've done a statistical test on a transformed variable, such as the log of fish abundance, it is not a good idea to report your means, standard errors, etc. in transformed units. A graph that showed that the mean of the log of fish per 75 meters of stream was 1.044 would not be very informative for someone who can't do fractional exponents in their head. Instead, you should back-transform your results. This involves doing the opposite of the mathematical function you used in the data transformation. For the log transformation, you would back-transform by raising 10 to the power of your number. For example, the log transformed data above has a mean of 1.044 and a 95 percent confidence interval of 0.344 log-transformed fish. The back-transformed mean would be $10^{1.044}$=11.1 fish. The upper confidence limit would be $10^{(1.044+0.344)}$=24.4 fish, and the lower confidence limit would be $10^{(1.044-0.344)}$=5.0 fish. Note that the confidence limits are no longer symmetrical; the upper limit is 13.3 fish above the mean, while the lower limit is 6.1 fish below the mean. Also note that you can't just back-transform the confidence interval and add or subtract that from the back-transformed mean; you can't take $10^{0.344}$ and add or subtract that to 11.1.

Choosing the right transformation

Data transformations are an important tool for the proper statistical analysis of biological data. To those with a limited knowledge of statistics, however, they may seem a bit fishy, a form of playing around with your data in order to get the

answer you want. It is therefore essential that you be able to defend their use. There are an infinite number of transformations you could use, but it is better to use a transformation that is commonly used in your field, such as the square-root transformation for count data or the log transformation for size data, than an obscure transformation that not many people have heard of. It is also important that you decide which transformation to use before you do the statistical test. If you have a large number of observations, compare the effects of different transformations on the normality and the homoscedasticity of the variable. If you have a small number of observations, you may not be able to see much effect of the transformations on the normality and homoscedasticity; in that case, your decision to use a transformation will be based on the convention in your field for that kind of variable.

Common transformations

There are many transformations that are used occasionally in biology; here are three of the most common:

Log transformation**.** This consists of taking the log of each observation. You can use either base-10 logs (LOG in a spreadsheet, LOG10 in SAS) or base-*e* logs, also known as natural logs (LN in a spreadsheet, LOG in SAS). It makes no difference for a statistical test whether you use base-10 logs or natural logs, because they differ by a constant factor; the base-10 log of a number is just 2.303…× the natural log of the number. You should specify which log you're using when you write up the results, as it will affect things like the slope and intercept in a regression. I prefer base-10 logs, because it's possible to look at them and see the magnitude of the original number: log(1)=0, log(10)=1, log(100)=2, etc.

The back transformation is to raise 10 or *e* to the power of the number. If you have zeros or negative numbers, you can't take the log; you should add a constant to each number to make them positive and non-zero. If you have count data, and some of the counts are zero, the convention is to add 0.5 to each number.

Many variables in biology have log-normal distributions, meaning that after log-transformation, the values are normally distributed. This is because if you take a bunch of independent factors and multiply them together, the resulting product is log-normal. For example, let's say you've planted a bunch of maple seeds, then 10 years later you see how tall the trees are. The height of an individual tree would be affected by the nitrogen in the soil, the amount of water, amount of sunlight, amount of insect damage, etc. Having more nitrogen might make a tree 10% larger than one with less nitrogen; the right amount of water might make it 30% larger than one with too much or too little water; more sunlight might make it 20% larger; less insect damage might make it 15% larger, etc. Thus the final size of a tree would be a function of nitrogen×water×sunlight×insects, and mathematically, this kind of function turns out to be log-normal.

Square-root transformation**.** This consists of taking the square root of each observation. The back transformation is to square the number. If you have negative numbers, you can't take the square root; you should add a constant to each number to make them all positive.

The square-root transformation is commonly used when the variable is a count of something, such as bacterial colonies per petri dish, blood cells going through a capillary per minute, mutations per generation, etc.

Arcsine transformation**.** This consists of taking the arcsine of the square root of a number. (The result is given in radians, not degrees, and can range from $-\pi/2$ to $\pi/2$.) The numbers to be arcsine transformed must be in the range −1 to 1. This is commonly used for proportions, which range from 0 to 1, such as the proportion of female Eastern mudminnows that are infested by a parasite. Note that this kind of proportion is really a nominal variable, so it is incorrect to treat it as a measurement variable, whether or not you arcsine tranform it. For example, it would be incorrect to count the number of mudminnows that are or are not parasitized each of several streams in Maryland, treat the arcsine-transformed proportion of parasitized females in each stream as a measurement variable, then perform a linear regression on these data vs. stream depth. This is because the proportions from streams with a smaller sample size of fish will have a higher variance than proportions from streams with larger samples of fish, information that is disregarded when treating the arcsine-transformed proportions as measurement variables. Instead, you should use a test designed for nominal variables; in this example, you should do logistic regression instead of linear regression. If you insist on using the arcsine transformation, despite what I've just told you, the back-transformation is to square the sine of the number.

How to transform data

Spreadsheet

In a blank column, enter the appropriate function for the transformation you've chosen. For example, if you want to transform numbers that start in cell A2, you'd go to cell B2 and enter =LOG(A2) or =LN(A2) to log transform, =SQRT(A2) to square-root transform, or =ASIN(SQRT(A2)) to arcsine transform. Then copy cell B2 and paste into all the cells in column B that are next to cells in column A that contain data. To copy and paste the transformed values into another spreadsheet, remember to use the "Paste Special..." command, then choose to paste "Values." Using the "Paste Special...Values" command makes Excel copy the numerical result of an equation, rather than the equation itself. (If your spreadsheet is Calc, choose "Paste Special" from the Edit menu, uncheck the boxes labelled "Paste All" and "Formulas," and check the box labelled "Numbers.")

To back-transform data, just enter the inverse of the function you used to transform the data. To back-transform log transformed data in cell B2, enter

=10^B2 for base-10 logs or =EXP^B2 for natural logs; for square-root transformed data, enter =B2^2; for arcsine transformed data, enter =(SIN(B2))^2.

Web pages

I'm not aware of any web pages that will do data transformations.

SAS

To transform data in SAS, read in the original data, then create a new variable with the appropriate function. This example shows how to create two new variables, square-root transformed and log transformed, of the mudminnow data.

```
data mudminnow;
   input location $ banktype $ count;
   countlog=log10(count);
   countsqrt=sqrt(count);
   cards;
Gwynn_1       forest 38
Gwynn_2       urban   1
Gwynn_3       urban  13
Jones_1       urban   2
Jones_2       forest 13
LGunpowder_1  forest 20
LGunpowder_2  field  50
LGunpowder_3  forest  9
BGunpowder_1  forest 28
BGunpowder_2  forest  6
BGunpowder_3  forest  4
BGunpowder_4  field  43
;
```

The dataset "mudminnow" contains all the original variables (LOCATION, BANKTYPE and COUNT) plus the new variables (COUNTLOG and COUNTSQRT). You then run whatever PROC you want and analyze these variables just like you would any others. Of course, this example does two different transformations only as an illustration; in reality, you should decide on one transformation before you analyze your data.

The function for arcsine-transforming X is ARSIN(SQRT(X)).

You'll probably find it easiest to backtransform using a spreadsheet or calculator, but if you really want to do everything in SAS, the function for taking 10 to the X power is 10**X; the function for taking *e* to a power is EXP(X); the function for squaring X is X**2; and the function for backtransforming an arcsine transformed number is SIN(X)**2.

Further reading

Sokal and Rohlf, pp. 409-422.

Zar, pp. 273-280.

Kruskal–Wallis test and Mann–Whitney U test

When to use them

The Kruskal–Wallis test is most commonly used when there is one nominal variable and one measurement variable, and the measurement variable does not meet the normality assumption of an anova. It is the non-parametric analogue of a one-way anova. A one-way anova may yield inaccurate estimates of the P-value when the data are very far from normally distributed. The Kruskal–Wallis test does not make assumptions about normality. Like most non-parametric tests, it is performed on ranked data, so the measurement observations are converted to their ranks in the overall data set: the smallest value gets a rank of 1, the next smallest gets a rank of 2, and so on. The loss of information involved in substituting ranks for the original values can make this a less powerful test than an anova, so the anova should be used if the data meet the assumptions.

If the original data set actually consists of one nominal variable and one ranked variable, you cannot do an anova and must use the Kruskal–Wallis test.

The Mann–Whitney U-test (also known as the Mann–Whitney–Wilcoxon test, the Wilcoxon rank-sum test, or the Wilcoxon two-sample test) is limited to nominal variables with only two values; it is the non-parametric analogue to Student's t-test. It uses a different test statistic (*U* instead of the *H* of the Kruskal–Wallis test), but the P-value is mathematically identical to that of a Kruskal–Wallis test. For simplicity, I will only refer to Kruskal–Wallis on the rest of this web page, but everything also applies to the Mann–Whitney U-test.

Null hypothesis

The null hypothesis is that the samples come from populations such that the probability that a random observation from one group is greater than a random observation from another group is 0.5.

The Kruskal–Wallis test does *not* test the null hypothesis that the populations have identical means, which is the null hypothesis of a one-way anova. It is therefore incorrect to say something like "The mean amount of substance X was significantly higher in muscle tissue than in liver (Kruskal–Wallis test, P=0.012)." It

also does not test the null hypothesis that the populations have equal medians, although you will see this error many places, including some statistics textbooks. To illustrate this point, I made up these three sets of numbers. They have identical means (43.5), and identical medians (27.5), but the mean ranks are different (34.6, 27.5, and 20.4, respectively), resulting in a significant (P=0.025) Kruskal–Wallis test:

Group 1	Group 2	Group 3
1	10	19
2	11	20
3	12	21
4	13	22
5	14	23
6	15	24
7	16	25
8	17	26
9	18	27
46	37	28
47	58	65
48	59	66
49	60	67
50	61	68
51	62	69
52	63	70
53	64	71
342	193	72

Assumptions

The Kruskal–Wallis test does NOT assume that the data are normally distributed; that is its big advantage. It DOES, however, assume that the observations in each group come from populations with the same shape of distribution, so if different groups have have different shapes (one is skewed to the right and another is skewed to the left, for example, or they have different variances), the Kruskal–Wallis test may give inaccurate results (Fagerland and Sandvik 2009). I don't know what to suggest in that situation; maybe you could look into some kind of bootstrap analysis.

Heteroscedasticity is one way in which different groups can have different shaped distributions. If the distributions are normally shaped but highly heteroscedastic, you can use Welch's t-test for two groups, or Welch's anova for more than two groups. If the distributions are both non-normal and highly heteroscedastic, I don't know what to recommend.

How the test works

When working with a measurement variable, the Kruskal–Wallis test starts by substituting the rank in the overall data set for each measurement value. The

smallest value gets a rank of 1, the second-smallest gets a rank of 2, etc. Tied observations get average ranks; thus if there were four identical values occupying the fifth, sixth, seventh and eighth smallest places, all would get a rank of 6.5.

The sum of the ranks is calculated for each group, then the test statistic, H, is calculated. H is given by a rather formidable formula that basically represents the variance of the ranks among groups, with an adjustment for the number of ties. H is approximately chi-square distributed, meaning that the probability of getting a particular value of H by chance, if the null hypothesis is true, is the P value corresponding to a chi-square equal to H; the degrees of freedom is the number of groups minus 1.

If the sample sizes are too small, H does not follow a chi-squared distribution very well, and the results of the test should be used with caution. N less than 5 in each group seems to be the accepted definition of "too small."

A significant Kruskal–Wallis test may be followed up by unplanned comparisons of mean ranks, analogous to the Tukey-Kramer method for comparing means. There is an online calculator for computing the Least Significant Difference in ranks.

Examples

Bolek and Coggins (2003) collected multiple individuals of the toad *Bufo americanus*,, the frog *Rana pipiens,* and the salamander *Ambystoma laterale* from a small area of Wisconsin. They dissected the amphibians and counted the number of parasitic helminth worms in each individual. There is one measurement variable (worms per individual amphibian) and one nominal variable (species of amphibian), and the authors did not think the data fit the assumptions of an anova. The results of a Kruskal–Wallis test were significant ($H=63.48$, 2 d.f., $P=1.6 \times 10^{-14}$); the mean ranks of worms per individual are significantly different among the three species.

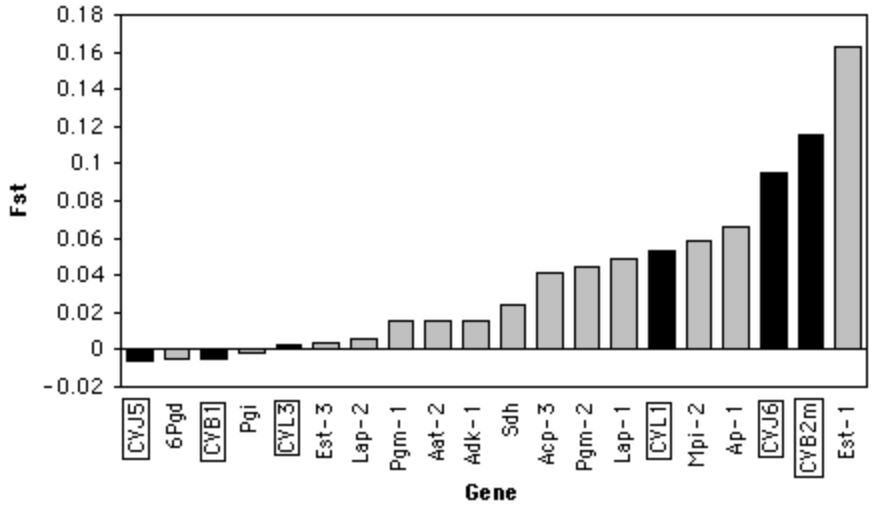
McDonald et al. (1996) examined geographic variation in anonymous DNA polymorphisms (variation in random bits of DNA of no known function) in the American oyster, *Crassostrea virginica*. They used an estimator of Wright's F_{ST} as a measure of geographic variation. They compared the F_{ST} values of the six DNA polymorphisms to F_{ST} values on 13 proteins from Buroker (1983). The biological question was whether protein polymorphisms would have generally lower or higher F_{ST} values than anonymous DNA polymorphisms; if so, it would suggest that natural selection could be affecting the protein polymorphisms. F_{ST} has a theoretical distribution that is highly skewed, so the data were analyzed with a Mann–Whitney U-test.

gene	class	F_{ST}
CVB1	DNA	-0.005
CVB2m	DNA	0.116
CVJ5	DNA	-0.006
CVJ6	DNA	0.095
CVL1	DNA	0.053
CVL3	DNA	0.003
6Pgd	protein	-0.005
Aat-2	protein	0.016
Acp-3	protein	0.041
Adk-1	protein	0.016
Ap-1	protein	0.066
Est-1	protein	0.163
Est-3	protein	0.004
Lap-1	protein	0.049
Lap-2	protein	0.006
Mpi-2	protein	0.058
Pgi	protein	-0.002
Pgm-1	protein	0.015
Pgm-2	protein	0.044
Sdh	protein	0.024

The results were not significant (U=0.21, P=0.84), so the null hypothesis that the F_{ST} of DNA and protein polymorphisms have the same mean ranks is not rejected.

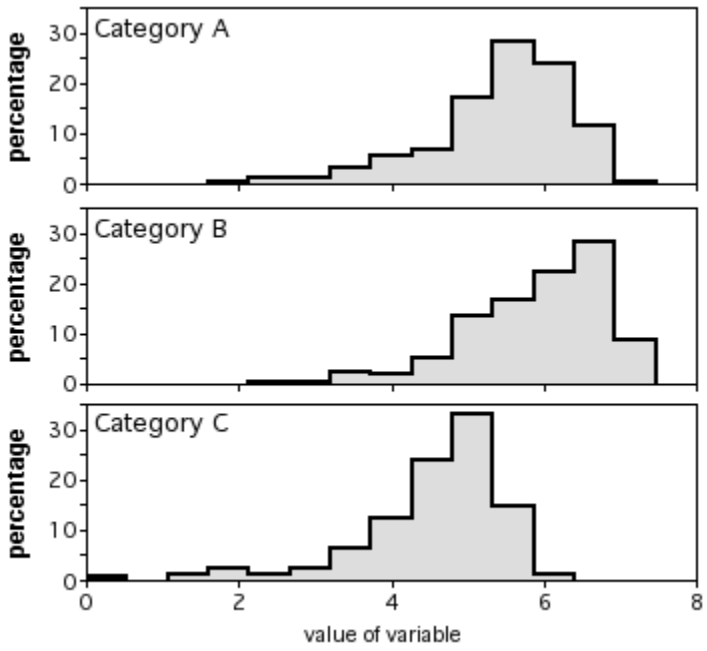
Graphing the results

It is tricky to know how to visually display the results of a Kruskal–Wallis test. It would be misleading to plot the means or medians on a bar graph, as the Kruskal–Wallis test is not a test of the difference in means or medians. If there are relatively small number of observations, you could put the individual observations on a bar graph, with the value of the measurement variable on the Y axis and its rank on the X axis, and use a different pattern for each value of the nominal variable. Here's an example using the oyster F_{st} data:



F_{st} values for DNA and protein polymorphisms in the American oyster. Names of DNA polymorphisms have a box around them.

If there are larger numbers of observations, you could plot a histogram for each category, all with the same scale, and align them vertically. I don't have suitable data for this handy, so here's an illustration with imaginary data:



Histograms of three sets of numbers.

Similar tests

One-way anova is more powerful and easier to understand than the Kruskal–Wallis test, so it should be used unless the data are severely non-normal. There is no firm rule about how non-normal data can be before an anova becomes inappropriate.

If the data are normally distributed but heteroscedastic, you can use Welch's t-test for two groups, or Welch's anova for more than two groups.

How to do the test

Spreadsheet

I have put together a spreadsheet to do the Kruskal–Wallis test on up to 20 groups, with up to 1000 observations per group.

Web pages

Richard Lowry has web pages for performing the Kruskal–Wallis test for two groups (http://faculty.vassar.edu/lowry/utest.html) , three groups (http://faculty.vassar.edu/lowry/kw3.html) , or four groups (http://faculty.vassar.edu/lowry/kw4.html) .

SAS

To do a Kruskal–Wallis test in SAS, use the NPAR1WAY procedure (that's the numeral "one," not the letter "el," in NPAR1WAY). "Wilcoxon" tells the procedure to only do the Kruskal–Wallis test; if you leave that out, you'll get several other statistical tests as well, tempting you to pick the one whose results you like the best. The nominal variable that gives the group names is given with the "class" parameter, while the measurement or rank variable is given with the "var" parameter. Here's an example, using the oyster data from above:

```
data oysters;
 input markername $ markertype $ fst;
 cards;
CVB1   DNA      -0.005
CVB2m  DNA       0.116
CVJ5   DNA      -0.006
CVJ6   DNA       0.095
CVL1   DNA       0.053
CVL3   DNA       0.003
6Pgd   protein  -0.005
Aat-2  protein   0.016
Acp-3  protein   0.041
Adk-1  protein   0.016
Ap-1   protein   0.066
Est-1  protein   0.163
Est-3  protein   0.004
```

```
Lap-1  protein   0.049
Lap-2  protein   0.006
Mpi-2  protein   0.058
Pgi    protein  -0.002
Pgm-1  protein   0.015
Pgm-2  protein   0.044
Sdh    protein   0.024
;
proc npar1way data=oysters wilcoxon;
 class markertype;
 var fst;
 run;
```

The output contains a table of "Wilcoxon scores"; the "mean score" is the mean rank in each group, which is what you're testing the homogeneity of. "Chi-square" is the H-statistic of the Kruskal–Wallis test, which is approximately chi-square distributed. The "Pr > Chi-Square" is your P-value. You would report these results as "H=0.04, 1 d.f., P=0.84."

```
          Wilcoxon Scores (Rank Sums) for Variable fst
            Classified by Variable markertype

                     Sum of   Expected       Std Dev         Mean
markertype    N      Scores   Under H0      Under H0        Score
-----------------------------------------------------------------
DNA           6       60.50       63.0     12.115236     10.083333
protein      14      149.50      147.0     12.115236     10.678571


                  Kruskal-Wallis Test

                Chi-Square         0.0426
                DF                      1
                Pr > Chi-Square    0.8365
```

Power analysis

I am not aware of a technique for estimating the sample size needed for a Kruskal–Wallis test.

Further reading

Sokal and Rohlf, pp. 424-426.

Zar, pp. 195-200.

References

Bolek, M.G., and J.R. Coggins. 2003. Helminth community structure of sympatric eastern American toad, *Bufo americanus americanus,* northern leopard frog, *Rana pipiens,* and blue-spotted salamander, *Ambystoma laterale,* from southeastern Wisconsin. J. Parasit. 89: 673-680.

Buroker, N. E. 1983. Population genetics of the American oyster *Crassostrea virginica* along the Atlantic coast and the Gulf of Mexico. Mar. Biol. 75:99-112.

Fagerland, M.W., and L. Sandvik. 2009. The Wilcoxon-Mann-Whitney test under scrutiny. Statist. Med. 28: 1487-1497.

McDonald, J.H., B.C. Verrelli and L.B. Geyer. 1996. Lack of geographic variation in anonymous nuclear polymorphisms in the American oyster, *Crassostrea virginica.* Molecular Biology and Evolution 13: 1114-1118.

Nested anova

When to use it

You use a nested anova when you have one measurement variable and two or more nominal variables. The nominal variables are nested, meaning that each value of one nominal variable (the subgroups) is found in combination with only one value of the higher-level nominal variable (the groups). The top-level nominal variable may be either Model I or Model II, but the lower-level nominal variables must all be Model II.

Nested analysis of variance is an extension of one-way anova in which each group is divided into subgroups. In theory, these subgroups are chosen randomly from a larger set of possible subgroups. For example, let's say you are testing the null hypothesis that stressed and unstressed rats have the same glycogen content in their gastrocnemius muscle. If you had one cage containing several stressed rats, another cage containing several unstressed rats, and one glycogen measurement from each rat, you would analyze the data using a one-way anova. However, you wouldn't know whether a difference in glycogen levels was due to the difference in stress, or some other difference between the cages--maybe the cage containing stressed rats gets more food, or is warmer, or happens to contain a mean rat who uses rat mind-control techniques to enslave all the other rats in the cage and get them to attack humans.

If, however, you had several cages of stressed rats and several cages of unstressed rats, with several rats in each cage, you could tell how much variation was among cages and how much was between stressed and unstressed. The groups would be stressed vs. unstressed, and each cage of several rats would be a subgroup; each glycogen level of a rat would be one observation within a subgroup.

The above is an example of a two-level nested anova; one level is the groups, stressed vs. unstressed, while another level is the subgroups, the different cages. If you worry about the accuracy of your glycogen assay, you might make multiple assays on each rat. In that case you would have a three-level nested anova, with groups (stressed vs. unstressed), subgroups (cages), and subsubgroups (the set of observations on each rat would be a subsubgroup). You can have more levels, too.

Note that if the subgroups, subsubgroups, etc. are distinctions with some interest (Model I), rather than random, you should not use a nested anova. For

example, you might want to divide the stressed rats into male and female subgroups, and the same for the unstressed rats. Male and female are not distinctions without interest; you would be interested to know that one sex had higher glycogen levels than the other. In this case you would use a two-way anova to analyze the data, rather than a nested anova.

Sometimes the distinction can be subtle. For example, let's say you measured the glycogen content of the right gastrocnemius muscle and left gastrocnemius muscle from each rat. If you think there might be a consistent right vs. left difference, you would use a two-way anova to analyze right vs. left and stressed vs. unstressed. If, however, you think that any difference between the two muscles of an individual rat is due to random variation in your assay technique, not a real difference between right and left, you could use a nested anova, with muscles as one level. Think of it this way: if you dissected out the muscles, labeled the tubes "A" and "B," then forgot which was right and which was left, it wouldn't matter if you were doing a nested anova; it would be a disaster if you were doing a two-way anova.

Null hypotheses

A nested anova has one null hypothesis for each level. In a two-level nested anova, one null hypothesis would be that the subgroups within each group have the same means; the second null hypothesis would be that the groups have the same means.

Assumptions

Nested anova, like all anovas, assumes that the observations within each subgroup are normally distributed and have equal variances.

How the test works

Remember that in a one-way anova, the test statistic, F_s, is the ratio of two mean squares: the mean square among groups divided by the mean square within groups. If the variation among groups (the group mean square) is high relative to the variation within groups, the test statistic is large and therefore unlikely to occur by chance. In a two-level nested anova, there are two F statistics, one for subgroups ($F_{subgroup}$) and one for groups (F_{group}). The subgroup F-statistic is found by dividing the among-subgroup mean square, $MS_{subgroup}$ (the average variance of subgroup means within each group) by the within-subgroup mean square, MS_{within} (the average variation among individual measurements within each subgroup). The group F-statistic is found by dividing the among-group mean square, MS_{group} (the variation among group means) by $MS_{subgroup}$. The P-value is then calculated for the F-statistic at each level.

For a nested anova with three or more levels, the F-statistic at each level is calculated by dividing the MS at that level by the MS at the level immediately below it.

If the subgroup F-statistic is not significant, it is possible to calculate the group F-statistic by dividing MS_{group} by MS_{pooled}, a combination of $MS_{subgroup}$ and MS_{within}. The conditions under which this is acceptable are complicated, and some statisticians think you should never do it; for simplicity, I suggest always using MS_{group} / $MS_{subgroup}$ to calculate F_{group}.

Partitioning variance

In addition to testing the equality of the means at each level, a nested anova also partitions the variance into different levels. This can be a great help in designing future experiments. For example, let's say you did a four-level nested anova with stressed vs. unstressed as groups, cages as subgroups, individual rats as subsubgroups, and the two gastrocnemius muscles as subsubsubgroups, with multiple glycogen assays per muscle. If most of the variation is among rats, with relatively little variation among muscles or among assays on each muscle, you might want to do just one assay per rat and use a lot more rats in your next experiment. This would give you greater statistical power than taking repeated measurements on a smaller number of rats. If the nested anova tells you there is variation among cages, you would either want to use more cages or try to control whatever variable is causing the cages to differ in the glycogen content of their rats; maybe the exercise wheel is broken in some of the cages, or maybe some cages have more rats than others. If you had an estimate of the relative cost of different parts of the experiment, such as keeping more rats vs. doing more muscle preps, formulas are available to help you design the most statistically powerful experiment for a given amount of money; see Sokal and Rohlf, pp. 309-317.

Mixed-model vs. pure Model II nested anova

All of the subgroups, subsubgroups, etc. in a nested anova should be based on distinctions of no inherent interest, of the kind analyzed with a Model II one-way anova. The groups at the top level may also be of no inherent interest, in which case it is a pure Model II nested anova. This often occurs in quantitative genetics. For example, if you are interested in estimating the heritability of ammonia content in chicken manure, you might have several roosters, each with several broods of chicks by different hens, with each chick having several ammonia assays of its feces. The offspring of each rooster would be the groups, the offspring of each hen would be the subgroups, and the set of ammonia assays on each chick would be subsubgroups. This would be a pure Model II anova, because you would want to know what proportion of the total variation in ammonia content was due to variation among roosters, as a way of estimating heritability; you wouldn't be interested in which rooster had offspring with the lowest or highest ammonia

content in their feces. In a pure model II nested anova, partitioning the variance is of primary importance.

If the top-level groups are of inherent interest, of the kind analyzed with a Model I one-way anova, then it is a mixed-model nested anova. The stressed vs. unstressed rat example is a mixed-model anova, because stressed vs. unstressed is what you are interested in. The ammonia in chicken feces example could also be analyzed using a mixed-model nested anova, *if* you were really interested in knowing which rooster had offspring with the lowest ammonia in their feces. This might be the case if you were going to use the best rooster to sire the next generation of chickens at your farm. In a mixed-model nested anova, partitioning the variance is of less interest than the significance test of the null hypothesis that the top-level groups have the same mean. You can then do planned comparisons among the top-level means, just as you would for a one-way anova, or the Tukey-Kramer test or Gabriel comparison intervals for unplanned comparisons. Even in a mixed model nested anova, partitioning the variance may help you design better experiments by revealing which level needs to be controlled better or replicated more.

Unequal sample sizes

When the sample sizes in a nested anova are unequal, the P-values corresponding to the F-statistics may not be very good estimates of the actual probability. For this reason, you should try to design your experiments with a "balanced" design, equal sample sizes in each subgroup. Often this is impractical; if you do have unequal sample sizes, you may be able to get a better estimate of the correct P-value by using modified mean squares at each level, found using a correction formula called the Satterthwaite approximation. Under some situations, however, the Satterthwaite approximation will make the P-values *less* accurate. If the Satterthwaite approximation cannot be used, the P-values will be conservative (less likely to be significant than they ought to be). Note that the Satterthwaite approximation results in fractional degrees of freedom, such as 2.87.

Examples

Keon and Muir (2002) wanted to know whether habitat type affected the growth rate of the lichen *Usnea longissima*. They weighed and transplanted 30 individuals into each of 12 sites in Oregon. The 12 sites were grouped into 4 habitat types, with 3 sites in each habitat. One year later, they collected the lichens, weighed them again, and calculated the change in weight. There are two nominal variables (site and habitat type), with sites nested within habitat type. One could analyze the data using two measurement variables, beginning weight and ending weight, but because the lichen individuals were chosen to have similar beginning weights, it makes more sense to use the change in weight as a single measurement

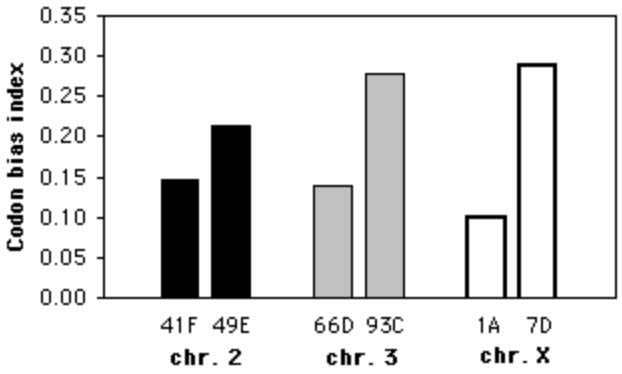
variable. The results of a mixed-model nested anova are that there is significant variation among sites within habitats ($F_{8, 200}$=8.11, P=1.8 x 10^{-9}) and significant variation among habitats ($F_{3, 8}$=8.29, P=0.008). When the Satterthwaite approximation is used, the test of the effect of habitat is only slightly different ($F_{3, 8.13}$=8.76, P=0.006)

Students in my section of Advanced Genetics Lab collected data on the codon bias index (CBI), a measure of the nonrandom use of synonymous codons from genes in *Drosophila melanogaster*. The groups are three chromosomes, and the subgroups are small regions within each chromosome. Each observation is the CBI value for a single gene in that chromosome region, and there were several genes per region. The data are shown below in the SAS program.

The results of the nested anova were $F_{3, 30}$=6.92, P=0.001 for subgroups and $F_{2, 3}$=0.10, P=0.91 for groups, without the Satterthwaite correction; using the correction changes the results only slightly. The among-subgroup variation is 49.9% of the total, while the among-group variation is 0%. The conclusion is that there is a lot of variation in CBI among different regions within a chromosome, so in order to see whether there is any difference among the chromosomes, it will be necessary to sample a lot more regions on each chromosome. Since 50.1% of the variance is among genes within regions, it will be necessary to sample several genes within each region, too.

Graphing the results

The way you graph the results of a nested anova depends on the outcome and your biological question. If the variation among subgroups is not significant and the variation among groups is significant—you're really just interested in the groups, and you used a nested anova to see if it was okay to combine subgroups—you might just plot the group means on a bar graph, as shown for one-way anova. If the variation among subgroups is interesting, you can plot the means for each subgroup, with different patterns or colors indicating the different groups. Here's an example for the codon bias data:



Graph of mean codon bias index in different regions of *Drosophila melanogaster* chromosomes. Solid black bars are regions in chromosome 2, gray bars are chromosome 3, and empty bars are the X chromosome.

Similar tests

Both nested anova and two-way anova (and higher level anovas) have one measurement variable and more than one nominal variable. The difference is that in a two-way anova, the values of each nominal variable are found in all combinations with the other nominal variable; in a nested anova, each value of one nominal variable (the subgroups) is found in combination with only one value of the other nominal variable (the groups).

There doesn't seem to have been a lot of work done on non-parametric alternatives to nested anova. You could convert the measurement variable to ranks (replace each observation with its rank over the entire data set), then do a nested anova on the ranks; see Conover and Iman (1981).

How to do the test

Spreadsheet

I have made an spreadsheet to do a two-level nested anova, with equal or unequal sample sizes, on up to 50 subgroups with up to 1000 observations per subgroup. It does significance tests and partitions the variance. The spreadsheet tells you whether the Satterthwaite approximation is appropriate, using the rules on p. 298 of Sokal and Rohlf (1983), and gives you the option to use it. F_{group} is calculated as $MS_{group}/MS_{subgroup}$. The spreadsheet gives the variance components as percentages of the total. If the estimate of the group component would be negative (which can happen in unbalanced designs), it is set to zero.

Web page

Rweb (http://bayes.math.montana.edu/cgi-bin/Rweb/buildModules.cgi) lets you do nested anovas. To use it, choose "ANOVA" from the Analysis Menu and

choose "External Data: Use an option below" from the Data Set Menu, then either select a file to analyze or enter your data in the box. On the next page (after clicking on "Submit"), select the two nominal variables under "Choose the Factors" and select the measurement variable under "Choose the response." F_{group} is calculated as MS_{group}/MS_{within}, which is not a good idea if $F_{subgroup}$ is significant. Rweb does not partition the variance.

SAS

PROC GLM will handle both balanced and unbalanced designs. List all the nominal variables in the CLASS statement. In the MODEL statement, give the name of the measurement variable, then after the equals sign give the name of the group variable, then the name of the subgroup variable followed by the group variable in parentheses, etc. The TEST statement tells it to calculate the F-statistic for groups by dividing the group mean square by the subgroup mean square, instead of the within-group mean square ("h" stands for "hypothesis" and "e" stands for "error"). "htype=1 etype=1" tells SAS to use "type I sums of squares"; I couldn't tell you the difference between them and types II, III and IV, but I'm pretty sure that type I is appropriate for a nested anova.

Here is an example using data on the codon bias index (CBI), a measure of the nonrandom use of synonymous codons. The groups are two chromosomes in *Drosophila melanogaster*, and the subgroups are small regions within each chromosome. Each observation is the CBI value for a single gene in that chromosome region.

```
data flies;
   input gene $ chrom $ reg $ cbi;
   cards;
singed          X       7D      0.366
```

====See the web page for the full data set====

```
sepia           3      66D      0.245
;
proc glm data=flies;
  class chrom reg;
  model cbi=chrom reg(chrom) / ss1;
  test h=chrom e=reg(chrom) / htype=1 etype=1;
run;
```

The output includes F_{group} calculated two ways, as MS_{group}/MS_{within} and as $MS_{group}/MS_{subgroup}$.

```
Source      DF  Type I SS   Mean Sq.  F Value  Pr > F

chrom        2  0.0103259   0.005163   0.66   0.5255 MSgroup/MSwithin
reg(chrom)   3  0.1629461   0.054315   6.92   0.0011


Tests of Hypotheses Using Type I MS for reg(chrom) as an Error Term

Source      DF  Type I SS   Mean Sq.  F Value  Pr > F

chrom        2  0.0103259   0.005163   0.10   0.9120 MSgroup/MSsubgroup
```

To do the Tukey-Kramer test or Gabriel comparison intervals, add a MEANS statement, as shown here:

```
proc glm data=flies;
  class chrom reg;
  model cbi=chrom reg(chrom) / ss1;
  test h=chrom e=reg(chrom) / htype=1 etype=1;
  means chrom /lines tukey;
  means chrom /clm gabriel;
run;
```

Here is the output from these two MEANS statements when applied to the example data set:

```
        Tukey's Studentized Range (HSD) Test for cbi

  Means with the same letter are not significantly different.

 Tukey Grouping          Mean      N    chrom

              A       0.21975     12    3
              A
              A       0.19330     10    X
              A
              A       0.18014     14    2

                          .
                          .
                          .

        Gabriel's Comparison Intervals for cbi

                                        95% Comparison
 chrom              N          Mean           Limits

 3                 12       0.21975     0.17412     0.26538
 X                 10       0.19330     0.14331     0.24329
```

```
 2                14       0.18014     0.13790     0.22239
```

PROC GLM does not partition the variance. There is a PROC NESTED that will partition the variance, but it only does the hypothesis testing for a balanced nested anova, so if you have an unbalanced design you'll want to run both PROC GLM and PROC NESTED. PROC NESTED requires that the data be sorted by groups and subgroups. To use it on the above data, add the following to the end of the above SAS program. PROC SORT sorts the data by the first variable in the BY statement, then by the second variable. In PROC NESTED, the group is given first in the CLASS statement, then the subgroup.

```
proc sort data=flies;
   by chrom reg;
proc nested data=flies;
   class chrom reg;
   var cbi;
run;
```

As you can see, PROC NESTED didn't calculate the F-statistics or P-values, since the fly data are unbalanced.

```
Variance      Sum of    F         Error  Mean      Variance    Percent
Source    DF Squares Value Pr>F  Term Square    Component   of Total

Total     35 0.40891                  0.01168    0.015670  100.0000
chrom      2 0.01033                  0.00516   -0.004171    0.0000
reg        3 0.16295                  0.05431    0.007816   49.8765
Error     30 0.23564                  0.00785    0.007855   50.1235
```

Further reading

Sokal and Rohlf, pp. 272-308.

Zar, pp. 303-311.

References

Conover, W.J., and R.L. Iman. 1981. Rank transformations as a bridge between parametric and nonparametric statistics. Am. Statistician 35: 124-129.

Keon, D.B., and P.S. Muir. 2002. Growth of *Usnea longissima* across a variety of habitats in the Oregon coast range. Bryologist 105: 233-242.

Two-way anova

When to use it

You use a two-way anova (also known as a factorial anova, with two factors) when you have one measurement variable and two nominal variables. The nominal variables (often called "factors" or "main effects") are found in all possible combinations. For example, let's say you are testing the null hypothesis that stressed and unstressed rats have the same glycogen content in their gastrocnemius muscle, and you are worried that there might be sex-related differences in glycogen content as well. The two factors are stress level (stressed vs. unstressed) and sex (male vs. female). Unlike a nested anova, each grouping extends across the other grouping. In a nested anova, you might have "cage 1" and "cage 2" nested entirely within the stressed group, while "cage 3" and "cage 4" were nested within the unstressed group. In a two-way anova, the stressed group contains both male and female rats, and the unstressed group also contains both male and female rats. The factors used to group the observations may both be model I, may both be model II, or may be one of each ("mixed model").

A two-way anova may be done with replication (more than one observation for each combination of the nominal variables) or without replication (only one observation for each combination of the nominal variables).

Assumptions

Two-way anova, like all anovas, assumes that the observations within each cell are normally distributed and have equal variances.

Two-way anova with replication

Null hypotheses: The results of a two-way anova with replication include tests of three null hypotheses: that the means of observations grouped by one factor are the same; that the means of observations grouped by the other factor are the same; and that there is no interaction between the two factors. The interaction test tells you whether the effects of one factor depend on the other factor. In the rat example, imagine that stressed and unstressed female rats have about the same glycogen level, while stressed male rats had much lower glycogen levels than

unstressed male rats. The different effects of stress on female and male rats would result in a significant interaction term in the anova. When the interaction term is significant, the usual advice is that you should *not* test the effects of the individual factors. In this example, it would be misleading to examine the individual factors and conclude "Stressed rats have lower glycogen than unstressed," when that is only true for male rats, or "Male rats have lower glycogen than female rats," when that is only true when they are stressed.

What you can do, if the interaction term is significant, is look at each factor separately, using a one-way anova. In the rat example, you might be able to say that for female rats, the mean glycogen levels for stressed and unstressed rats are not significantly different, while for male rats, stresed rats have a significantly lower mean glycogen level than unstressed rats. Or, if you're more interested in the sex difference, you might say that male rats have a significantly lower mean glycogen level than female rats under stress conditions, while the mean glycogen levels do not differ significantly under unstressed conditions.

How the test works: When the sample sizes in each subgroup are equal (a "balanced design"), the mean square is calculated for each of the two factors (the "main effects"), for the interaction, and for the variation within each combination of factors. Each F-statistic is found by dividing a mean square by the within-subgroup mean square.

When the sample sizes for the subgroups are not equal (an "unbalanced design"), the analysis is much more complicated, and there are several different techniques for testing the main and interaction effects. The details of this are beyond the scope of this handbook. If you're doing a two-way anova, your statistical life will be a lot easier if you make it a balanced design.

Two-way anova without replication

Null hypotheses: When there is only a single observation for each combination of the nominal variables, there are only two null hypotheses: that the means of observations grouped by one factor are the same, and that the means of observations grouped by the other factor are the same. It is impossible to test the null hypothesis of no interaction. Testing the two null hypotheses about the main effects requires assuming that there is no interaction.

How the test works: The mean square is calculated for each of the two main effects, and a total mean square is also calculated by considering all of the observations as a single group. The remainder mean square (also called the discrepance or error mean square) is found by subtracting the two main effect mean squares from the total mean square. The F-statistic for a main effect is the main effect mean square divided by the remainder mean square.

Repeated measures: One experimental design that is analyzed by a two-way anova is repeated measures, where an observation has been made on the same individual more than once. This usually involves measurements taken at different

time points. For example, you might measure running speed before, one week into, and three weeks into a program of exercise. Because individuals would start with different running speeds, it is better to analyze using a two-way anova, with "individual" as one of the factors, rather than lumping everyone together and analyzing with a one-way anova. Sometimes the repeated measures are repeated at different places rather than different times, such as the hip abduction angle measured on the right and left hip of individuals. Repeated measures experiments are often done without replication, although they could be done with replication.

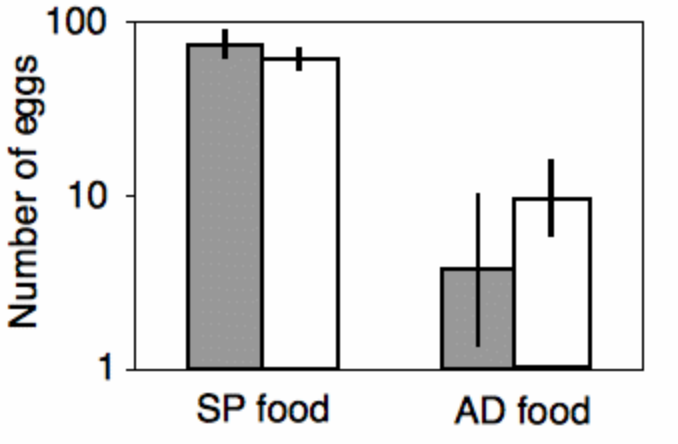
In a repeated measures design, one of main effects is usually uninteresting and the test of its null hypothesis may not be reported. If the goal is to determine whether a particular exercise program affects running speed, there would be little point in testing whether individuals differed from each other in their average running speed; only the change in running speed over time would be of interest.

Randomized blocks: Another experimental design that is analyzed by a two-way anova is randomized blocks. This often occurs in agriculture, where you may want to test different treatments on small plots within larger blocks of land. Because the larger blocks may differ in some way that may affect the measurement variable, the data are analyzed with a two-way anova, with the block as one of the nominal variables. Each treatment is applied to one or more plot within the larger block, and the positions of the treatments are assigned at random. This is most commonly done without replication (one plot per block), but it can be done with replication as well.

Examples

Shimoji and Miyatake (2002) raised the West Indian sweetpotato weevil for 14 generations on an artificial diet. They compared these artificial diet weevils (AD strain) with weevils raised on sweet potato roots (SP strain), the weevil's natural food. Multiple females of each strain were placed on either the artificial diet or sweet potato root, and the number of eggs each female laid over a 28-day period was counted. There are two nominal variables, the strain of weevil (AD or SP) and the oviposition test food (artificial diet or sweet potato), and one measurement variable (the number of eggs laid).

The results of the two-way anova with replication include a significant interaction term ($F_{1, 117}$=17.02, P=7 x 10^{-5}). Looking at the graph, the interaction can be interpreted this way: on the sweet potato diet, the SP strain laid more eggs than the AD strain; on the artificial diet, the AD strain laid more eggs than the SP strain. Each main effect is also significant: weevil strain ($F_{1, 117}$=8.82, P=0.0036) and oviposition test food ($F_{1, 117}$=345.92, P=9 x 10^{-37}). However, the significant effect of strain is a bit misleading, as the direction of the difference between strains depends on which food they ate. This is why it is important to look at the interaction term first.

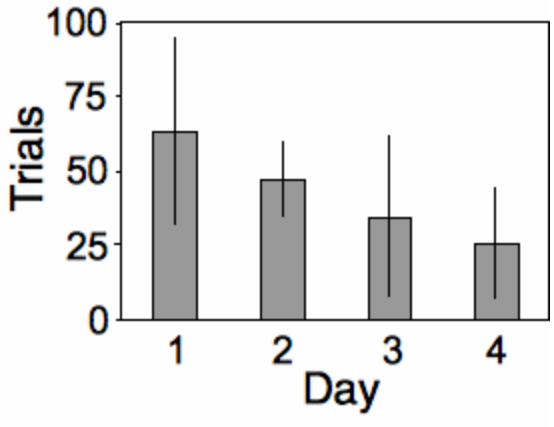


Mean total numbers of eggs of females from the SP strain (gray bars) and AD strain (white bars). Values are mean ±SEM. (Adapted from Fig. 4 of Shimoji and Miyatake [2002]).

I assayed the activity of the enzyme mannose-6-phosphate isomerase (MPI) in the amphipod crustacean *Platorchestia platensis* (McDonald, unpublished data). There are three genotypes at the locus for MPI, Mpi^{ff}, Mpi^{fs}, and Mpi^{ss}, and I wanted to know whether the genotypes had different activity. Because I didn't know whether sex would affect activity, I also recorded the sex. Each amphipod was lyophilized, weighed, and homogenized; then MPI activity of the soluble portion was assayed. The data (in ΔO.D. units/sec/mg dry weight) are shown below as part of the SAS example. The results indicate that the interaction term, the effect of sex and the effect of genotype are all non-significant.

Place and Abramson (2008) put diamondback rattlesnakes (*Crotalus atrox*) in a "rattlebox," a box with a lid that would slide open and shut every 5 minutes. At first, the snake would rattle its tail each time the box opened. After a while, the snake would become habituated to the box opening and stop rattling its tail. They counted the number of box openings until a snake stopped rattling; fewer box openings means the snake was more quickly habituated. They repeated this experiment on each snake on four successive days. Place and Abramson (2008) used 10 snakes, but some of them never became habituated; to simplify this example, I'll use data from the 6 snakes that did become habituated on each day:

Snake ID	Day	Trials to habituation
D1	1	85
	2	58
	3	15
	4	57
D3	1	107
	2	51
	3	30
	4	12
D5	1	61
	2	60
	3	68
	4	36
D8	1	22
	2	41
	3	63
	4	21
D11	1	40
	2	45
	3	28
	4	10
D12	1	65
	2	27
	3	3
	4	16

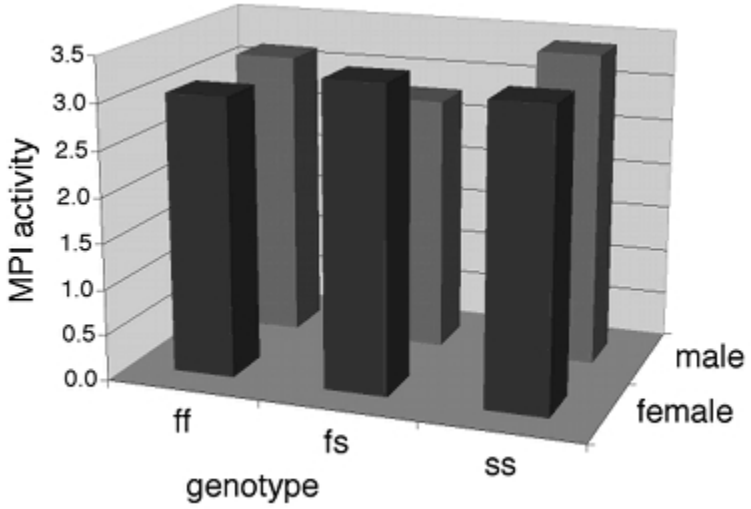


Mean number of trials before rattlesnakes stopped rattling, on four successive days. Values are mean ±95% confidence intervals. Data from Place and Abramson (2008).

The measurement variable is trials to habituation, and the two nominal variables are day (1 to 4) and snake ID. This is a repeated measures design, as the measurement variable is measured repeatedly on each snake. It is analyzed using a two-way anova without replication. The effect of snake is not significant ($F_{5, 15}$=1.24, P=0.34), while the effect of day is significant ($F_{3, 15}$=3.32, P=0.049).

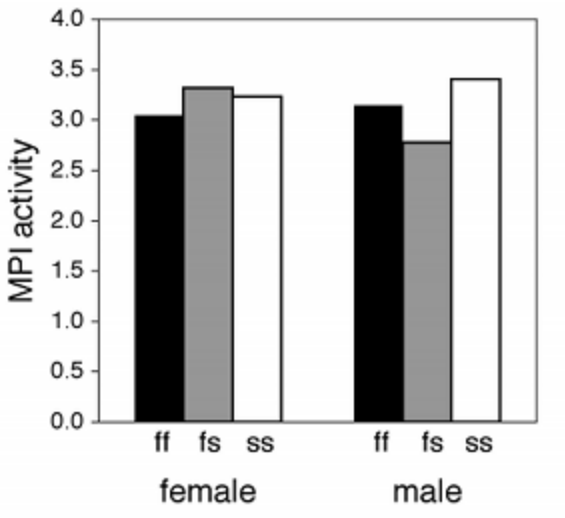
Graphing the results

Sometimes the results of a two-way anova are plotted on a 3-D graph, with the measurement variable on the Y-axis, one nominal variable on the X-axis, and the other nominal variable on the Z-axis (going into the paper). This makes it difficult to visually compare the heights of the bars in the front and back rows, so I don't recommend this. Instead, I suggest you plot a bar graph with the bars clustered by one nominal variable, with the other nominal variable identified using the color or pattern of the bars.



Don't use this kind of graph. Which bar is higher: *fs* in females or *ss* in males?

If one of the nominal variables is the interesting one, and the other is just a possible confounder, I'd group the bars by the possible confounder and use different patterns for the interesting variable. For the amphipod data described above, I was interested in seeing whether MPI phenotype affected enzyme activity, with any difference between males and females as an annoying confounder, so I group the bars by sex.



Mannose-6-phosphate isomerase activity in three MPI genotypes in the amphipod crustacean *Platorchestia platensis*. Isn't this graph much better?

Similar tests

A two-way anova without replication and only two values for the interesting nominal variable may be analyzed using a paired t-test. The results of a paired t-test are mathematically identical to those of a two-way anova, but the paired t-test is easier to do. Data sets with one measurement variable and two nominal variables, with one nominal variable nested under the other, are analyzed with a nested anova.

Data in which the measurement variable is severely non-normal or heteroscedastic may be analyzed using the non-parametric Friedman's method (http://www.fon.hum.uva.nl/Service/Statistics/Friedman.html) (for a two-way design without replication) or the Scheirer–Ray–Hare technique (for a two-way design with replication). See Sokal and Rohlf (1995), pp. 440-447. I don't know how to tell whether the non-normality or heteroscedasticity in your data are so bad that a two-way anova would be inappropriate.

Three-way and higher order anovas are possible, as are anovas combining aspects of a nested and a two-way or higher order anova. The number of interaction terms increases rapidly as designs get more complicated, and the interpretation of any significant interactions can be quite difficult. It is better, when possible, to design your experiments so that as many factors as possible are controlled, rather than collecting a hodgepodge of data and hoping that a sophisticated statistical analysis can make some sense of it.

How to do the test

Spreadsheet

I haven't put together a spreadsheet to do two-way anovas.

Web pages

Web pages are available to perform: a 2x2 two-way anova with replication, (http://faculty.vassar.edu/lowry/vsanova.html) up to a 6x4 two-way anova without replication (http://home.ubalt.edu/ntsbarsh/Business-stat/otherapplets/ANOVATwo.htm) or with up to 4 replicates. (http://home.ubalt.edu/ntsbarsh/Business-stat/otherapplets/ANOVA2Rep.htm)

Rweb (http://bayes.math.montana.edu/cgi-bin/Rweb/buildModules.cgi) lets you do two-way anovas with or without replication. To use it, choose "ANOVA" from the Analysis Menu and choose "External Data: Use an option below" from the Data Set Menu, then either select a file to analyze or enter your data in the box. On the next page (after clicking on "Submit"), select the two nominal variables under "Choose the Factors" and select the measurement variable under "Choose the response."

SAS

Use PROC GLM for a two-way anova. Here is an example using the MPI activity data described above:

```
data amphipods;
   input ID $ sex $ genotype $ activity;
   cards;
1       male     ff   1.884
```

====See the web page for the full data set====

```
49      male     ss   3.110
;
proc glm data=amphipods;
  class sex genotype;
  model activity=sex genotype sex*genotype;
run;
```

The results indicate that the interaction term is not significant (P=0.60), the effect of genotype is not significant (P=0.84), and the effect of sex concentration not significant (P=0.77).

Source	DF	Type I SS	Mean Square	F Value	Pr > F
sex	1	0.06808050	0.06808050	0.09	0.7712
genotype	2	0.27724017	0.13862008	0.18	0.8400
sex*genotype	2	0.81464133	0.40732067	0.52	0.6025

If you are using SAS to do a two-way anova without replication, do not put an interaction term in the model statement (sex*genotype is the interaction term in the example above).

Further reading

Sokal and Rohlf, pp. 321-342.

Zar, pp. 231-271.

References

Place, A.J., and C.I. Abramson. 2008. Habituation of the rattle response in western diamondback rattlesnakes, *Crotalus atrox*. Copeia 2008: 835-843.

Shimoji, Y., and T. Miyatake. 2002. Adaptation to artificial rearing during successive generations in the West Indian sweetpotato weevil, *Euscepes postfasciatus* (Coleoptera: Curuculionidae). Annals of the Entomological Society of America 95: 735-739.

Paired t-test

When to use it

You use the paired t-test when there is one measurement variable and two nominal variables. One of the nominal variables has only two values. The most common design is that one nominal variable represents different individuals, while the other is "before" and "after" some treatment. Sometimes the pairs are spatial rather than temporal, such as left vs. right, injured limb vs. uninjured limb, above a dam vs. below a dam, etc.

An example would be the performance of undergraduates on a test of manual dexterity before and after drinking a cup of tea. For each student, there would be two observations, one before the tea and one after. I would expect the students to vary widely in their performance, so if the tea decreased their mean performance by 5 percent, it would take a very large sample size to detect this difference if the data were analyzed using a Student's t-test. Using a paired t-test has much more statistical power when the difference between groups is small relative to the variation within groups.

The paired t-test is only appropriate when there is just one observation for each combination of the nominal values. For the tea example, that would be one measurement of dexterity on each student before drinking tea, and one measurement after drinking tea. If you had multiple measurements of dexterity on each student before and after drinking tea, you would do a two-way anova with replication.

Null hypothesis

The null hypothesis is that the mean difference between paired observations is zero. This is mathematically equivalent to the null hypothesis of a one-way anova or t-test, that the means of the groups are equal, but because of the paired design of the data, the null hypothesis of a paired t-test is usually expressed in terms of the mean difference.

Assumption

The paired t-test assumes that the differences between pairs are normally distributed; you can use the histogram spreadsheet on that page to check the normality. If the differences between pairs are severely non-normal, it would be better to use the Wilcoxon signed-rank test. I don't think the test is very sensitive to deviations from normality, so unless the deviation from normality is really obvious, I wouldn't worry about it.

How the test works

The difference between the observations is calculated for each pair, and the mean and standard error of these differences are calculated. Dividing the mean by the standard error of the mean yields a test statistic, t_s, that is t-distributed with degrees of freedom equal to one less than the number of pairs.

Examples

Wiebe and Bortolotti (2002) examined color in the tail feathers of northern flickers. Some of the birds had one "odd" feather that was different in color or length from the rest of the tail feathers, presumably because it was regrown after being lost. They measured the yellowness of one odd feather on each of 16 birds and compared it with the yellowness of one typical feather from the same bird. There are two nominal variables, type of feather (typical or odd) and the individual bird, and one measurement variable, yellowness. Because these birds were from a hybrid zone between red-shafted flickers and yellow-shafted flickers, there was a lot of variation among birds in color, making a paired analysis more appropriate. The difference was significant (P=0.001), with the odd feathers significantly less yellow than the typical feathers (higher numbers are more yellow).

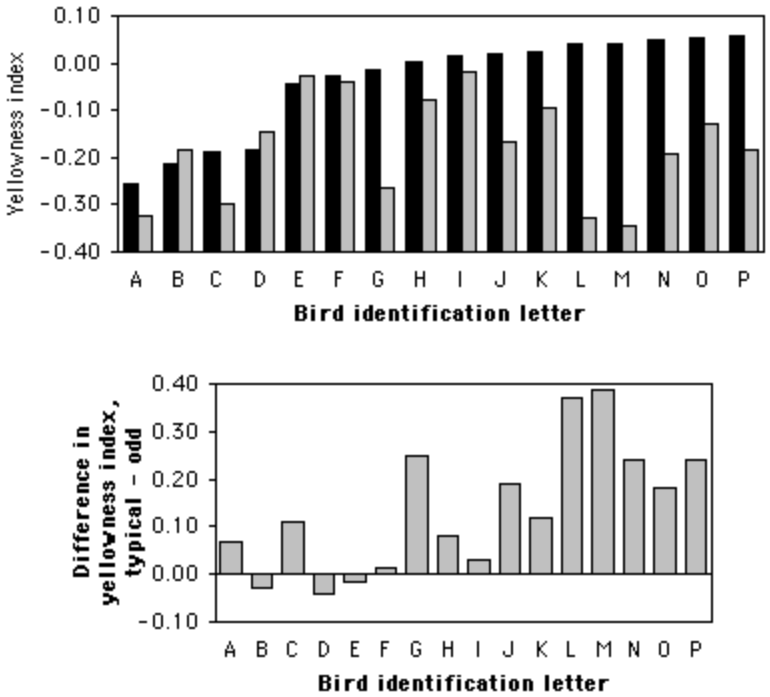
	Yellowness index	
Bird	Typical feather	Odd feather
A	-0.255	-0.324
B	-0.213	-0.185
C	-0.190	-0.299
D	-0.185	-0.144
E	-0.045	-0.027
F	-0.025	-0.039
G	-0.015	-0.264
H	0.003	-0.077
I	0.015	-0.017
J	0.020	-0.169
K	0.023	-0.096
L	0.040	-0.330

M	0.040	-0.346
N	0.050	-0.191
O	0.055	-0.128
P	0.058	-0.182

Wilder and Rypstra (2004) tested the effect of praying mantis excrement on the behavior of wolf spiders. They put 12 wolf spiders in individual containers; each container had two semicircles of filter paper, one semicircle that had been smeared with praying mantis excrement and one without excrement. They observed each spider for one hour, and measured its walking speed while it was on each half of the container. There are two nominal variables, filter paper type (with or without excrement) and the individual spider, and one measurement variable (walking speed). Different spiders may have different overall walking speed, so a paired analysis is appropriate to test whether the presence of praying mantis excrement changes the walking speed of a spider. The mean change in walking speed is almost, but not quite, significantly different from 0 (t=2.11, 11 d.f., P=0.053).

Graphing the results

If there are a moderate number of pairs, you could either plot each individual value on a bar graph, or plot the differences. Here is one graph in each format for the flicker data:



Colors of tail feathers in the northern flicker. The graph on the top shows the yellowness index for a "typical" feather with a black bar and an "odd" feather with a gray bar. The graph on the bottom shows the difference (typical − odd).

Related tests

The paired t-test is mathematically equivalent to one of the hypothesis tests of a two-way anova without replication. The paired t-test is simpler to perform and may be more familiar. A two-way anova would be better if both null hypotheses (equality of means of the two treatments and equality of means of the individuals) were of interest; in a paired t-test, the means of individuals are so likely to be different that there's no point in testing them. A two-way anova would have to be used if the measurements are replicated for the treatment/individual combinations.

If the paired nature of the data is ignored, the data would be analyzed using a one-way anova or a regular t-test. The loss in statistical power can be quite dramatic, however, so this is not a good idea.

One non-parametric analogue of the paired t-test is Wilcoxon signed-rank test. A simpler and even less powerful test is the sign test, which considers only the direction of difference between pairs of observations, not the size of the difference.

How to do the test

Spreadsheet

Spreadsheets have a built-in function to perform paired t-tests. Put the "before" numbers in one column, and the "after" numbers in the adjacent column, with the before and after observations from each individual on the same row. Then enter =TTEST(*array1, array2, tails, type*), where *array1* is the first column of data, *array2* is the second column of data, *tails* is normally set to 2 for a two-tailed test, and *type* is set to 1 for a paired t-test. The result of this function is the P-value of the paired t-test.

Web pages

There are web pages to do paired t-tests here (http://www.fon.hum.uva.nl/Service/Statistics/Student_t_Test.html) , here (http://faculty.vassar.edu/lowry/t_corr_stats.html) , here (http://www.physics.csbsju.edu/stats/Paired_t-test_NROW_form.html) , and here (http://graphpad.com/quickcalcs/ttest1.cfm) .

SAS

To do a paired t-test in SAS, you use PROC TTEST with the PAIRED option. Here is an example using the feather data from above:

```
data feathers;
   input bird typical odd;
   cards;
 A     -0.255           -0.324
 B     -0.213           -0.185
 C     -0.190           -0.299
 D     -0.185           -0.144
 E     -0.045           -0.027
 F     -0.025           -0.039
 G     -0.015           -0.264
 H      0.003           -0.077
 I      0.015           -0.017
 J      0.020           -0.169
 K      0.023           -0.096
 L      0.040           -0.330
 M      0.040           -0.346
 N      0.050           -0.191
 O      0.055           -0.128
 P      0.058           -0.182
;
proc ttest data=feathers;
   paired typical*odd;
run;
```

The results include the following, which shows that the P-value is 0.0010:

```
                    T-Tests
Difference            DF    t Value    Pr > |t|
typical - odd         15       4.06      0.0010
```

Power analysis

To estimate the sample sizes needed to detect a mean difference that is significantly different from zero, you need the following:

- the effect size, or the mean difference. In the feather data used above, the mean difference between typical and odd feathers is 0.176 yellowness units.
- the standard deviation of differences. Note that this is *not* the standard deviation within each group. For example, in the feather data, the standard deviation of the differences is 0.111; this is not the standard deviation among typical feathers, or the standard deviation among odd feathers, but the standard deviation of the differences;
- alpha, or the significance level (usually 0.05);
- beta, the probability of accepting the null hypothesis when it is false (0.80 and 0.90 are common values).

As an example, let's say you want to do a study comparing the redness of typical and odd tail feathers in cardinals. The closest you can find to preliminary data is the Weibe and Bortolotti (2002) paper on yellowness in flickers. They found a mean difference of 0.176 yellowness units, with a standard deviation of 0.111; you arbitrarily decide you want to be able to detect a mean difference of 0.10 redness units in your cardinals. On the form shown on the web page, you enter 0.10 for "Mean difference", 0.111 for "Standard deviation of differences", 0.05 for the alpha, and 0.80 for the power. The result is 12, so you'll need a minimum of 12 birds (with two observations per bird).

You can also do a power analysis for a paired t-test using the free program G*Power (http://www.psycho.uni-duesseldorf.de/abteilungen/aap/gpower3/) . Choose "t tests" from the "Test family" menu and "Means: difference between dependent means (matched pairs)" from the "Statistical test" menu. To determine the effect size, click on the Determine button and enter the mean difference and the standard deviation of the difference. Then click on the "Calculate and transfer to main window" button; it calculates the effect size and enters it into the main window. Enter your alpha (usually 0.05) and power (typically 0.80 or 0.90) and hit the Calculate button. The result is the number of pairs of observations.

Further reading

Sokal and Rohlf, pp. 698-703, 729-730.

Zar, pp. 161-164.

References

Wiebe, K.L., and G.R. Bortolotti. 2002. Variation in carotenoid-based color in northern flickers in a hybrid zone. Wilson Bull. 114: 393-400.

Wilder, S.M., and A.L. Rypstra. 2004. Chemical cues from an introduced predator (Mantodea, Mantidae) reduce the movement and foraging of a native wolf spider (Araneae, Lycosidae) in the laboratory. Environ. Entom. 33: 1032-1036.

Wilcoxon signed-rank test

When to use it

You use the Wilcoxon signed-rank test when there are two nominal variables and one measurement variable. One of the nominal variables has only two values, such as "before" and "after," and the other nominal variable often represents individuals. This is the non-parametric analogue to the paired t-test, and should be used if the distribution of differences between pairs may be non-normally distributed.

Null hypothesis

The null hypothesis is that the median difference between pairs of observations is zero. Note that this is different from the null hypothesis of the paired t-test, which is that the *mean* difference between pairs is zero, or the null hypothesis of the sign test, which is that the numbers of differences in each direction are equal.

How it works

The absolute value of the differences between observations are ranked from smallest to largest, with the smallest difference getting a rank of 1, then next larger difference getting a rank of 2, etc. Ties are given average ranks. The ranks of all differences in one direction are summed, and the ranks of all differences in the other direction are summed. The smaller of these two sums is the test statistic, W (sometimes symbolized T_s). Unlike most test statistics, *smaller* values of W are less likely under the null hypothesis.

Examples

Laureysens et al. (2004) measured metal content in the wood of 13 poplar clones growing in a polluted area, once in August and once in November. Concentrations of aluminum (in micrograms of Al per gram of wood) are shown below.

Clone	Aug	Nov
Balsam Spire	8.1	11.2
Beaupre	10.0	16.3
Hazendans	16.5	15.3
Hoogvorst	13.6	15.6
Raspalje	9.5	10.5
Unal	8.3	15.5
Columbia River	18.3	12.7
Fritzi Pauley	13.3	11.1
Trichobel	7.9	19.9
Gaver	8.1	20.4
Gibecq	8.9	14.2
Primo	12.6	12.7
Wolterson	13.4	36.8

There are two nominal variables: time of year (August or November) and poplar clone (Balsam Spire, Beaupre, etc.), and one measurement variable (micrograms of aluminum per gram of wood). There are not enough observations to confidently test whether the differences between August and November are normally distributed, but they look like they might be a bit skewed; the Wolterson clone, in particular, has a much larger difference than any other clone. To be safe, the authors analyzed the data using a signed-rank test. The median change from August to November (3.1 micrograms Al/g wood) is significantly different from zero (W=16, P=0.040).

Buchwalder and Huber-Eicher (2004) wanted to know whether turkeys would be less aggressive towards unfamiliar individuals if they were housed in larger pens. They tested 10 groups of three turkeys that had been reared together, introducing an unfamiliar turkey and then counting the number of times it was pecked during the test period. Each group of turkeys was tested in a small pen and in a large pen. There are two nominal variables, size of pen (small or large) and the group of turkeys, and one measurement variable (number of pecks per test). The median difference between the number of pecks per test in the small pen vs. the large pen was significantly greater than zero (W=10, P=0.04).

Ho et al. (2004) inserted a plastic implant into the soft palate of 12 chronic snorers to see if it would reduce the volume of snoring. Snoring loudness was judged by the sleeping partner of the snorer on a subjective 10-point scale. There are two nominal variables, time (before the operations or after the operation) and individual snorer, and one measurement variable (loudness of snoring). One person left the study, and the implant fell out of the palate in two people; in the remaining nine people, the median change in snoring volume was significantly different from zero (W=0, P=0.008).

Graphing the results

You should graph the data for a signed rank test the same way you would graph the data for a paired t-test, a bar graph with either the values side-by-side for each pair, or the differences at each pair.

Similar tests

Paired observations of a measurement variable may be analyzed using a paired t-test, if the null hypothesis is that the mean difference between pairs of observations is zero and the differences are normally distributed. If you have a large number of paired observations, you can plot a histogram of the differences to see if they look normally distributed. I do not know how severe the deviation from normality has to be to make the paired t-test inappropriate.

The sign test is used when the null hypothesis is that there are equal number of differences in each direction.

How to do the test

Spreadsheet

I have prepared a spreadsheet to do the Wilcoxon signed-rank test. It will handle up to 1000 pairs of observations.

Web page

There is a web page (http://www.fon.hum.uva.nl/Service/Statistics/Signed_Rank_Test.html) that will perform the Wilcoxon signed-rank test. You may enter your paired numbers directly onto the web page; it will be easier if you enter them into a spreadsheet first, then copy them and paste them into the web page.

SAS

To do Wilcoxon signed-rank test in SAS, you first create a new variable that is the difference between the two observations. You then run PROC UNIVARIATE on the difference, which automatically does the Wilcoxon signed-rank test along with several others. Here's an example using the poplar data from above:

```
data poplars;
   input clone $ aug_al nov_al;
   diff=aug_al - nov_al;
   cards;
Balsam_Spire     8.1  11.2
Beaupre         10.0  16.3
Hazendans       16.5  15.3
Hoogvorst       13.6  15.6
Raspalje         9.5  10.5
Unal             8.3  15.5
Columbia_River  18.3  12.7
```

```
Fritzi_Pauley   13.3  11.1
Trichobel        7.9  19.9
Gaver            8.1  20.4
Gibecq           8.9  14.2
Primo           12.6  12.7
Wolterson       13.4  36.8
;
proc univariate data=poplars;
   var diff;
run;
```

PROC UNIVARIATE returns a bunch of descriptive statistics that you don't need; the result of the Wilcoxon signed-rank test is shown in the row labelled "Signed rank":

```
            Tests for Location: Mu0=0

Test           -Statistic-    -----p Value------

Student's t    t   -2.3089    Pr > |t|    0.0396
Sign           M      -3.5    Pr >= |M|   0.0923
Signed Rank    S     -29.5    Pr >= |S|   0.0398
```

Further reading

Sokal and Rohlf, pp. 440-444.

Zar, pp. 165-169.

References

Buchwalder, T., and B. Huber-Eicher. 2004. Effect of increased floor space on aggressive behaviour in male turkeys (*Melagris gallopavo*). Appl. Anim. Behav. Sci. 89: 207-214.

Ho, W.K., W.I. Wei, and K.F. Chung. 2004. Managing disturbing snoring with palatal implants: a pilot study. Arch. Otolaryngology Head and Neck Surg. 130: 753-758.

Laureysens, I., R. Blust, L. De Temmerman, C. Lemmens and R. Ceulemans. 2004. Clonal variation in heavy metal accumulation and biomass production in a poplar coppice culture. I. Seasonal variation in leaf, wood and bark concentrations. Environ. Pollution 131: 485-494.

Sign test

When to use it

You use the sign test when there are two nominal variables and one measurement variable or ranked variable. One of the nominal variables has only two values, such as "before" and "after" or "left" and "right," and the other nominal variable identifies the pairs of observations. The data could be analyzed using a paired t-test or a Wilcoxon signed-rank test, if the null hypothesis is that the mean or median difference between pairs of observations is zero. The sign test is used to test the null hypothesis that there are equal numbers of differences in each direction.

One situation in which a sign test is appropriate is when the biological null hypothesis is that there may be large differences between pairs of observations, but they are random in direction. For example, let's say you want to know whether copper pennies will reduce the number of mosquito larvae in backyard ponds. You measure the abundance of larvae in your pond, add some pennies, then measure the abundance of larvae again a month later. You do this for several other backyard ponds, with the before- and after-pennies measurements at different times in the summer for each pond. Based on prior research, you know that mosquito larvae abundance varies a lot throughout the summer, due to variation in the weather and random fluctuations in the number of adult mosquitoes that happen to find a particular pond; even if the pennies have no effect, you expect big differences in the abundance of larvae between the before and after samples. The random fluctuations in abundance would be random in direction, however, so if the pennies have no effect, you'd expect half the ponds to have more larvae before adding the pennies, and half the ponds to have more larvae after adding pennies.

To see why a paired t-test would be inappropriate for the mosquito experiment, imagine that you've done the experiment in a neighborhood with 100 backyard ponds. Due to changes in the weather, etc., the abundance of mosquito larvae increases in half the ponds and decreases in half the ponds; in other words, the probability that a random pond will decrease in mosquito larvae abundance is 0.5. If you do the experiment on four ponds picked at random, and all four happen show the same direction of difference (all four increase or all four decrease) even though the pennies really have no effect, you'll probably get a significant paired t-test. However, the probability that all four ponds will show the same direction of

change is 2×0.5^4, or 0.125. Thus you'd get a "significant" P-value from the paired t-test 12.5% of the time, which is much higher than the $P<0.05$ you want.

The other time you'd use a sign test is when you don't know the size of the difference, only its direction; in other words, you have a ranked variable with only two values, "greater" and "smaller." For example, let's say you're comparing the abundance of adult mosquitoes between your front yard and your back yard. You stand in your front yard for five minutes, swatting at every mosquito that lands on you, and then you stand in your back yard for five minutes. You intend to count every mosquito that lands on you, but they are so abundant that soon you're dancing around, swatting yourself wildly, with no hope of getting an accurate count. You then repeat this in your back yard and rate the mosquito abundance in your back yard as either "more annoying" or "less annoying" than your front yard. You repeat this on several subsequent days. You don't have any numbers for mosquito abundance, but you can do a sign test and see whether there are significantly more times where your front yard has more mosquitoes than your back yard, or vice versa.

Null hypothesis

The null hypothesis is that an equal number of pairs of observations have a change in each direction. If the pairs are "before" and "after," the null hypothesis would be that the number of pairs showing an increase equals the number showing a decrease.

Note that this is different from the null hypothesis tested by a paired t-test, which is that the mean difference between pairs is zero. The difference would be illustrated by a data set in which 19 pairs had an increase of 1 unit, while one pair had a decrease of 19 units. The 19: 1 ratio of increases to decreases would be highly significant under a sign test, but the mean change would be zero.

Examples

Farrell et al. (2001) estimated the evolutionary tree of two subfamilies of beetles that burrow inside trees as adults. They found ten pairs of sister groups in which one group of related species, or "clade," fed on angiosperms and one fed on gymnosperms, and they counted the number of species in each clade. There are two nominal variables, food source (angiosperms or gymnosperms) and pair of clades (Corthylina vs. Pityophthorus, etc.) and one measurement variable, the number of species per clade.

The biological null hypothesis is that although the number of species per clade may vary widely due to a variety of unknown factors, whether a clade feeds on angiosperms or gymnosperms will not be one of these factors. In other words, you expect that each pair of related clades will differ in number of species, but half the

time the angiosperm-feeding clade will have more species, and half the time the gymnosperm-feeding clade will have more species.

Applying a sign test, there are 10 pairs of clades in which the angiosperm-specialized clade has more species, and 0 pairs with more species in the gymnosperm-specialized clade; this is significantly different from the null expectation (P=0.002), and you can reject the null hypothesis and conclude that in these beetles, clades that feed on angiosperms tend to have more species than clades that feed on gymnosperms.

Angiosperm-feeding	Spp.	Gymonsperm-feeding	Spp.
Corthylina	458	Pityophthorus	200
Scolytinae	5200	Hylastini + Tomacini	180
Acanthotomicus + Premnobious	123	Orhotomicus	11
Xyleborini/Dryocoetini	1500	Ipini	195
Apion	1500	Antliarhininae	12
Belinae	150	Allocoryninae + Oxycorinae	30
Higher Curculionidae	44002	Nemonychidae	85
Higher Cerambycidae	25000	Aseminae + Spondylinae	78
Megalopodinae	400	Palophaginae	3
Higher Chrysomelidae	33400	Aulocoscelinae + Orsodacninae	26

Sherwin (2004) wanted to know whether laboratory mice preferred having a mirror in their cage. He set up 16 pairs of connected cages, one with a mirror and one without, and put a solitary mouse in each pair of cages. He then measured the amount of time each mouse spent in each of its two cages. There are two nominal variables, mirror (present or absent) and the individual mouse, and one measurement variable, the time spent in each cage. Three of the 16 mice spent more time in the cage with a mirror, and 13 mice spent more time in the cage without a mirror. The result or a sign test is P=0.021, so you can reject the null hypothesis that the number of mice that prefer a mirror equals the number of mice that prefer not having a mirror.

McDonald (1991) counted allele frequencies at the mannose-6-phosphate (MPI) locus in the amphipod crustacean *Orchestia grillus* from six bays on the north shore of Long Island, New York. At each bay two sites were sampled, one outside the bay ("exposed") and one inside the bay ("protected"). There are three nominal variables: allele ("fast" or "slow"), habitat ("exposed" or "protected"), and bay. The allele frequencies at each bay were converted to a ranked variable with two values: Mpi^{fast} more common at the exposed site than the protected site, or Mpi^{fast} less common at the exposed site. At all six bays, Mpi^{fast} was less common at the exposed site, which is significant by a sign test (P=0.03).

Note that this experimental design is identical to the study of Lap allele frequencies in the mussel *Mytilus trossulus* inside and outside of Oregon estuaries that was used as an example for the Cochran–Mantel–Haenszel test. Although the experimental designs are the same, the biological questions are different, which

makes the Cochran–Mantel–Haenszel test appropriate for the mussels and the sign test appropriate for the amphipods.

Two evolutionary processes can cause allele frequencies to be different between different locations, natural selection or random genetic drift. Mussels have larvae that float around in the water for a few weeks before settling onto rocks, so I considered it very unlikely that random genetic drift would cause a difference in allele frequencies between locations just a few kilometers apart. Therefore the biological null hypothesis is that the absence of natural selection keeps the allele frequencies the same inside and outside of estuaries; any difference in allele frequency between marine and estuarine habitats would be evidence for natural selection. The Cochran–Mantel–Haenszel test is a test of the statistical null hypothesis that the allele frequencies are the same in the two habitats, so the significant result is evidence that Lap in the mussels is affected by natural selection.

The amphipod *Orchestia grillus* does not have larvae that float in the water; the female amphipods carry the young in a brood pouch until they're ready to hop around on their own. The amphipods live near the high tide line in marshes, so the exposed and protected sites are likely to be well isolated populations with little migration between them. Therefore differences between exposed and protected sites due to random genetic drift are quite likely, and it wouldn't have been very interesting to find them. Random genetic drift, however, is random in direction; if it were the only process affecting allele frequencies, the Mpi*fast* allele would be more common inside half the bays and less common inside half the bays. The significant sign test indicates that the direction of difference in allele frequency is not random, so the biological null hypothesis of differences due to random drift can be rejected and the alternative hypothesis of differences due to natural selection can be accepted.

Graphing the results

You should graph the data for a sign test the same way you would graph the data for a paired t-test, a bar graph with either the values side-by-side for each pair, or the differences at each pair.

Similar tests

Paired observations of a measurement variable may be analyzed using a paired t-test, if the null hypothesis is that the mean difference between pairs of observations is zero, or a Wilcoxon signed-rank test, if the null hypothesis is that the median difference between pairs of observations is zero. The sign test is used when the null hypothesis is that there are equal number of differences in each direction.

How to do the test

Spreadsheet

First, count the number of pairs of observations with an increase (plus signs) and the number of pairs with a decrease (minus signs). Ignore pairs with no difference. Compare the ratio of plus signs: minus signs to the expected 1:1 ratio using the exact binomial test spreadsheet.

Web page

You can use Richard Lowry's exact binomial test web page (http://faculty.vassar.edu/lowry/binomialX.html) to do a sign test, once you've counted the number of differences in each direction by hand.

SAS

PROC UNIVARIATE automatically does a sign test; see the example on the Wilcoxon signed-rank web page.

Power analysis

Because the sign test is just an application of the exact binomial test, you can use the sample size calculator for the exact binomial test.

Further reading

Sokal and Rohlf, pp. 444-445.

Zar, pp. 538-539.

References

Farrell, B.D., A.S. Sequeira, B.C. O'Meara, B.B. Normark, J.H. Chung, and B.H. Jordal. 2001. The evolution of agriculture in beetles (Curculionidae: Scolytinae and Platypodinae). Evolution 55: 2011-2027.

McDonald, J.H. 1991. Contrasting amounts of geographic variation as evidence for direct selection: the *Mpi* and *Pgm* loci in eight crustacean species. Heredity 67:215-219.

Sherwin, C.M. 2004. Mirrors as potential environmental enrichment for individually housed laboratory mice. Appl. Anim. Behav. Sci. 87: 95-103.

Correlation and linear regression

Introduction

I find the descriptions of correlation and regression in most textbooks to be unnecessarily confusing. Some statistics textbooks have correlation and linear regression in separate chapters, and make it seem as if it is important to pick one technique or the other, based on subtle differences in the design and assumptions of the experiment. I think this overemphasizes the differences between them. Other books muddle correlation and regression together, leading the reader puzzled about what the difference is.

My understanding of the two techniques, as they are practiced, is that they primarily differ in goals. The goal of a correlation analysis is to see whether two measurement variables covary, and to measure the strength of any relationship between the variables. The results of correlation are expressed as a P-value (for the hypothesis test) and an r-value (correlation coefficient) or r^2 value (coefficient of determination). The goal of linear regression is to find the equation (slope and intercept) of the line that best fits the points; this line is then used as a visual summary of the relationship between the variables, or for estimating unknown values of one variable when given the value of the other.

When you have two measurement variables in biology, you'll usually want to do *both* correlation and regression—you'll want the P-value of the hypothesis test, *and* the r^2 that describes the strength of the relationship, *and* the regression line that illustrates the relationship. It would be less confusing if there were a single name for the whole process, just like "anova" includes testing hypotheses, partitioning variance, and estimating means. Since there isn't a single name, one option is to refer to the P-value and r^2 as resulting from a correlation analysis, while the equation of the line results from linear regression: "The correlation of variables X and Y is significant (r^2=0.89, P=0.007); the linear regression line is shown in the figure." It is also common to say something like "The linear regression of Y on X is significant (r^2=0.89, P=0.007)"; either seems appropriate. The one thing you should not do is call a linear regression line a "correlation line"; if that means anything, it means something different from a regression line.

Here I'll treat correlation and linear regression as different aspects of a single analysis. Be aware that this approach will probably be different from what you'll see elsewhere.

When to use them

Correlation and linear regression are used when you have two measurement variables, such as food intake and weight, drug dosage and blood pressure, air temperature and metabolic rate, etc.

There's also one nominal variable that keeps the two measurements together in pairs, such as the name of an individual organism. I'm not aware that anyone else considers this nominal variable to be part of correlation and regression, and it's not something you need to know the value of—you could indicate that a food intake measurement and weight measurement came from the same rat by putting both numbers on the same line, without ever giving the rat a name. For that reason, I'll call it a "hidden" nominal variable.

The data are typically plotted as a scatter of points on a graph, with one variable on the X axis and the other variable on the Y axis. The goals are to find the equation for the line that best fits these points, and to determine whether the slope of this line is significantly different from zero. If the slope is significantly different from zero, there is a significant relationship between the two variables: as the values of one variable increase, the values of the other variable either tend to increase (if the slope is positive) or tend to decrease (if the slope is negative).

There are three main uses for correlation and regression in biology. One is to test hypotheses about cause-and-effect relationships. In this case, the experimenter determines the values of the X-variable and sees whether variation in X causes variation in Y. An example would be giving people different amounts of a drug and measuring their blood pressure. The null hypothesis would be that there was no relationship between the amount of drug and the blood pressure. If the null hypothesis is rejected, the conclusion would be that the amount of drug *causes* changes in the blood pressure.

The second main use for correlation and regression is to see whether two variables are associated, without necessarily inferring a cause-and-effect relationship. In this case, neither variable is determined by the experimenter; both are naturally variable. If an association is found, the inference is that variation in X may cause variation in Y, or variation in Y may cause variation in X, or variation in some other factor may affect both X and Y. An example would be measurements of the amount of a particular protein on the surface of some cells and the pH of the cytoplasm of those cells. If the protein amount and pH are correlated, it may be that the amount of protein affects the internal pH; or the internal pH affects the amount of protein; or some other factor, such as oxygen concentration, affects both protein concentration and pH. Often, a significant correlation suggests further experiments to test for a cause and effect relationship; if protein concentration and

pH were correlated, you might want to manipulate protein concentration and see what happens to pH, or manipulate pH and measure protein, or manipulate oxygen and see what happens to both.

The third common use of linear regression is estimating the value of one variable corresponding to a particular value of the other variable. For example, if you were doing a protein assay you would start by constructing a standard curve. You would add the reagent to known amounts of protein (10, 20, 30 mg, etc.) and measure the absorbance. You would then find the equation for the regression line, with protein amount as the X variable and absorbance as the Y variable. Then when you measure the absorbance of a sample with an unknown amount of protein, you can rearrange the equation of the regression line to solve for X and estimate the amount of protein in the sample.

Null hypothesis

The null hypothesis is that the slope of the best-fit line is equal to zero; in other words, as the X variable gets larger, the associated Y variable gets neither higher nor lower.

It is also possible to test the null hypothesis that the Y value predicted by the regression equation for a given value of X is equal to some theoretical expectation; the most common would be testing the null hypothesis that the Y intercept is 0. This is rarely necessary in biological experiments, so I won't cover it here, but be aware that it is possible.

Independent vs. dependent variables

If a cause-and-effect relationship is being tested, the variable that causes the relationship is called the independent variable and is plotted on the X axis, while the effect is called the dependent variable and is plotted on the Y axis. In some cases the experimenter determines the value of the independent variable, such as putting frogs in temperature-controlled chambers and measuring their calling rate. In other cases, both variables exhibit natural variation, but any cause-and-effect relationship would be in one way; if you measure the air temperature and frog calling rate at a pond on several different nights, both the air temperature and the calling rate would display natural variation, but if there's a cause-and-effect relationship, it's temperature affecting calling rate; the rate at which frogs call does not affect the air temperature.

Sometimes it's not clear which is the independent variable and which is the dependent. For example, if you measure the salt content of people's food and their blood pressure to test whether eating more salt causes higher blood pressure, you'd want to make salt content the independent variable. But if you thought that high blood pressure caused people to crave high-salt foods, you'd make blood pressure the independent variable.

Sometimes, you're not looking for a cause-and-effect relationship at all; if you measure the range-of-motion of the hip and the shoulder, you're not trying to see whether more flexible hips cause more flexible shoulders, or vice versa, you're just trying to see if people with more flexible hips also tend to have more flexible shoulders, presumably due to some factor (age, diet, exercise, genetics) that affects overall flexibility. In this case, it would be completely arbitrary which variable you put on the X axis and which you put on the Y axis.

Fortunately, the P-value and the r^2 are not affected by which variable you call the X and which you call the Y; you'll get mathematically identical values either way. The regression line *does* depend on which variable is the X and which is the Y; the two lines can be quite different if the r^2 is low. If you're truly interested only in whether the two variables covary, and you are not trying to infer a cause-and-effect relationship, you may want to avoid using the linear regression line as decoration on your graph.

In some fields, the independent variable is traditionally plotted on the Y axis. In oceanography, for example, depth is often plotted on the Y axis (with 0 at the top) and a variable that is directly or indirectly affected by depth, such as chlorophyll concentration, is plotted on the X axis. I wouldn't recommend this unless it's a really strong tradition in your field, as it could lead to confusion about which variable is the independent variable in a linear regression.

Assumptions

No "error" in X variable. One assumption of linear regression is that the X variable is set by the experimenter and there is no error, either measurement error or biological variation. If you're only using the regression line to illustrate the relationship between two variables (basically, it's decoration on your graph), violation of this assumption doesn't matter much. If you're trying to accurately predict Y from X or predict X from Y and the X variable has a lot of measurement error or biological variation, you may want to look into different techniques for "model II regression," such as "major axis regression" or "reduced major axis regression," which are not covered here.

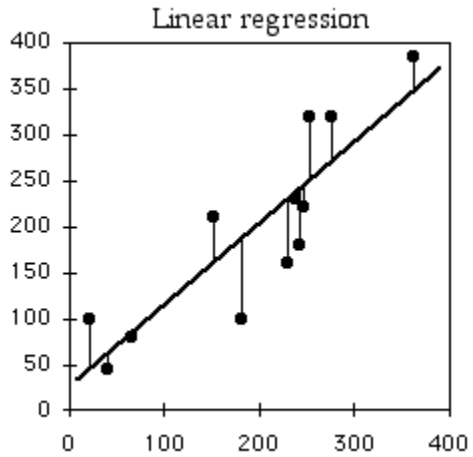
Normality and homoscedasticity. Two more assumptions, similar to those for anova, are that for any value of X, the Y values will be normally distributed and they will be homoscedastic. Although you will rarely have enough data to test these assumptions, they are often violated, especially homoscedasticity. If there is a significant regression, X values with higher mean Y values will often have higher variances of Y as well. A data transformation of the Y variable may fix this problem, but if that doesn't work, you can use the non-parametric Spearman rank correlation instead. I don't know how much non-normality or heteroscedasticity are enough to make linear regression and correlation inappropriate.

Linearity. Linear regression assumes that the data fit to a straight line. If this isn't the case, a data transformation may help, or it may be necessary to use polynomial regression.

Independence. Linear regression assumes that the data points are independent of each other, meaning that the value of one data point does not depend on what the value of any other data point is. The most common violation of this assumption is in time series data, where some Y variable has been measured at different times. For example, let's say you've counted the number of elephants in a park in Africa every year. The population either goes up by 10% or goes down by 10% each year, and the direction of change is completely random. The number of elephants in one year is not independent of the number of elephants in the previous year, it is highly dependent on it; if the number of elephants in one year is high, the number in the next year will still be pretty high, even if it goes down by 10%. The direction of change from year to the next is completely random, so you wouldn't expect a significant regression, but this kind of non-independence can give you a "significant" regression much more often than 5% of the time, even when the null hypothesis of no relationship between X and Y is true.

There are special statistical tests for time-series data and other non-independent data (such as data showing spatial autocorrelation). I will not cover them here.

How the test works



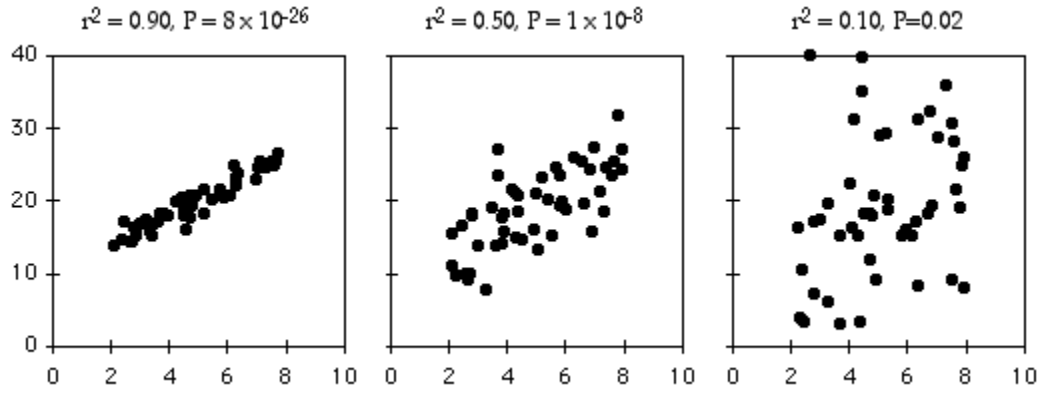
The graph shows the data points (dots), linear regression line (thick line), and data points connected to the point on the regression line with the same X value (thin lines). The regression line is the line that minimizes the sum of the squared vertical distances between the points and the line.

Regression line

Linear regression finds the line that best fits the data points. In this case, the "best" fit is defined as the line that minimizes the squared vertical distances between the data points and the line. For a data point with an X value of X_1 and a Y value of Y_1, the difference between Y_1 and the value of Y on the line at X_1 is calculated, then squared. This squared deviate is calculated for each data point, and the sum of these squared deviates measures how well a line fits the data. The regression line is the one for which this sum of squared deviates is smallest.

The equation for the regression line is usually expressed as Y=intercept+slope×X. This equation can be used to predict the value of Y for a given value of X. You can also predict X from Y, using the equation X=(Y−intercept)/slope. These predictions are best done within the range of X and Y values observed in the data (interpolation). Predicting Y or X values outside the range of observed values (extrapolation) is sometimes interesting, but it can easily yield ridiculous results. For example, in the frog example below, you could mathematically predict that the inter-call interval would be about 16 seconds at -40°C. Actually, the frogs would not call at that temperature; they'd be dead.

Coefficient of determination (r^2)



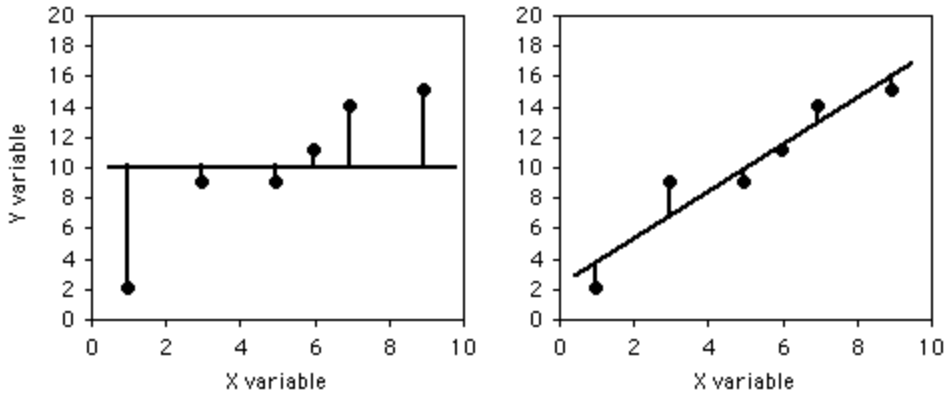
Three relationships with the same slope, same intercept, and different amounts of scatter around the best-fit line.

The coefficient of determination, or r^2, expresses the strength of the relationship between the X and Y variables. It is the proportion of the variation in the Y variable that is "explained" by the variation in the X variable. r^2 can vary from 0 to 1; values near 1 mean the Y values fall almost right on the regression line, while values near 0 mean there is very little relationship between X and Y. As you can see, regressions can have a small r^2 and not look like there's any relationship, yet they still might have a slope that's significantly different from zero.

To illustrate the meaning of r^2, here are six pairs of X and Y values:

X	Y	deviate from mean	squared deviate
1	2	8	64
3	9	1	1
5	9	1	1
6	11	1	1
7	14	4	16
9	15	5	25
		sum of squares:	108

If you didn't know anything about the X value and were told to guess what a Y value was, your best guess would be the mean Y; for this example, the mean Y is 10. The squared deviates of the Y values from their mean is the total sum of squares, familiar from analysis of variance. The vertical lines on the left graph below show the deviates from the mean; the first point has a deviate of 8, so its squared deviate is 64, etc. The total sum of squares for these numbers is 64+1+1+1+16+25=108.



Deviations from the mean Y and from the regression line.

If you did know the X value and were told to guess what a Y value was, you'd calculate the regression equation and use it. The regression equation for these numbers is Y=1.5429×X+2.0286, so for the first X value you'd predict a Y value of 1.5429×1+2.0286=3.5715, etc. The vertical lines on the right graph above show the deviates of the actual Y values from the predicted Y values. As you can see, most of the points are closer to the regression line than they are to the overall mean. Squaring these deviates and taking the sum gives us the regression sum of squares, which for these numbers is 10.8.

X	Y	predicted Y-value	deviate from predicted	squared deviate
1	2	3.57	1.57	2.46
3	9	6.66	2.34	5.48
5	9	9.74	0.74	0.55
6	11	11.29	0.29	0.08
7	14	12.83	1.17	1.37
9	15	15.91	0.91	0.83
				10.8

regression sum of squares: 10.8

The regression sum of squares is 10.8, which is 90% smaller than the total sum of squares (108). This difference between the two sums of squares, expressed as a fraction of the total sum of squares, is the r^2. In this case we would say that r^2=0.90; the X-variable "explains" 90% of the variation in the Y-variable.

The r^2 value is formally known as the "coefficient of determination," although it is usually just called r^2. The square root of r^2, with a negative sign if the slope is negative, is the Pearson product-moment correlation coefficient, or just "correlation coefficient." Either r or r^2 can be used to describe the strength of the association between two variables, but I recommend r^2, because it has a more understandable meaning (the proportional difference between total sum of squares and regression sum of squares) and doesn't have those annoying negative values.

Test statistic

The test statistic for a linear regression is $t_s = d.f. \times r^2/(1-r^2)$. It gets larger as the degrees of freedom (n−2) get larger or the r^2 gets larger. Under the null hypothesis, t_s is t-distributed with n−2 degrees of freedom. When reporting the results of a linear regression, it is conventional to report just the r^2 and degrees of freedom, not the t_s value. Anyone who really needs the t_s value can calculate it from the r^2 and degrees of freedom.

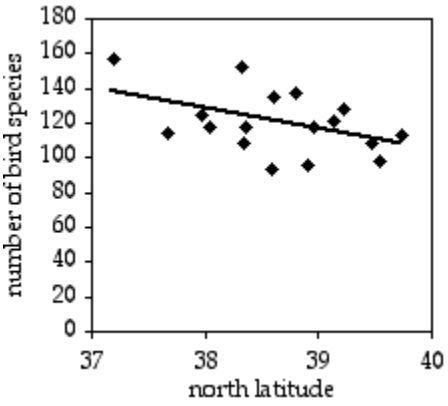
It is also possible to square t_s and get an F-statistic with 1 degree of freedom in the numerator and n−2 degrees of freedom in the denominator. The resulting P-value is mathematically identical to that calculated with t_s.

Examples

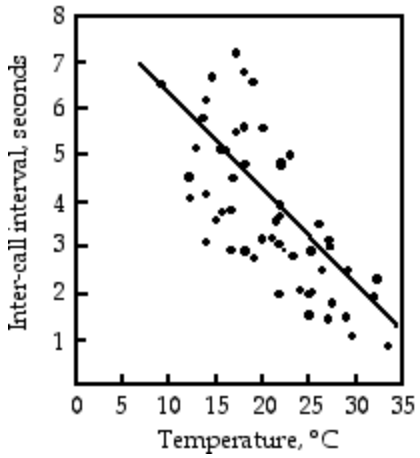
A common observation in ecology is that species diversity decreases as you get further from the equator. To see whether this pattern could be seen on a small scale, I used data from the Audobon Society's Christmas Bird Count (http://www.audubon.org/bird/cbc/index.html) , in which birders try to count all the birds in a 15-mile diameter area during one winter day. I looked at the total number of species seen in each area on the Delmarva Peninsula during the 2005 count. Latitude and number of bird species are the two measurement variables; location is the hidden nominal variable.

Location	Latitude	Number of species
Bombay Hook, DE	39.217	128
Cape Henlopen, DE	38.800	137
Middletown, DE	39.467	108
Milford, DE	38.958	118
Rehoboth, DE	38.600	135
Seaford-Nanticoke, DE	38.583	94
Wilmington, DE	39.733	113
Crisfield, MD	38.033	118
Denton, MD	38.900	96
Elkton, MD	39.533	98
Lower Kent County, MD	39.133	121
Ocean City, MD	38.317	152
Salisbury, MD	38.333	108
S. Dorchester County, MD	38.367	118
Cape Charles, VA	37.200	157
Chincoteague, VA	37.967	125
Wachapreague, VA	37.667	114

The result is r^2=0.214, with 15 d.f., so the P-value is 0.061. The trend is in the expected direction, but it is not quite significant. The equation of the regression line is number of species=−12.039(latitude)+585.14. Even if it were significant, I don't know what you'd do with the equation; I suppose you could extrapolate and use it to predict that above the 49th parallel, there would be fewer than zero bird species.



Latitude and bird species on the Delmarva Peninsula.



Relationship of body temperature and inter-call interval in the gray tree frog.

Gayou (1984) measured the intervals between male mating calls in the gray tree frog, *Hyla versicolor*, at different temperatures. The regression line is interval=−0.205(temperature)+8.36, and it is highly significant (r^2=0.29, 45 d.f., $p=9\times10^{-5}$). You could rearrange the equation, temperature=(interval−8.36)/(−0.205), measure the interval between frog mating calls, and estimate the air temperature. Or you could buy a thermometer.

Goheen et al. (2003) captured 14 female northern grasshopper mice (*Onchomys leucogaster*) in north-central Kansas, measured the body length, and counted the number of offspring. There are two measurement variables, body length and number of offspring, and the authors were interested in whether larger body size causes an increase in the number of offspring, so they did a linear regression. The results are significant: r^2=0.46, 12 d.f., P=0.008. The equation of the regression line is offspring=−7.88+0.108(length).

Graphing the results

In a spreadsheet, you show the results of a regression on a scatter graph, with the independent variable on the X axis. To add the regression line to the graph, finish making the graph, then select the graph and go to the Chart menu. Choose "Add Trendline" and choose the straight line. If you want to show the regression line extending beyond the observed range of X-values, choose "Options" and adjust the "Forecast" numbers until you get the line you want.

If you have transformed your data for the regression, don't plot the untransformed data; instead, plot the transformed data. See the Excel or Calc graph instructions for details on how to do this.

Similar tests

Sometimes it is not clear whether an experiment includes one measurement variable and two nominal variables, and should be analyzed with a two-way anova or paired t-test, or includes two measurement variables and one nominal variable, and should be analyzed with correlation and regression. In that case, your choice of test is determined by the biological question you're interested in. For example, let's say you've measured the range of motion of the right shoulder and left shoulder of a bunch of right-handed people. If your question is "Is there an association between the range of motion of people's right and left shoulders--do people with more flexible right shoulders also tend to have more flexible left shoulders?", you'd treat "right shoulder range-of-motion" and "left shoulder range-of-motion" as two different measurement variables, and individual as one nominal variable, and analyze with correlation and regression. If your question is "Is the right shoulder more flexible than the left shoulder?", you'd treat "range of motion" as one measurement variable, "right vs. left" as one nominal variable, individual as one nominal variable, and you'd analyze with two-way anova or a paired t-test.

If the dependent variable is a percentage, such as percentage of people who have heart attacks on different doses of a drug, it's really a nominal variable, not a measurement. Each individual observation is a value of the nominal variable ("heart attack" or "no heart attack"); the percentage is not really a single observation, it's a way of summarizing a bunch of observations. One approach for percentage data is to arcsine transform the percentages and analyze with correlation and linear regression. You'll see this in the literature, and it's not horrible, but it's better to analyze using logistic regression.

If the relationship between the two measurement variables is best described by a curved line, not a straight one, one possibility is to try different transformations on one or both of the variables. The other option is to use polynomial regression (also known as curvilinear regression).

Linear regression assumes that the Y variables for any value of X would be normally distributed and homoscedastic; if these assumptions are violated, Spearman rank correlation, the non-parametric analog of linear regression, may be used.

To compare two or more regression lines to each other, use ancova. If there are more than two measurement variables, use multiple regression.

How to do the test

Spreadsheet

I have put together a spreadsheet to do linear regression on up to 1000 pairs of observations. It provides the following:

- The regression coefficient (the slope of the regression line).
- The Y-intercept. With the slope and the intercept, you have the equation for the regression line: Y=a+bX, where *a* is the Y intercept and *b* is the slope.
- The r^2 value.
- The degrees of freedom. There are n−2 degrees of freedom in a regression, where *n* is the number of observations.
- The P-value. This gives you the probability of finding a slope that is as large or larger than the observed slope, under the null hypothesis that the true slope is 0.
- A Y-estimator and an X-estimator. This enables you to enter a value of X and find the corresponding value of Y on the best-fit line, or vice-versa. This would be useful for constructing standard curves, such as used in protein assays for example.

Web pages

Web pages that will perform linear regression are here, (http://faculty.vassar.edu/lowry/corr_stats.html) here, (http://home.ubalt.edu/ntsbarsh/Business-stat/otherapplets/Regression.htm) and here. (http://www.physics.csbsju.edu/stats/QF_NROW_form.html) They all require you to enter each number individually, and thus are inconvenient for large data sets. This web page (http://www.fon.hum.uva.nl/Service/Statistics/Correlation_coefficient.html) does linear regression and lets you paste in a set of numbers, which is more convenient for large data sets.

SAS

You can use either PROC GLM or PROC REG for a simple linear regression; since PROC REG is also used for multiple regression, you might as well learn to use it. Here's an example using the bird data from above.

```
data birds;
   input town $ state $ latitude species;
   cards;
Bombay_Hook          DE     39.217     128
Cape_Henlopen        DE     38.800     137
Middletown           DE     39.467     108
Milford              DE     38.958     118
Rehoboth             DE     38.600     135
Seaford-Nanticoke    DE     38.583      94
Wilmington           DE     39.733     113
```

```
Crisfield            MD     38.033     118
Denton               MD     38.900      96
Elkton               MD     39.533      98
Lower_Kent_County    MD     39.133     121
Ocean_City           MD     38.317     152
Salisbury            MD     38.333     108
S_Dorchester_County  MD     38.367     118
Cape_Charles         VA     37.200     157
Chincoteague         VA     37.967     125
Wachapreague         VA     37.667     114
;
proc reg data=birds;
   model species=latitude;
   run;
```

The output includes an analysis of variance table. Don't be alarmed by this; if you dig down into the math, regression is just another variety of anova. Below the anova table are the r^2, slope, intercept, and P-value:

```
Root MSE              16.37357    R-Square     0.2143 r^2
Dependent Mean       120.00000    Adj R-Sq     0.1619
Coeff Var             13.64464

                      Parameter Estimates

               Parameter   Standard
Variable   DF   Estimate      Error  t Value  Pr > |t|

               intercept
Intercept   1  585.14462  230.02416     2.54    0.0225
latitude    1  -12.03922    5.95277    -2.02    0.0613 P-value
                slope
```

These results indicate an r^2 of 0.21, intercept of 585.1, a slope of −12.04, and a P-value of 0.061.

Power analysis

The G*Power (http://www.psycho.uni-duesseldorf.de/abteilungen/aap/gpower3/) program will calculate the sample size needed for a regression/correlation. The effect size is the absolute value of the correlation coefficient *r*; if you have r^2, take the positive square root of it. Choose "t tests" from the "Test family" menu and "Correlation: Point biserial model" from the "Statistical test" menu. Enter the r-value you hope to see, your alpha (usually 0.05) and your power (usually 0.80 or 0.90).

For example, let's say you want to look for a relationship between calling rate and temperature in the barking tree frog, *Hyla gratiosa*. Gayou (1984) found an r^2 of 0.29 in the *H. versicolor*, so you decide you want to be able to detect an r^2 of 0.25 or more. The square root of 0.25 is 0.5, so you enter 0.5 for "Effect size", 0.05 for alpha,

and 0.8 for power. The result is 26 observations of temperature and frog calling rate.

It's important to note that the distribution of X variables, in this case air temperatures, should be the same for the proposed study as for the pilot study the sample size calculation was based on. Gayou (1984) measured frog calling rate at temperatures that were fairly evenly distributed from 10°C to 34°C. If you looked at a narrower range of temperatures, you'd need a lot more observations to detect the same kind of relationship.

Further reading

Sokal and Rohlf, pp. 451-471, 486-493.

Zar, pp. 324-358, 377-386.

References

Gayou, D.C. 1984. Effects of temperature on the mating call of *Hyla versicolor*. Copeia 1984: 733-738.

Goheen, J.R., G.A. Kaufman, and D.W. Kaufman. 2003. Effect of body size on reproductive characteristics of the northern grasshopper mouse in north-central Kansas. Southwest. Naturalist 48: 427-431.

Spearman rank correlation

When to use it

Spearman rank correlation is used when you have two measurement variables and one "hidden" nominal variable. The nominal variable groups the measurements into pairs; if you've measured height and weight of a bunch of people, "individual name" is a nominal variable. You want to see whether the two measurement variables covary; whether, as one variable increases, the other variable tends to increase or decrease. It is the non-parametric alternative to correlation, and it is used when the data do not meet the assumptions about normality, homoscedasticity and linearity. Spearman rank correlation is also used when one or both of the variables consists of ranks.

You will rarely have enough data in your own data set to test the normality and homoscedasticity assumptions of regression and correlation; your decision about whether to do linear regression and correlation or Spearman rank correlation will usually depend on your prior knowledge of whether the variables are likely to meet the assumptions.

Null hypothesis

The null hypothesis is that the ranks of one variable do not covary with the ranks of the other variable; in other words, as the ranks of one variable increase, the ranks of the other variable are not more likely to increase (or decrease).

How the test works

Spearman rank correlation works by converting each variable to ranks. Thus, if you you're doing a Spearman rank correlation of blood pressure vs. body weight, the lightest person would get a rank of 1, second-lightest a rank of 2, etc. The lowest blood pressure would get a rank of 1, second lowest a rank of 2, etc. If one or both variables is already ranks, they remain unchanged, of course. When two or more observations are equal, the average rank is used. For example, if two observations are tied for the second-highest rank, they would get a rank of 2.5 (the average of 2 and 3).

Once the two variables are converted to ranks, a correlation analysis is done on the ranks. The correlation coefficient is calculated for the two columns of ranks, and the significance of this is tested in the same way as the correlation coefficient for a regular correlation. (This Spearman's correlation coefficient is also called Spearman's rho). The P-value from the correlation of ranks is the P-value of the Spearman rank correlation. The ranks are rarely graphed against each other, and a line is rarely used for either predictive or illustrative purposes, so you don't calculate a non-parametric equivalent of the regression line.

Example

Males of the magnificent frigatebird (*Fregata magnificens*) have a large red throat pouch. They visually display this pouch and use it to make a drumming sound when seeking mates. Madsen et al. (2004) wanted to know whether females, who presumably choose mates based on their pouch size, could use the pitch of the drumming sound as an indicator of pouch size. The authors estimated the volume of the pouch and the fundamental frequency of the drumming sound in 18 males:

```
Volume, cm^3  Frequency, Hz
1760           529
```
====See the web page for the full data set====
```
7960           416
```

There are two measurement variables, pouch size and pitch; the identity of each male is the hidden nominal variable. The authors analyzed the data using Spearman rank correlation, which converts the measurement variables to ranks, and the relationship between the variables is significant (Spearman's rho=-0.76, 16 d.f., P=0.0002). The authors do not explain why they used Spearman rank correlation; if they had used regular correlation, they would have obtained r=-0.82, P=0.00003.

Graphing the results

If you have measurement data for both of the X and Y variables, you could plot the results the same way you would for a linear regression. Don't put a regression line on the graph, however; you can't plot a rank correlation line on a graph with measurement variables on the axes, and it would be misleading to put a linear regression line on a graph when you've analyzed it with rank correlation.

If you actually have true ranked data for both variables, you could plot a line through them, I suppose. I'm not sure what the point would be, however.

How to do the test

Spreadsheet

I've put together a spreadsheet that will perform a Spearman rank correlation on up to 1000 observations. With small numbers of observations (10 or fewer), the P-value based on the equation using r^2 is inaccurate, so the spreadsheet looks up the P-value in a table of critical values.

Web page

This web page (http://faculty.vassar.edu/lowry/corr_rank.html) will do Spearman rank correlation.

SAS

Use PROC CORR with the SPEARMAN option to do Spearman rank correlation. Here is an example using the bird data from the correlation and regression web page:

```
proc corr data=birds spearman;
   var species latitude;
   run;
```

The results include the Spearman correlation coefficient, analagous to the r-value of a regular correlation, and the P-value:

```
Spearman Correlation Coefficients, N = 17
     Prob > |r| under H0: Rho=0

           species   latitude

species    1.00000   -0.36263 Spearman correlation coefficient
                      0.1526  P-value

latitude -0.36263    1.00000
           0.1526
```

Further reading

Sokal and Rohlf, pp. 598, 600.

Zar, pp. 395-398.

Reference

Madsen, V., T.J.S. Balsby, T. Dabelsteen, and J.L. Osorno. 2004. Bimodal signaling of a sexually selected trait: gular pouch drumming in the magnificent frigatebird. Condor 106: 156-160.

Polynomial regression

When to use it

Sometimes, when you analyze data with correlation and linear regression, you notice that the relationship between the independent (X) variable and dependent (Y) variable looks like it follows a curved line, not a straight line. In that case, the linear regression line will not be very good for describing and predicting the relationship, and the P-value may not be an accurate test of the hypothesis that the variables are not associated.

Your first choice when faced with a curved relationship between two measurement variables should be to try data transformations on one or both of the variables. Often, this will straighten out a simple J-shaped curve. If that doesn't work, you can try curvilinear regression, in which a more complicated equation than the linear regression equation is fit to the data. Equations with a variety of terms will produce curved lines, including exponential (involving b^X, where *b* is a constant), power (involving X^b), logarithmic (involving log(X)), and trigonometric (involving sine, cosine, or or other trigonometric functions). For any particular form of equation involving such terms, it is possible to find the equation for the curved line that best fits the data points, and to compare the fit of the more complicated equation to that of a simpler equation (such as the equation for a straight line).

Here I will use polynomial regression as one example of curvilinear regression. A polynomial equation has X raised to integer powers such as X^2 and X^3. A quadratic equation has the form $Y=a+b_1X+b_2X^2$, where *a* is the Y-intercept and b_1 and b_2 are constants. It produces a parabola. A cubic equation has the form $Y=a+b_1X+b_2X^2+b_3X^3$ and produces an S-shaped curve, while a quartic equation has the form $Y=a+b_1X+b_2X^2+b_3X^3+b_4X^4$ and can produce M or W shaped curves. You can fit higher-order polynomial equations, but it is very unlikely that you would want to use anything more than the cubic in biology.

Null hypotheses

Several null hypotheses are tested while doing polynomial regression. The first null hypothesis is that a quadratic equation does not fit the data significantly better than a linear equation; the next null hypothesis may be that a cubic equation does

not fit the data significantly better than a quadratic equation, and so on. There is also a null hypothesis for each equation that says that it does not fit the data significantly better than a horizontal line; in other words, that there is no relationship between the X and Y variables.

How the test works

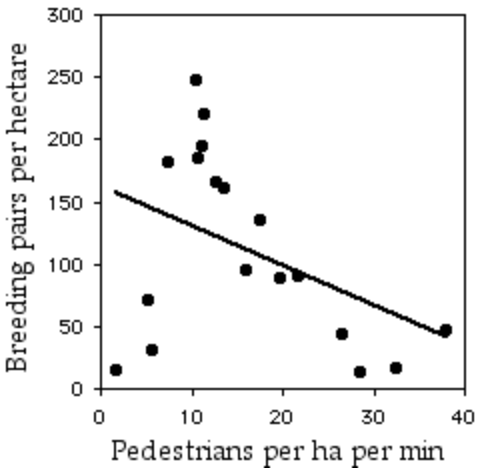
In polynomial regression, different powers of the X variable (X, X^2, X^3…) are added to an equation to see whether they increase the r^2 significantly. First a linear regression is done, fitting an equation of the form $Y=a+bX$ to the data. Then an equation of the form $Y=a+b_1X+b_2X^2$, which produces a parabola, is fit to the data. The r^2 will always increase when you add a higher-order term, but the question is whether the increase in r^2 is significantly greater than expected due to chance. Next, an equation of the form $Y=a+b_1X+b_2X^2+b_3X^3$, which produces an S-shaped line, is fit and the increase in r^2 is tested. This can continue until adding another term does not increase r^2 significantly, although in most cases it is hard to imagine a biological meaning for exponents greater than 3. Once the best-fitting equation is chosen, it is tested to see whether it fits the data significantly better than an equation of the form $Y=a$; in other words, a horizontal line.

Even though the usual procedure is to test the linear regression first, then the quadratic, then the cubic, you don't need to stop if one of these is not significant. For example, if the graph looks U-shaped, the linear regression may not be significant, but the quadratic will be.

Examples

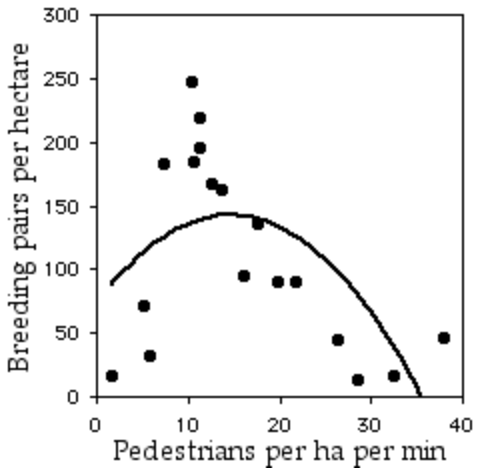
Fernandez-Juricic et al. (2003) examined the effect of human disturbance on the nesting of house sparrows (*Passer domesticus*). They counted breeding sparrows per hectare in 18 parks in Madrid, Spain, and also counted the number of people per minute walking through each park (both measurement variables); the identity of the park is the hidden nominal variable.

The linear regression is not significant (r^2=0.174, 16 d.f., P=0.08).



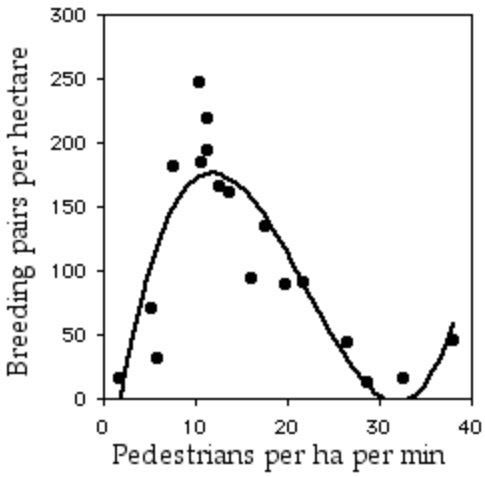
Graph of sparrow abundance vs. human disturbance with linear regression line.

The quadratic regression is significant (r^2=0.372, 15 d.f., P=0.03), and it is significantly better than the linear regression (P=0.03). This seems biologically plausible; the data suggest that there is some intermediate level of human traffic that is best for house sparrows. Perhaps areas with too many humans scare the sparrows away, while areas with too few humans favor other birds that outcompete the sparrows for nest sites or something.



Graph of sparrow abundance vs. human disturbance with quadratic regression line.

The cubic graph is significant (r^2=0.765, 14 d.f., P=0.0001), and the increase in r^2 between the cubic and the quadratic equation is highly significant ($P=1\times10^{-5}$). The cubic equation is $Y=0.0443x^3-2.916x^2+50.601x-87.765$. The quartic equation does not fit significantly better than the cubic equation (P=0.80). Even though the cubic equation fits significantly better than the quadratic, it's more difficult to imagine a plausible biological explanation for this. I'd want to see more samples from areas with more than 35 people per hectare per minute before I accepted that the sparrow abundance really starts to increase again above that level of pedestrian traffic.

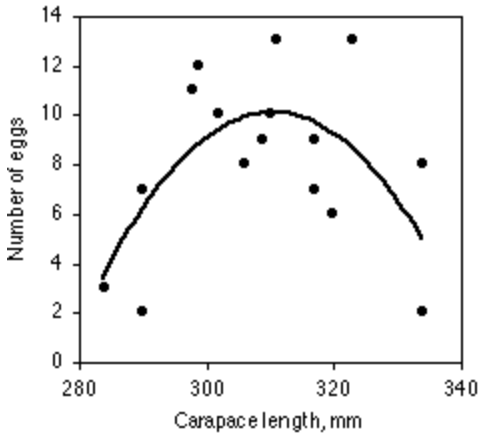


Graph of sparrow abundance vs. human disturbance with cubic regression line.

Ashton et al. (2007) measured the carapace length (in mm) of 18 female gopher tortoises (*Gopherus polyphemus*) in Okeeheelee County Park, Florida, and X-rayed them to count the number of eggs in each. The data are shown below in the SAS example. The linear regression is not significant (r^2=0.015, 16 d.f., P=0.63), but the quadratic is significant (r^2=0.43, 15 d.f., P=0.014). The increase in r^2 from linear to quadratic is significant (P= 0.001). The best-fit quadratic equation is $Y=-899.9+5.857X-0.009425X^2$. Adding the cubic and quartic terms does not significantly increase the r^2.

The first part of the graph is not surprising; it's easy to imagine why bigger tortoises would have more eggs. The decline in egg number above 310 mm carapace length is the interesting result; it suggests that egg production declines in these tortoises as they get old and big.

Graphing the results



Graph of clutch size (number of eggs) vs. carapace length, with best-fit quadratic line.

As shown above, you graph a polynomial regression the same way you would a linear regression, a scattergraph with the independent variable on the X-axis and the dependent variable on the Y-axis. In general, you shouldn't show the regression line for values outside the range of observed X values, as extrapolation with polynomial regression is even more likely than linear regression to yield ridiculous results. For example, extrapolating the quadratic equation relating tortoise carapace length and number of eggs predicts that tortoises with carapace length less than 279 mm or greater than 343 mm would have negative numbers of eggs.

Similar tests

Before performing a polynomial regression, you should try different transformations when faced with an obviously curved relationship between an X and a Y variable. A linear equation relating transformed variables is simpler and more elegant than a curvilinear equation relating untransformed variables. You should also remind yourself of your reason for doing a regression. If your purpose is prediction of unknown values of Y corresponding to known values of X, then you need an equation that fits the data points well, and a polynomial regression may be appropriate if transformations do not work. However, if your purpose is testing the null hypothesis that there is no relationship between X and Y, and a linear regression gives a significant result, you may want to stick with the linear regression even if polynomial gives a significantly better fit. Using a less-familiar technique that yields a more-complicated equation may cause your readers to be a bit suspicious of your results; they may feel you went fishing around for a statistical test that supported your hypothesis, especially if there's no obvious biological reason for an equation with terms containing exponents.

Spearman rank correlation is a nonparametric test of the association between two variables. It will work well if there is a steady increase or decrease in Y as X increases, but not if Y goes up and then goes down.

Polynomial regression is a form of multiple regression. In multiple regression, there is one dependent (Y) variable and multiple independent (X) variables, and the X variables (X_1, X_2, X_3...) are added to the equation to see whether they

increase the R^2 significantly. In polynomial regression, the independent "variables" are just X, X^2, X^3, etc.

How to do the test

Spreadsheet

I have prepared a spreadsheet that will help you perform a polynomial regression. It tests equations up to fourth order, and it will handle up to 1000 observations.

Web pages

There is a very powerful web page (http://StatPages.org/nonlin.html) that will fit just about any equation you can think of to your data (not just polynomial). Another web page that will fit any of 15 commonly used equations is here; (http://www.colby.edu/chemistry/PChem/scripts/lsfitpl.html") it is easier to use, and even draws a graph. This web page (http://www3.sympatico.ca/mcomeau/webpublic/javapage/reg/reg.htm) only does polynomial regression, but is very fast and easy to use.

SAS

To do polynomial regression in SAS, you create a data set containing the square of the independent variable, the cube, etc. You then use PROC REG for models containing the higher-order variables. It's possible to do this as a multiple regression, but I think it's less confusing to use multiple model statements, adding one term to each model. There doesn't seem to be an easy way to test the significance of the increase in r^2 in SAS, so you'll have to do that by hand. If r^2_i is the r^2 for the i_{th} order, and r^2_j is the r^2 for the next higher order, and $d.f._j$ is the degrees of freedom for the higher-order equation, the F-statistic is $d.f._j \times (r^2_j - r^2_i)/(1 - r^2_j)$. It has *j* degrees of freedom in the numerator and $d.f._j = n - j - 1$ degrees of freedom in the denominator.

Here's an example, using the data on tortoise carapace length and clutch size from Ashton et al. (2007).

```
data turtles;
   input length clutch;
   cards;
284     3
290     2
290     7
290     7
298     11
299     12
302     10
306     8
306     8
```

```
309     9
310     10
311     13
317     7
317     9
320     6
323     13
334     2
334     8
;
data turtlepower; set turtles;
   length2=length*length;
   length3=length*length*length;
   length4=length*length*length*length;
proc reg data=turtlepower;
   model clutch=length;
   model clutch=length length2;
   model clutch=length length2 length3;
   run;
```

In the output, first look for the r^2 values under each model:

```
                    The REG Procedure
                     Model: MODEL1
                  Dependent Variable: clutch
                           .
                           .
                           .
Root MSE              3.41094    R-Square    0.0148 linear r-sq
Dependent Mean        8.05556    Adj R-Sq   -0.0468
Coeff Var            42.34268
                           .
                           .
                           .
                    The REG Procedure
                       Model: MODEL2
                  Dependent Variable: clutch
                           .
                           .
                           .
Root MSE              2.67050    R-Square    0.4338 quadratic r-sq
Dependent Mean        8.05556    Adj R-Sq    0.3583
Coeff Var            33.15104
```

For this example, n=18. The F-statistic for the increase in r^2 from linear to quadratic is $15\times(0.4338-0.0148)/(1-0.4338)=11.10$ with d.f.=2, 15. Using a spreadsheet (enter =FDIST(11.10, 2, 15)) or an online F-statistic calculator, (http://www.danielsoper.com/statcalc/calc07.as) this gives a P-value of 0.0011.

Once you've figured out which equation is best (the quadratic, for our example, since the cubic and quartic equations do not significantly increase the r^2), look for the parameters in the output:

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-899.93459	270.29576	-3.33	0.0046
length	1	5.85716	1.75010	3.35	0.0044
length2	1	-0.00942	0.00283	-3.33	0.0045

This tells you that the equation for the best-fit quadratic curve is $Y=-899.9+5.857X-0.00942X^2$.

Further reading

Sokal and Rohlf, pp. 665-670.

Zar, pp. 452-459.

References

Ashton, K.G., R.L. Burke, and J.N. Layne. 2007. Geographic variation in body and clutch size of gopher tortoises. Copeia 2007: 355-363.

Fernandez-Juricic, E., A. Sallent, R. Sanz, and I. Rodriguez-Prieto. 2003. Testing the risk-disturbance hypothesis in a fragmented landscape: non-linear responses of house sparrows to humans. Condor 105: 316-326.

Analysis of covariance

When to use it

Analysis of covariance (ancova) is used when you have two measurement variables and two nominal variables. One of the nominal variables groups is the "hidden" nominal variable that groups the measurement observations into pairs, and the other nominal variable divides the regressions into two or more sets.

The purpose of ancova to compare two or more linear regression lines. It is a way of comparing the Y variable among groups while statistically controlling for variation in Y caused by variation in the X variable. For example, let's say you want to know whether the Cope's gray treefrog, *Hyla chrysoscelis,* has a different calling rate than the eastern gray treefrog, *Hyla versicolor*, which has twice as many chromosomes as *H. chrysoscelis* but is morphologically identical. As shown on the regression web page, the calling rate of eastern gray treefrogs is correlated with temperature, so you need to control for that. One way to control for temperature would be to bring the two species of frogs into a lab and keep them all at the same temperature, but you'd have to worry about whether their behavior in an artificial lab environment was really the same as in nature. In addition, you'd want to know whether one species had a higher calling rate at some temperatures, while the other species had a higher calling rate at other temperatures. It might be better to measure the calling rate of each species of frog at a variety of temperatures in nature, then use ancova to see whether the regression line of calling rate on temperature is significantly different between the two species.

Null hypotheses

Two null hypotheses are tested in an ancova. The first is that the slopes of the regression lines are all the same. If this hypothesis is not rejected, the second null hypothesis is tested: that the Y-intercepts of the regression lines are all the same.

Although the most common use of ancova is for comparing two regression lines, it is possible to compare three or more regressions. If their slopes are all the same, it is then possible to do planned or unplanned comparisions of Y-intercepts, similar to the planned or unplanned comparisons of means in an anova. I won't cover that here.

How it works

The first step in performing an ancova is to compute each regression line. In the frog example, there are two values of the species nominal variable, *Hyla chrysoscelis* and *H. versicolor,* so the regression line is calculated for calling rate vs. temperature for each species of frog.

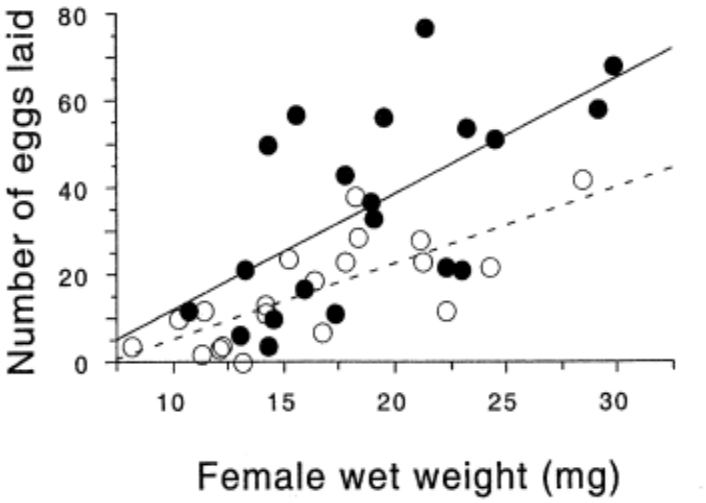
Next, the slopes of the regression lines are compared; the null hypothesis that the slopes are the same is tested. The final step of the anova, comparing the Y-intercepts, cannot be performed if the slopes are significantly different from each other. If the slopes of the regression lines are different, the lines cross each other somewhere, and one group has higher Y values in one part of the graph and lower Y values in another part of the graph. (If the slopes are different, there are techniques for testing the null hypothesis that the regression lines have the same Y-value for a particular X-value, but they're not used very often and I won't consider them here.)

If the slopes are significantly different, the ancova is done, and all you can say is that the slopes are significantly different. If the slopes are not significantly different, the next step in an ancova is to draw a regression line through each group of points, all with the same slope. This common slope is a weighted average of the slopes of the different groups.

The final test in the ancova is to test the null hypothesis that all of the Y-intercepts of the regression lines with a common slope are the same. Because the lines are parallel, saying that they are significantly different at one point (the Y-intercept) means that the lines are different at any point.

Examples

In the firefly species *Photinus ignitus,* the male transfers a large spermatophore to the female during mating. Rooney and Lewis (2002) wanted to know whether the extra resources from this "nuptial gift" enable the female to produce more offspring. They collected 40 virgin females and mated 20 of them to one male and 20 to three males. They then counted the number of eggs each female laid. Because fecundity varies with the size of the female, they analyzed the data using ancova, with female weight (before mating) as the independent measurement variable and number of eggs laid as the dependent measurement variable. Because the number of males has only two values ("one" or "three"), it is a nominal variable, not measurement.

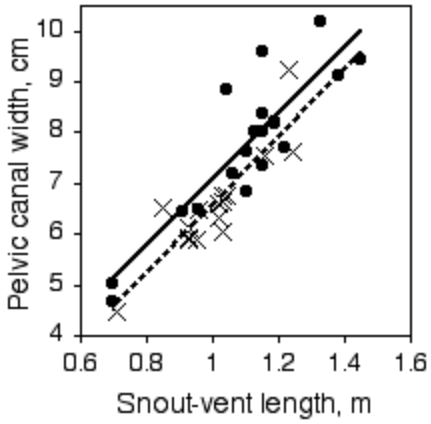


Eggs laid vs. female weight in the firefly *Photinus ignitus*.. Filled circles are females that have mated with three males; open circles are females that have mated with one male.

The slopes of the two regression lines (one for single-mated females and one for triple-mated females) are not significantly different ($F_{1, 36}$=1.1, P=0.30). The Y-intercepts are significantly different ($F_{1, 36}$=8.8, P=0.005); females that have mated three times have significantly more offspring than females mated once.

Paleontologists would like to be able to determine the sex of dinosaurs from their fossilized bones. To see whether this is feasible, Prieto-Marquez et al. (2007) measured several characters that are thought to distinguish the sexes in alligators (*Alligator mississipiensis*), which are among the closest living relatives of dinosaurs. One of the characters was pelvic canal width, which they wanted to standardize using snout-vent length. The raw data are shown in the SAS example below.

The slopes of the regression lines are not significantly different (P=0.9101). The Y-intercepts are significantly different (P=0.0267), indicating that male alligators of a given length have a significantly greater pelvic canal width. However, inspection of the graph shows that there is a lot of overlap between the sexes even after standardizing for sex, so it would not be possible to reliably determine the sex of a single individual with this character alone.



Pelvic canal width vs. snout-vent length in the American alligator. Circles and solid line are males; X's and dashed line are females.

Graphing the results

Data for an ancova are shown on a scattergraph, with the independent variable on the X-axis and the dependent variable on the Y-axis. A different symbol is used for each value of the nominal variable, as in the firefly graph above, where filled circles are used for the thrice-mated females and open circles are used for the once-mated females. To get this kind of graph in a spreadsheet, you would put all of the X-values in column A, one set of Y-values in column B, the next set of Y-values in column C, and so on.

Most people plot the individual regression lines for each set of points, as shown in the firefly graph, even if the slopes are not significantly different. This lets people see how similar or different the slopes look. This is easy to do in a spreadsheet; just click on one of the symbols and choose "Add Trendline" from the Chart menu.

Similar tests

One alternative technique that is sometimes possible is to take the ratio of the two measurement variables, then use a one-way anova. For the mussel example I used for testing the homogeneity of means in one-way anova, I standardized the length of the anterior adductor muscle by dividing by the total length. There are technical problems with doing statistics on ratios of two measurement variables (the ratio of two normally distributed variables is not normally distributed), but if you can safely assume that the regression lines all pass through the origin (in this case, that a mussel that was 0 mm long would have an AAM length of 0 mm), this is not an unreasonable thing to do, and it simplifies the statistics. It would be

important to graph the association between the variables and analyze it with linear regression to make sure that the relationship is linear and does pass through the origin.

Sometimes the two measurement variables are just the same variable measured at different times or places. For example, if you measured the weights of two groups of individuals, put some on a new weight-loss diet and the others on a control diet, then weighed them again a year later, you could treat the difference between final and initial weights as a single variable, and compare the mean weight loss for the control group to the mean weight loss of the diet group using a one-way anova. The alternative would be to treat final and initial weights as two different variables and analyze using an ancova: you would compare the regression line of final weight vs. initial weight for the control group to the regression line for the diet group. The anova would be simpler, and probably perfectly adequate; the ancova might be better, particularly if you had a wide range of initial weights, because it would allow you to see whether the change in weight depended on the initial weight.

One nonparametric alternative to ancova is to convert the measurement variables to ranks, then do a regular ancova on the ranks; see Conover and Iman (1982) for the details. There are several other versions of nonparametric ancova, but they appear to be less popular, and I don't know the advantages and disadvantages of each.

How to do the test

Spreadsheet and web pages

Richard Lowry has made web pages (http://faculty.vassar.edu/lowry/vsancova.html) that allow you to perform ancova with two, three or four groups, and a downloadable spreadsheet for ancova with more than four groups. You may cut and paste data from a spreadsheet to the web pages. In the results, the P-value for "adjusted means" is the P-value for the difference in the intercepts among the regression lines; the P-value for "between regressions" is the P-value for the difference in slopes. One bug in the web pages is that very small values of P are not represented correctly. If the web page gives you a strange P-value (negative, greater than 1, "5e-7"), use the FDIST function of a spreadsheet along with the F value and degrees of freedom from the web page to calculate the correct P value. For example, if the F-value for the adjusted means is 281.37, the d.f. for the adjusted means is 1 and the d.f. for the adjusted error is 84, go to a spreadsheet and enter "=FDIST(281.37, 1, 84)" to get the correct P-value. To get the P-value for the slopes, use the d.f. for "between regressions" and "remainder."

SAS

Here's an illustration of how to do analysis of covariance in SAS, using the data from Prieto-Marquez et al. (2007) on snout-vent length and pelvic canal width in alligators:

```
data gators;
   input sex $ snoutvent pelvicwidth;
   cards;
male      1.10    7.62
```

====See the web page for the full data set====

```
female    1.23    9.23
;
proc glm data=gators;
   class sex;
   model pelvicwidth=snoutvent sex snoutvent*sex;
proc glm data=gators;
   class sex;
   model pelvicwidth=snoutvent sex;
   run;
```

The first time you run PROC GLM, the MODEL statement includes the interaction term (SNOUTVENT*SEX). This tests whether the slopes of the regression lines are significantly different:

```
                    Type III   Mean
Source         DF      SS     Square  F Value   Pr > F

snoutvent       1   33.949   33.949    88.05   <.0001
sex             1    0.079    0.079     0.21   0.6537
snoutvent*sex   1    0.005    0.005     0.01   0.9101 slope P-value
```

If the P-value of the slopes is significant, you'd be done. In this case it isn't, so you look at the output from the second run of PROC GLM. This time, the MODEL statement doesn't include the interaction term, so the model assumes that the slopes of the regression lines are equal. This P-value tells you whether the Y-intercepts are significantly different:

```
                 Type III   Mean
Source      DF      SS     Square  F Value   Pr > F

snoutvent    1   41.388   41.388   110.76   <.0001
sex          1    2.016    2.016     5.39   0.0267 intercept P-value
```

Power analysis

The form on the web version of this handbook calculates the sample size needed for an ancova, using the method of Borm et al. (2007). It only works for

ancova with two groups, and it assumes each group has the same standard deviation and the same r^2. To use it, enter:

- the effect size, or the difference in Y-intercepts you hope to detect;
- the standard deviation. This is the standard deviation of all the Y values within each group (without controlling for the X variable). For example, in the alligator data above, this would be the standard deviation of pelvic width among males, or the standard deviation of pelvic width among females.
- alpha, or the significance level (usually 0.05);
- power, the probability of rejecting the null hypothesis when the given effect size is the true difference (0.80 and 0.90 are common values);
- the r^2 within groups. For the alligator data, this would be the r^2 of pelvic width vs. snout-vent length among males, or the r^2 among females.

As an example, let's say you want to do a study with an ancova on pelvic width vs. snout-vent length in male and female crocodiles, and since you don't have any preliminary data on crocodiles, you're going to base your sample size calculation on the alligator data. You want to detect a difference in adjusted means of 0.2 cm. The standard deviation of pelvic width in the male alligators is 1.45 and for females is 1.02; taking the average, enter 1.23 for standard deviation. The r^2 in males is 0.774 and for females it's 0.780, so enter the average (0.777) for r^2 in the form. With 0.05 for the alpha and 0.80 for the power, the result is that you'll need 133 male crocodiles and 133 female crocodiles.

Further reading

Sokal and Rohlf, pp. 499-521.

References

Borm, G.F., J. Fransen, and W.A.J.G. Lemmens. 2007. A simple sample size formula for analysis of covariance in randomized clinical trials. J. Clin. Epidem. 60: 1234-1238.

Conover, W.J., and R.L. Iman. Analysis of covariance using the rank transformation. Biometrics 38: 715-724.

Prieto-Marquez, A., P.M. Gignac, and S. Joshi. 2007. Neontological evaluation of pelvic skeletal attributes purported to reflect sex in extinct non-avian archosaurs. J. Vert. Paleontol. 27: 603-609.

Rooney, J., and S.M. Lewis. 2002. Fitness advantage from nuptial gifts in female fireflies. Ecol. Entom. 27: 373-377.

Multiple regression

When to use it

You use multiple regression when you have three or more measurement variables. One of the measurement variables is the dependent (Y) variable. The rest of the variables are the independent (X) variables. The purpose of a multiple regression is to find an equation that best predicts the Y variable as a linear function of the X variables. There is also a"hidden" nominal variable that groups the measurement variables together.

Multiple regression for prediction

One use of multiple regression is prediction or estimation of an unknown Y value corresponding to a set of X values. For example, let's say you're interested in finding suitable habitat to reintroduce the rare beach tiger beetle, *Cicindela dorsalis dorsalis*, which lives on sandy beaches on the Atlantic coast of North America. You've gone to a number of beaches that already have the beetles and measured the density of tiger beetles (the dependent variable) and several biotic and abiotic factors, such as wave exposure, sand particle size, beach steepness, density of amphipods and other prey organisms, etc. Multiple regression would give you an equation that would relate the tiger beetle density to a function of all the other variables. Then if you went to a beach that doesn't have tiger beetles and measured all the independent variables (wave exposure, sand particle size, etc.) you could use the multiple regression equation to predict the density of tiger beetles that could live there if you introduced them.

Multiple regression for understanding causes

A second use of multiple regression is to try to understand the functional relationships between the dependent and independent variables, to try to see what might be causing the variation in the dependent variable. For example, if you did a regression of tiger beetle density on sand particle size by itself, you would probably see a significant relationship. If you did a regression of tiger beetle density on wave exposure by itself, you would probably see a significant relationship. However, sand particle size and wave exposure are correlated; beaches with bigger waves tend to have bigger sand particles. Maybe sand particle size is really important, and the correlation between it and wave exposure is the

only reason for a significant regression between wave exposure and beetle density. Multiple regression is a statistical way to try to control for this; it can answer questions like "If sand particle size (and every other measured variable) were the same, would the regression of beetle density on wave exposure be significant?"

Null hypothesis

The main null hypothesis of a multiple regression is that there is no relationship between the X variables and the Y variables, that the fit of the observed Y values to those predicted by the multiple regression equation is no better than what you would expect by chance. As you are doing a multiple regression, there is also a null hypothesis for each X variable, that adding that X variable to the multiple regression does not improve the fit of the multiple regression equation any more than expected by chance.

How it works

The basic idea is that an equation is found, like this:

$$Y_{exp}=a+b_1X_1+b_2X_2+b_3X_3...$$

The Y_{exp} is the expected value of Y for a given set of X values. b_1 is the estimated slope of a regression of Y on X_1, if all of the other X variables could be kept constant, and so on for b_2, b_3, etc; *a* is the intercept. Values of b_1, etc. (the "partial regression coefficients") and the intercept are found that minimize the squared deviations between the expected and observed values of Y.

How well the equation fits the data is expressed by R^2, the "coefficient of multiple determination." This can range from 0 (for no relationship between the X and Y variables) to 1 (for a perfect fit, no difference between the observed and expected Y values). The P-value is a function of the R^2, the number of observations, and the number of X variables.

When the purpose of multiple regression is prediction, the important result is an equation containing partial regression coefficients. If you had the partial regression coefficients and measured the X variables, you could plug them into the equation and predict the corresponding value of Y. The magnitude of the partial regression coefficient depends on the unit used for each variable, so it does not tell you anything about the relative importance of each variable.

When the purpose of multiple regression is understanding functional relationships, the important result is an equation containing *standard* partial regression coefficients, like this:

$$y'_{exp}=a+b'_1x'_1+b'_2x'_2+b'_3x'_3...$$

where b'_1 is the standard partial regression coefficient of y on X_1. It is the number of standard deviations that Y would change for every one standard deviation change in X_1, if all the other X variables could be kept constant. The magnitude of the standard partial regression coefficients tells you something about the relative importance of different variables; X variables with bigger standard partial regression coefficients have a stronger relationship with the Y variable.

Selecting variables in multiple regression

Every time you add a variable to a multiple regression, the R^2 increases (unless the variable is a simple linear function of one of the other variables, in which case R^2 will stay the same). The best-fitting model is therefore the one that includes all of the X variables. However, whether the purpose of a multiple regression is prediction or understanding functional relationships, it is often useful to decide which are important and unimportant variables. In the tiger beetle example, if your purpose was prediction it would be useful to know that your prediction would be almost as good if you measured only sand particle size and amphipod density, rather than measuring a dozen difficult variables. If your purpose was understanding possible causes, knowing that certain variables did not explain much of the variation in tiger beetle density could suggest that they are probably not important causes of the variation in beetle density.

One way to choose variables, called forward selection, is to do a linear regression for each of the X variables, one at a time, then pick the X variable that had the highest R^2. Next you do a multiple regression with the X variable from step 1 and each of the other X variables. The X variable that increases the R^2 by the greatest amount is added, if the P-value of the increase in R^2 is below the desired cutoff. This procedure continues until adding another X variable does not significantly increase the R^2.

To calculate the P-value of an increase in R^2 when increasing the number of X variables from *d* to *e*, where the total sample size is *n*, use the formula:

$$F_s = \frac{(R^2_e - R^2_d)/(e-d)}{(1-R^2_e)/(n-e-1)}$$

A second technique, called backward elimination, is to start with a multiple regression using all of the X variables, then perform multiple regressions with each X variable removed in turn. The X variable whose removal causes the smallest decrease in R^2 is eliminated. This process continues until removal of any X variable would cause a significant decrease in R^2.

Odd things can happen when using either of the above techniques. You could add variables X_1, X_2, X_3, and X_4, with a significant increase in R^2 at each step, then

find that once you've added X_3 and X_4, you can remove X_1 with little decrease in R^2. It is possible to do multiple regression with independent variables A, B, C, and D, and have forward selection choose variables A and B, and backward elimination choose variables C and D. To avoid this, many people use stepwise multiple regression. After adding each X variable, the effects of removing any of the other X variables is tested. This continues until adding new X variables does not significantly increase R^2 and removing X variables does not significantly decrease it.

Important warning

It is easy to throw a big data set at a multiple regression and get an impressive-looking output. However, many people are skeptical of the usefulness of multiple regression, especially for variable selection, and you should view the results with caution. You should examine the linear regression of the dependent variable on each independent variable, one at a time, examine the linear regressions between each pair of independent variables, and consider what you know about the biology. You should probably treat multiple regression as a way of suggesting patterns in your data, rather than rigorous hypothesis testing.

If independent variables A and B are both correlated with Y, and A and B are highly correlated with each other, only one may contribute significantly to the model, but it would be incorrect to blindly conclude that the variable that was dropped from the model has no biological importance.For example, let's say you did a multiple regression on vertical leap in children five to 12 years old, with height, weight, age and score on a reading test as independent variables. All four independent variables are highly correlated in children, since older children are taller, heavier and read better, so it's possible that once you've added weight and age to the model, there is so little variation left that the effect of height is not significant. It would be biologically silly to conclude that height had no influence on vertical leap. Because reading ability is correlated with age, it's possible that it would contribute significantly to the model; that might suggest some interesting followup experiments on children all of the same age, but it would be unwise to conclude that there was a real effect of reading ability and vertical leap based solely on the multiple regression.

Example

I extracted some data from the Maryland Biological Stream Survey (http://www.dnr.state.md.us/streams/data/index.html) to practice multiple regression on; the data are shown below in the SAS example. The dependent variable is the number of longnose dace (*Rhinichthys cataractae*) per 75-meter section of stream. The independent variables are the area (in acres) drained by the stream; the dissolved oxygen (in mg/liter); the maximum depth (in cm) of the

75-meter segment of stream; nitrate concetration (mg/liter); sulfate concentration (mg/liter); and the water temperature on the sampling date (in degrees C).

One biological goal might be to measure the physical and chemical characteristics of a stream and be able to predict the abundance of longnose dace; another goal might be to generate hypotheses about the causes of variation in longnose dace abundance.

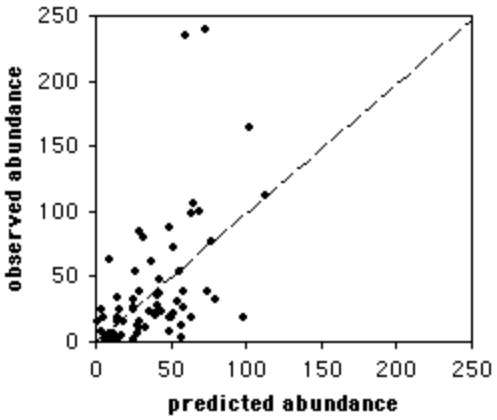
The results of a stepwise multiple regression, with P-to-enter and P-to-leave both equal to 0.15, is that acreage, nitrate, and maximum depth contribute to the multiple regression equation. The R^2 of the model including these three terms is 0.28, which isn't very high.

Graphing the results

If the multiple regression equation ends up with only two independent variables, you might be able to draw a three-dimensional graph of the relationship. Because most humans have a hard time visualizing four or more dimensions, there's no good visual way to summarize all the information in a multiple regression with three or more independent variables. It could be useful to plot a scattergraph with the predicted values on the X-axis and the observed values on the Y-axis. For the longnose dace, I set up a spreadsheet with acreage in column C, maximum depth in column E, and nitrate in column F. Then I put the following equation (in Excel format) in column J, row 2, and copied it into each cell in column J:

```
=0.00199*C2+0.3361*E2+8.67304*F2-23.82907
```

If the multiple regression were perfect, the points would fall on the diagonal dashed line; I made the graph square, with the same scale on the X and Y axis, to emphasize this. The graph makes it easy to see that the multiple regression equation doesn't do a very good job of predicting longnose dace abundance; either other factors that haven't been included in the model are important, or there's a lot of stochasticity in longnose dace abundance.



Observed abundance of longnose dace vs. the abundance predicted from the multiple regression equation.

Similar tests

There are dozens of other multivariate statistical techniques that have been developed, and picking the most appropriate one for your experiment, and interpreting the results, can be difficult. My goal here is mainly to help you understand the results of the most common technique, multiple regression; if you want to actually use multivariate techniques, you're going to have to do a lot of reading in more specialized texts and consult with experts.

How to do multiple regression

Spreadsheet

If you're serious about doing multiple regressions as part of your research, you're going to have to learn a specialized statistical program such as SAS or SPSS. I've written a spreadsheet that will enable you to do a multiple regression with up to 12 X variables and up to 1000 observations. It's fun to play with, but I'm not confident enough in it that I'd recommend using it for publishable results. The spreadsheet includes histograms to help you decide whether to transform your variables, and scattergraphs of the Y variable vs. each X variable so you can see if there are any non-linear relationships. It doesn't do variable selection automatically, you manually choose which variables to include.

Web pages

VassarStat, (http://faculty.vassar.edu/lowry/multU.html) Rweb (http://bayes.math.montana.edu/cgi-bin/Rweb/buildModules.cgi) and AutoFit

(http://www.eskimo.com/~brainy/) are three web pages that are supposed to perform multiple regression, but I haven't been able to get them to work on my computer.

SAS

You use PROC REG to do multiple regression in SAS. Here is an example using the data on longnose dace abundance described above.

```
data fish;
   var stream $ longnosedace acreage do2 maxdepth no3 so4 temp;
   cards;
BASIN_RUN  13  2528  9.6  80  2.28  16.75  15.3
====See the web page for the full data set====
WATTS_BR   19  510  6.7  82  5.25  14.19  26.5
;
proc reg data=fish;
   model longnosedace=acreage do2 maxdepth no3 so4 temp /
     selection=stepwise slentry=0.15 slstay=0.15 details=summary stb;
   run;
```

In the MODEL statement, the dependent variable is to the left of the equals sign, and all the independent variables are to the right. SELECTION determines which variable selection method is used; choices include FORWARD, BACKWARD, STEPWISE, and several others. You can omit the SELECTION parameter if you want to see the multiple regression model that includes all the independent variables. SLENTRY is the significance level for entering a variable into the model, if you're using FORWARD or STEPWISE selection; in this example, a variable must have a P-value less than 0.15 to be entered into the regression model. SLSTAY is the significance level for removing a variable in BACKWARD or STEPWISE selection; in this example, a variable with a P-value greater than 0.15 will be removed from the model. DETAILS=SUMMARY produces a shorter output file; you can omit it to see more details on each step of the variable selection process. The STB option causes the standard partial regression coefficients to be displayed.

Summary of Stepwise Selection

Step	Variable Entered	Variable Removed	Partial R-Square	Model R-Square	C(p)	F-Value	Pr > F
1	acreage		0.1201	0.1201	14.2427	9.01	0.0038
2	no3		0.1193	0.2394	5.6324	10.20	0.0022
3	maxdepth		0.0404	0.2798	4.0370	3.59	0.0625

The summary shows that acreage was added to the model first, yielding an R^2 of 0.1201. Next, no3 was added. The R^2 increased to 0.2394, and the increase in R^2 was significant (P=0.0022). Next, maxdepth was added. The R^2 increased to 0.2798,

which was not quite significant (P=0.0625); SLSTAY was set to 0.15, not 0.05, because you might want to include this variable in a predictive model even if it's not quite significant. None of the other variables increased R^2 enough to have a P-value less than 0.15, and removing any of the variables caused a decrease in R^2 big enough that P was less than 0.15, so the stepwise process is done.

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate
Intercept	1	-23.82907	15.27399	-1.56	0.1237	0
acreage	1	0.00199	0.00067421	2.95	0.0045	0.32581
maxdepth	1	0.33661	0.17757	1.90	0.0625	0.20860
no3	1	8.67304	2.77331	3.13	0.0027	0.33409

The "parameter estimates" are the partial regression coefficients; they show that the model is Y_{exp}=0.00199(acreage)+0.3361(maxdepth)+8.67304(no3)−23.82907. The "standardized estimates" are the standard partial regression coefficients; they show that no3 has the greatest contribution to the model, followed by acreage and then maxdepth. The value of this multiple regression would be that it suggests that the acreage of a stream's watershed is somehow important. Because watershed area wouldn't have any direct effect on the fish in the stream, I would carefully look at the correlations between the acreage and the other independent variables; I would also try to see if there are other variables that were not analyzed that might be both correlated with watershed area and directly important to fish, such as current speed, water clarity, or substrate type.

Further reading

Sokal and Rohlf, pp. 609-631.

Zar, pp. 413-450.

Logistic regression

When to use it

You use **simple logistic regression** when you have one nominal variable with two values (male/female, dead/alive, etc.) and one measurement variable. The nominal variable is the dependent variable, and the measurement variable is the independent variable.

Multiple logistic regression is used when the dependent variable is nominal and there is more than one independent variable. It is analogous to multiple linear regression, and all of the same caveats apply. If you're an epidemiologist, you'll probably need to take a whole course on multiple logistic regression; if you're any other kind of biologist, you'll probably never use it. I won't discuss it any more here; if I say "logistic regression," I'm referring to simple logistic regression.

Simple logistic regression is analogous to linear regression, except that the dependent variable is nominal, not a measurement. One goal is to see whether the probability of getting a particular value of the nominal variable is associated with the measurement variable; the other goal is to predict the probability of getting a particular value of the nominal variable, given the measurement variable.

Data with one nominal and one measurement variable can also be analyzed using a one-way anova or a Student's t-test, and the distinction can be subtle. One clue is that logistic regression allows you to predict the probability of the nominal variable. For example, imagine that you had measured the cholesterol level in the blood of a large number of 55-year-old women, then followed up ten years later to see who had had a heart attack. You could do a t-test, comparing the cholesterol levels of the women who did have heart attacks vs. those who didn't, and that would be a perfectly reasonable way to test the null hypothesis that cholesterol level is not associated with heart attacks; if the hypothesis test was all you were interested in, the t-test would probably be better than the less-familiar logistic regression. However, if you wanted to *predict* the probability that a 55-year-old woman with a particular cholesterol level would have a heart attack in the next ten years, so that doctors could tell their patients "If you reduce your cholesterol by 40 points, you'll reduce your risk of heart attack by X percent," you would have to use logistic regression.

Another situation that calls for logistic regression, rather than an anova or t-test, is when the values of the measurement variable are set by the experimenter,

while the values of the nominal variable are free to vary. For example, let's say you are studying the effect of incubation temperature on sex determination in Komodo dragons. You raise 10 eggs at 30 C, 30 eggs at 32 C, 12 eggs at 34 C, etc., then determine the sex of the hatchlings. It would be silly to compare the mean incubation temperatures between male and female hatchlings, and test the difference using an anova or t-test, because the incubation temperature does not depend on the sex of the offspring; you've set the incubation temperature, and if there is a relationship, it's that the sex of the offspring depends on the temperature.

When there are multiple observations of the nominal variable for each value of the measurement variable, as in the Komodo dragon example, you'll often sees the data analyzed using linear regression, with the proportions treated as a second measurement variable. Often the proportions are arc-sine transformed, because that makes the distributions of proportions more normal. This is not horrible, but it's not strictly correct. One problem is that linear regression treats all of the proportions equally, even if they are based on much different sample sizes. If 6 out of 10 Komodo dragon eggs raised at 30 C were female, and 15 out of 30 eggs raised at 32 C were female, the 60% female at 30 C and 50% at 32 C would get equal weight in a linear regression, which is inappropriate. Logistic regression analyzes each observation (in this example, the sex of each Komodo dragon) separately, so the 30 dragons at 32 C would have 3 times the weight of the 10 dragons at 30 C.

It is also possible to do logistic regression with two nominal variables, but to be honest, I don't see the advantage of this over a chi-squared or G-test of independence.

Null hypothesis

The statistical null hypothesis is that the probability of a particular value of the nominal variable is not associated with the value of the measurement variable; in other words, the line describing the relationship between the measurement variable and the probability of the nominal variable has a slope of zero.

How the test works

Simple logistic regression finds the equation that best predicts the value of the Y variable for each value of the X variable. What makes logistic regression different from linear regression is that the Y variable is not directly measured; it is instead the probability of obtaining a particular value of a nominal variable. If you were studying people who had heart attacks, the values of the nominal variable would be "did have a heart attack" vs. "didn't have a heart attack." The Y variable used in logistic regression would then be the probability of having a heart attack. This probability could take values from 0 to 1. The limited range of this probability would present problems if used directly in a regression, so the odds, Y/(1-Y), is used instead. (If the probability of a heart attack is 0.25, the odds of a heart attack

are 0.25/(1-0.25)=1/3. In gambling terms, this would be expressed as "3 to 1 odds *against* having a heart attack.") Taking the natural log of the odds makes the variable more suitable for a regression, so the result of a logistic regression is an equation that looks like this:

$$\ln[Y/(1-Y)]=a+bX$$

The slope (b) and intercept (a) of the best-fitting equation in a logistic regression are found using the maximum-likelihood method, rather than the least-squares method used for linear regression. Maximum likelihood is a computer-intensive technique; the basic idea is that it finds the values of the parameters under which you would be most likely to get the observed results.

There are several different ways of estimating the P-value. The Wald chi-square is fairly popular, but it may yield inaccurate results with small sample sizes. The likelihood ratio method may be better. It uses the difference between the probability of obtaining the observed results under the logistic model and the probability of obtaining the observed results in a model with no relationship between the independent and dependent variables. I recommend you use the likelihood-ratio method; be sure to specify which method you've used when you report your results.

Examples

McDonald (1985) counted allele frequencies at the mannose-6-phosphate isomerase (Mpi) locus in the amphipod crustacean *Megalorchestia californiana,* which lives on sandy beaches of the Pacific coast of North America. There were two common alleles, Mpi^{90} and Mpi^{100}. The latitude of each collection location, the count of each of the alleles, and the proportion of the Mpi^{100} allele, are shown here:

location	latitude	Mpi90	Mpi100	p, Mpi100
Port Townsend, WA	48.1	47	139	0.748
Neskowin, OR	45.2	177	241	0.577
Siuslaw R., OR	44.0	1087	1183	0.521
Umpqua R., OR	43.7	187	175	0.483
Coos Bay, OR	43.5	397	671	0.628
San Francisco, CA	37.8	40	14	0.259
Carmel, CA	36.6	39	17	0.304
Santa Barbara, CA	34.3	30	0	0.000

Allele (Mpi^{90} or Mpi^{100}) is the nominal variable, location is the hidden nominal variable, and latitude is the measurement variable. If the biological question were "Do different locations have different allele frequencies?", you would ignore latitude and do a chi-square or G-test of independence; here the biological question is "Are allele frequencies associated with latitude?"

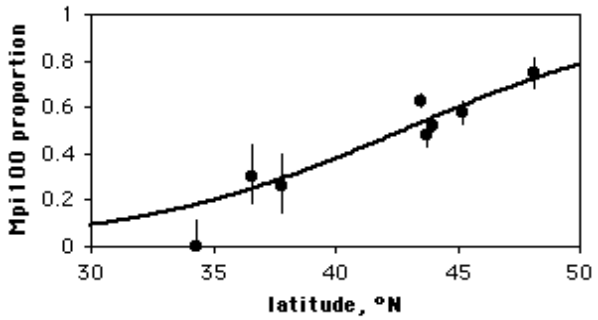
Note that although the proportion of the Mpi^{100} allele seems to increase with increasing latitude, the sample sizes for the northern and southern areas are pretty small. Doing a logistic regression, the result is $chi^2=83.3$, 1 d.f., $P=7\times10^{-20}$. The equation is

$$\ln(Y/(1-Y))=-7.6469+0.1786(\text{latitude}),$$

where Y is the predicted probability of getting an Mpi^{100} allele. Solving this for Y gives

$$Y=e^{-7.6469+0.1786(\text{lat})}/(1+e^{-7.6469+0.1786(\text{lat})}).$$

This logistic regression line is shown on the graph; note that it has a gentle S-shape.



Mpi allele frequencies vs. latitude in the amphipod *Megalorchestia californiana*. Error bars are 95% confidence intervals; the thick black line is the logistic regression line.

Imagine that you have measured antennal stroking speed for some male cucumber beetles. You then present each beetle to a female and see whether mating occurs. Mating would be the nominal variable (mated vs. not mated), and you would do a logistic regression with the probability of mating as the Y variable and anntennal stroking speed as the X variable. The result would tell you whether the stroking speed was significantly associated with the probability of mating.

Graphing the results

If you have multiple observations for each value of the measurement variable, as in the amphipod example above, you can plot a scattergraph with the measurement variable on the X-axis and the proportions on the Y-axis. You might want to put 95% confidence intervals on the points; this gives a visual indication of which points contribute more to the regression (the ones with larger sample sizes have smaller confidence intervals).

There's no automatic way in spreadsheets to add the logistic regression line. Here's how I got it onto the graph of the amphipod data. First, I put the latitudes in column A and the proportions in column B. Then, using the Fill: Series command, I added numbers 30, 30.1, 30.2,…50 to cells A10 through A210. In column C I entered the equation for the logistic regression line; in Excel format, it's

```
=exp(-7.6469+0.1786*(A10))/(1+exp(-7.6469+0.1786*(A10)))
```

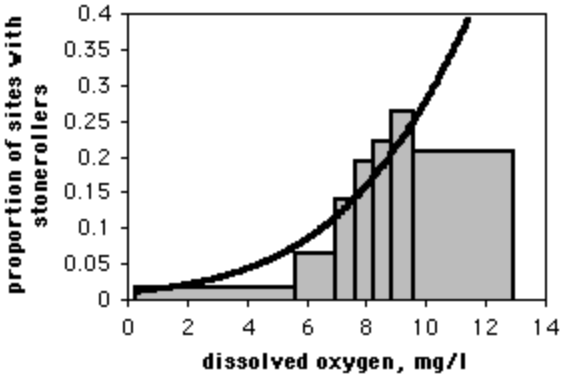
for row 10. I copied this into cells C11 through C210. Then when I drew a graph of the numbers in columns A, B, and C, I gave the numbers in column B symbols but no line, and the numbers in column C got a line but no symbols.

If you only have one observation of the nominal variable for each value of the measurement variable, it would be silly to draw a scattergraph, as each point on the graph would be at either 0 or 1 on the Y-axis. If you have lots of data points, you can divide the measurement values into intervals and plot the proportion for each interval on a bar graph. Here is data from the Maryland Biological Stream Survey (http://www.dnr.state.md.us/streams/mbss/) on 2180 sampling sites in Maryland streams. The measurement variable is dissolved oxygen concentration, and the nominal variable is the presence or absence of the central stoneroller, *Campostoma anomalum*. The intervals are different widths so that each interval includes roughly the same number of stream sites. If you use a bar graph to illustrate a logistic regression, you should explain that the grouping was for heuristic purposes only, and the logistic regression was done on the raw, ungrouped data.

Spreadsheets don't have an option to make bars of different widths, so I graphed these data as a scattergraph. The first bar covers the interval from 0.25 to 5.65, and the proportion is 0.018, so the first four rows on the spreadsheet are:

```
0.25      0
0.25  0.018
5.65  0.018
5.65      0
```

These values are connected with a red line, with no symbols. I got the heavy black line for the logistic regression as described above for the amphipod graph. I used a graphics program to paint the bars gray.



Proportion of streams with central stonerollers vs. dissolved oxygen. Dissolved oxygen intervals were set to have roughly equal numbers of stream sites. The thick black line is the logistic regression line; it is based on the raw data, not the data grouped into intervals.

Similar tests

It is possible to do logistic regression with a dependent variable that has more than two values, known as a multinomial, polytomous, or polychotomous variable. This subject is not covered here.

Multiple logistic regression is used when the dependent variable is nominal and there is more than one independent variable. It is analogous to multiple linear regression, and all of the same caveats apply.

Linear regression is used when the Y variable is a measurement variable. For example, if you measured the length of time it took for male beetles to make and wanted to relate that to stroking speed, you would use linear regression.

When there is just one measurement variable and one nominal variable, one-way anova or a t-test could also be used to compare the means of the measurement variable between the two groups. Conceptually, the difference is whether you think variation in the nominal variable causes variation in the measurement variable (use a t-test) or variation in the measurement variable causes variation in the probability of the nominal variable (use logistic regression). You should also consider who you are presenting your results to, and how they are going to use the information. For example, if you were only interested in stroking speed and mating success in cucumber beetles, you could do a t-test to compare average stroking speed between mated and unmated males. This would be simpler and more familiar than logistic regression; your conclusion would be something like "The mean stroking speed is 73 strokes per minute in mated males and 64 spm in unmated males, a significant difference." Logistic regression is more difficult and

less familiar, but you would be able to express your results with statements such as "A male beetle who strokes a female's antennae at 75 strokes per minute is twice as likely to be accepted by the female as one who strokes 61 strokes per minute." This might be easier to understand, and therefore more useful (especially if you're a male cucumber beetle).

How to do the test

Spreadsheet

I have written a spreadsheet to do simple logistic regression. You can enter the data either in summarized form (for example, saying that at 30 C there were 7 male and 3 female Komodo dragons) or non-summarized form (for example, entering each Komodo dragon separately, with "0" for a male and "1" for a female). It uses the likelihood-ratio method for calculating the P-value. The spreadsheet makes use of the "Solver" tool in Excel. **If you don't see Solver listed in the Tools menu, go to Add-Ins in the Tools menu and install Solver.**

Web page

There is a very nice web page that will do logistic regression, (http://statpages.org/logistic.html) with the likelihood-ratio chi-square. You can enter the data either in summarized form or non-summarized form, with the values separated by tabs (which you'll get if you copy and paste from a spreadsheet) or commas. The amphipod data would be entered like this:

```
48.1,47,139
45.2,177,241
44.0,1087,1183
43.7,187,175
43.5,397,671
37.8,40,14
36.6,39,17
34.3,30,0
```

SAS

Use PROC LOGISTIC for simple logistic regression. There are two forms of the MODEL statement. When you have multiple observations for each value of the measurement variable, your data set can have the measurement variable, the number of "successes" (this can be either value of the nominal variable), and the total. Here is an example using the amphipod data:

```
data amphipods;
   input location $ latitude mpi90 mpi100;
   total=mpi90+mpi100;
   cards;
Port_Townsend,_WA    48.1      47    139
```

```
Neskowin,_OR          45.2     177    241
Siuslaw_R.,_OR        44.0    1087   1183
Umpqua_R.,_OR         43.7     187    175
Coos_Bay,_OR          43.5     397    671
San_Francisco,_CA     37.8      40     14
Carmel,_CA            36.6      39     17
Santa_Barbara,_CA     34.3      30      0
;
proc logistic data=amphipods;
   model mpi100/total=latitude;
run;
```

Note that the new variable TOTAL is created in the DATA step by adding the number of Mpi90 and Mpi100 alleles. The MODEL statement uses the number of Mpi100 alleles out of the total as the dependent variable. The P-value would be the same if you used Mpi90; the equation parameters would be different.

There is a lot of output from PROC LOGISTIC that you don't need. The program gives you three different P-values; the likelihood ratio P-value is the most commonly used:

```
      Testing Global Null Hypothesis: BETA=0

Test               Chi-Square     DF   Pr > ChiSq

Likelihood Ratio      83.3007      1       <.0001 P-value
Score                 80.5733      1       <.0001
Wald                  72.0755      1       <.0001
```

The coefficients of the logistic equation are given under "estimate":

```
              Analysis of Maximum Likelihood Estimates

                            Standard          Wald
Parameter    DF   Estimate      Error   Chi-Square   Pr > ChiSq

Intercept     1    -7.6469     0.9249      68.3605       <.0001
latitude      1     0.1786     0.0210      72.0755       <.0001
```

Using these coefficients, the maximum likelihood equation for the proportion of Mpi100 alleles at a particular latitude is

$$Y=e^{-7.6469+0.1786(\text{latitude})}/(1+e^{-7.6469+0.1786(\text{latitude})})$$

It is also possible to use data in which each line is a single observation. In that case, you may use either words or numbers for the dependent variable. In this example, the data are height (in inches) of the 2004 students of my class, along with their favorite insect (grouped into beetles vs. everything else, where "everything else" includes spiders, which a biologist really should know are not insects):

```
data insect;
   input height insect $;
   cards;
62  beetle
66  other
```
===See the web page for the full data set===
```
74  other
;
proc logistic data=insect;
   model insect=height;
run;
```

The format of the results is the same for either form of the MODEL statement. In this case, the model would be the probability of BEETLE, because it is alphabetically first; to model the probability of OTHER, you would add an EVENT after the nominal variable in the MODEL statement, making it
MODEL INSECT (EVENT='OTHER')=HEIGHT;

Further reading

Sokal and Rohlf, pp. 767-778.

Reference

McDonald, J.H. 1985. Size-related and geographic variation at two enzyme loci in *Megalorchestia californiana* (Amphipoda: Talitridae). Heredity 54: 359-366.

Multiple comparisons

Any time you reject a null hypothesis because a P-value is less than your critical value, you might be wrong; the null hypothesis might really be true, and your significant result might be due to chance. A P-value of 0.05 means that there's a 5 percent chance of getting your observed result, *if* the null hypothesis were true. It does *not* mean that there's a 5 percent chance that the null hypothesis is true.

For example, if you do 200 statistical tests, and for all of them the null hypothesis is actually true, you'd expect 10 of the tests to be significant at the $P<0.05$ level, just due to chance. In that case, you'd have 10 statistically significant results, all of which were false positives. The cost, in time, effort and perhaps money, could be quite high if you based important conclusions on these false positives, and it would at least be embarrassing for you once other people did further research and found that you'd been mistaken.

This problem, that when you do multiple statistical tests, some fraction will be false positives, has received increasing attention in the last few years. This is important for such techniques as the use of microarrays, which make it possible to measure RNA quantities for tens of thousands of genes at once; brain scanning, in which blood flow can be estimated in 100,000 or more three-dimensional bits of brain; and evolutionary genomics, where the sequences of every gene in the genome of multiple species can be compared. There is no universally accepted approach for dealing with the problem of multiple comparisons; it is an area of active research, both in the mathematical details and broader epistomological questions.

Controlling the familywise error rate: Bonferroni correction

The classic approach to the multiple comparison problem is to control the familywise error rate. Instead of setting the critical P-level for significance, or alpha, to 0.05, a lower alpha is used. If the null hypothesis is true for all of the tests, the probability of getting *one* result that is significant at this new, lower alpha level is 0.05. In other words, if the null hypotheses are true, the probability that the family of tests includes one or more false positives due to chance is 0.05.

The most common way to control the familywise error rate is with the Bonferroni correction. The significance level (alpha) for an individual test is found

by dividing the familywise error rate (usually 0.05) by the number of tests. Thus if you are doing 100 statistical tests, the alpha level for an individual test would be 0.05/100=0.0005, and only individual tests with P<0.0005 would be significant.

The Bonferroni correction is appropriate when a single false positive in a set of tests would be a problem. For example, let's say you've developed a new chicken feed, MiracleChick^{TM}, and you're comparing it to traditional chicken feed. You give some chickens the traditional feed and some other chickens the MiracleChick, then you compare the following between the two groups of chickens: food consumption, growth rate, egg production, egg size, feces production, phosphorus content of feces, nitrogen content of feces, meat/bone ratio, white meat/dark meat ratio, and general prettiness. If you see a significant improvement in any of these quantities, you'll start marketing the MiracleChick on that basis, but if your significant result turns out to be a false positive, the farmers are going to sue you. You've got ten statistical tests, so there's a good chance that one will be significant at the 0.05 level, even if MiracleChick is exactly the same as traditional food. Using the Bonferroni correction, you'd require the P-value to be less than 0.005, which would reduce your chance of a false positive (and the resulting angry farmers).

The Bonferroni correction assumes that the tests are independent of each other, as when you are comparing sample A with sample B, C with D, E with F etc. If you are comparing sample A with sample B, A with C, A with D, etc., the comparisons are not independent. This occurs when doing unplanned comparisons of means in anova, for which a variety of other techniques have been developed.

While the Bonferroni correction does a good job of controlling the familywise error rate for multiple, independent comparisons, it may lead to a very high rate of false negatives. For example, if you are comparing the expression levels of 100,000 genes between two kinds of cells, using the Bonferroni correction would mean that a t-test for an individual gene would have to have P<0.0000005 to be considered significant. That could mean that only genes with gigantic differences in expression level would be significant; there might be a lot of genes with real, moderate-sized differences that would be overlooked, all because you wanted to be sure that your results did not include a single false negative.

An important issue with the Bonferroni correction is deciding what a "family" of statistical tests is. If you're testing 12 new chicken feeds, and you measure 10 different quantities on the chickens, is each set of 10 tests for a single chicken feed one "family," so your critical P-value is 0.05/10? Or is the whole set of 10 tests on 12 feeds one family, so your critical P-value is 0.05/120? And what if three months later, you test 5 more chicken feeds--now do you go back and test everything against 0.05/170? There is no firm rule on this; you'll have to use your judgement, based on just how bad a false positive would be. Obviously, you should make this decision before you look at the results, otherwise it would be too easy to unconsiously rationalize a family size that gives you the results you want.

Controlling the false discovery rate: Benjamini–Hochberg procedure

An alternative approach is to control the false discovery rate. This is the proportion of "discoveries" (significant results) that are actually false positives. For example, let's say you're using microarrays to compare expression levels for 100,000 genes between liver tumors and normal liver cells. You're going to do additional experiments on any genes that show a significant difference between the normal and tumor cells, and you're willing to accept up to 10 percent of the genes with significant results being false positives; you'll find out they're false positives when you do the followup experiments. In this case, you would set your false discovery rate to 10 percent.

One good technique for controlling the false discovery rate was briefly mentioned by Simes (1986) and developed in detail by Benjamini and Hochberg (1995). Put the individual P-values in order, from smallest to largest. The smallest P-value has a rank of i=1, the next has i=2, etc. Then compare each individual P-value to (i/m)Q, where *m* is the total number of tests and *Q* is the chosen false discovery rate. The largest P-value that has P<(i/m)Q is significant, and all P-values smaller than it are also significant.

To illustrate this, here are some data on genotype frequencies in the oyster *Crassostrea virginica*. McDonald et al. (1996) compared the genotype frequencies of 6 polymorphisms to the frequencies expected under Hardy-Weinberg equilibrium, using goodness-of-fit tests. There were two population samples, so there were a total of twelve P-values, shown here ordered from smallest to largest. The value of (i/m)Q is shown for a false discovery rate of Q=0.20.

Gene	Location	i	P-value	(i/m)Q
CV7.7	FL	1	0.010	0.017
CVJ5	FL	2	0.032	0.033
CVL1	SC	3	0.07	0.050
CVB2m	SC	4	0.07	0.067
CVB1	FL	5	0.20	0.083
CV7.7	SC	6	0.38	0.100
CVB2e	FL	7	0.48	0.117
CVB2m	FL	8	0.49	0.133
CVB2e	SC	9	0.60	0.150
CVB1	SC	10	0.68	0.167
CVJ5	SC	11	0.74	0.183
CVL1	FL	12	0.97	0.200

Reading down the column of P-values, the largest one with P<(i/m)Q is the second one, CVJ5 in Florida, where the individual P value (0.032) is less than the (i/m)Q value of 0.033. Thus the first two tests would be significant. If you used a Bonferroni correction and set the familywise error rate to 0.05, then each individual P-value would be compared to 0.05/12=0.0042, and none would have been significant.

Other, more complicated techniques, such as Reiner et al. (2003), have been developed for controlling false discovery rate that may be more appropriate when there is lack of independence in the data. If you're using microarrays, in particular, you need to become familiar with this topic.

When not to correct for multiple comparisons

The goal of multiple comparisons corrections is to reduce the number of false positives. An inevitable byproduct of this is that you increase the number of false negatives, where there really is an effect but you don't detect it as statistically significant. If false negatives are very costly, you may not want to correct for multiple comparisons. For example, let's say you've gone to a lot of trouble and expense to knock out your favorite gene, mannose-6-phosphate isomerase (MPI), in a strain of mice that spontaneously develop lots of tumors. Hands trembling with excitement, you get the first $MPI^{-/-}$ mice and start measuring things: blood pressure, growth rate, maze-learning speed, bone density, general prettiness, everything you can think of to measure on a mouse. You measure 50 things on $MPI^{-/-}$ mice and normal mice, run tests, and the smallest P-value is 0.013 for a difference in tumor size. If you use either a Bonferroni correction or the Benjamini and Hochberg procedure, that P=0.013 won't be close to significant. Should you conclude that there's no significant difference between the $MPI^{-/-}$ and $MPI^{+/+}$ mice, write a boring little paper titled "Lack of anything interesting in $MPI^{-/-}$ mice," and look for another project? No, your paper should be "Possible effect of MPI on cancer." You should be suitably cautious, of course, but the cost of a false positive--if further experiments show that MPI really has no effect on tumors--is just a few more experiments. The cost of a false negative, on the other hand, could be that you've missed out on a hugely important discovery.

References

Benjamini, Y., and Y. Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. Roy. Stat. Soc. B. 57: 289-300.

McDonald, J.H., B.C. Verrelli and L.B. Geyer. 1996. Lack of geographic variation in anonymous nuclear polymorphisms in the American oyster, *Crassostrea virginica*. Mol. Biol. Evol. 13: 1114-1118.

Reiner, A., D. Yekutieli and Y. Benjamini. 2003. Identifying differentially expressed genes using false discovery rate controlling procedures. Bioinformatics 19: 368-375.

Simes, R.J. 1986. An improved Bonferroni procedure for multiple tests of significance. Biometrika 73: 751-754.

Meta-analysis

Meta-analysis is a statistical technique for combining the results of different studies to see if the overall effect is significant. This is most often done when there are multiple studies with conflicting results—a drug does or does not work, reducing salt in food does or does not affect blood pressure, that sort of thing. Meta-analysis is a way of combining the results of all the studies; ideally, the result is the same as doing one study with a really big sample size, one large enough to conclusively demonstrate an effect if there is one, or conclusively reject an effect if there isn't one of an appreciable size.

I'm going to outline the general steps involved in doing a meta-analysis, but I'm not going to describe it in sufficient detail that you could do one yourself; if that's what you want to do, see the Further Reading list at the bottom of this page. Instead, I hope to explain some of the things you should look for when reading the results of a meta-analysis.

Decide which studies to include

Before you start collecting studies, it's important to decide which ones you're going to include and which you'll exclude. Your criteria should be as objective as possible; someone else should be able to look at your criteria and then include and exclude the exact same studies that you did. For example, if you're looking at the effects of a drug on a disease, you might decide that only double-blind, placebo-controlled studies are worth looking at, or you might decide that single-blind studies (where the investigator knows who gets the placebo, but the patient doesn't) are acceptable; or you might decide that any study at all on the drug and the disease should be included.

You shouldn't use sample size as a criterion for including or excluding studies. The statistical techniques used for the meta-analysis will give studies with smaller sample sizes the lower weight they deserve.

Finding studies

The next step in a meta-analysis is finding all of the studies on the subject. A critical issue in meta-analysis is what's known as the **file-drawer effect**; people who do a study and fail to find a significant result are less likely to publish it than if they find a significant result. Studies with non-significant results are generally boring; it's difficult to get up the enthusiasm to write them up, and it's difficult to

get them published in decent journals. It's very tempting for someone with a bunch of boring, non-significant data to quietly put it in a file drawer, say "I'll write that up when I get some free time," and then never actually get enough free time.

The reason the file-drawer effect is important to a meta-analysis is that even if there is no real effect, 5% of studies will show a significant result at the $P<0.05$ level; that's what $P<0.05$ means, after all, that there's a 5% probability of getting that result if the null hypothesis is true. So if 100 people did experiments to see whether thinking about long fingernails made your fingernails grow faster, you'd expect 95 of them to find non-significant results. They'd say to themselves, "Well, that didn't work out, maybe I'll write it up for the *Journal of Fingernail Science* someday," then go on to do experiments on whether thinking about long hair made your hair grow longer and never get around to writing up the fingernail results. The 5 people who did find a statistically significant effect of thought on fingernail growth would jump up and down in excitement at their amazing discovery, then get their papers published in *Science* or *Nature*. If you did a meta-analysis on the published results on fingernail thought and fingernail growth, you'd conclude that there was a strong effect, even though the null hypothesis is true.

To limit the file-drawer effect, it's important to do a thorough literature search, including really obscure journals, then try to see if there are unpublished experiments. To find out about unpublished experiments, you could look through summaries of funded grant proposals, which for government agencies such as NIH and NSF are searchable online; look through meeting abstracts in the approporate field; write to the authors of published studies; and send out appeals on e-mail mailing lists.

You can never be 100% sure that you've found every study on your topic ever done. Fortunately, if your meta-analysis shows a significant effect, there are ways to estimate how many unpublished, non-significant studies there would have to be to make the overall effect non-significant. If that number is absurdly large, you can be more confident that your significant meta-analysis is not due to the file-drawer effect.

Extract the information

If the goal of a meta-analysis is to estimate the mean difference between two treatments, you need the means, sample sizes, and a measure of the variation: standard deviation, standard error, or confidence interval. If the goal is to estimate the association between two measurement variables, you need the slope of the regression, the sample size, and the r^2. Hopefully this information is presented in the publication in numerical form. Boring, non-significant results are more likely to be presented in an incomplete form, so you shouldn't be quick to exclude papers from your meta-analysis just because all the necessary information isn't presented

in easy-to-use form in the paper. If it isn't, you might need to write the authors, or measure the size and position of features on published graphs.

Do the meta-analysis

The basic idea of a meta-analysis is that the difference in means, slope of a regression, or other statistic is averaged across the different studies. Experiments with larger sample sizes get more weight, as do experiments with smaller standard deviations or higher r^2 values. It is then possible to test whether this common estimate is significantly different from zero.

Interpret the results

Meta-analysis was invented to be a more objective way of surveying the literature on a subject. A traditional literature survey consists of an expert reading a bunch of papers, dismissing or ignoring those that they don't think are very good, then coming to some conclusion based on what they think are the good papers. The problem with this is that it's easier to see the flaws in papers that disagree with your preconceived ideas about the subject and dismiss them, while deciding that papers that agree with your position are acceptable.

The problem with meta-analysis is that a lot of scientific studies really are crap, and pushing a bunch of little piles of crap together just gives you one big pile of crap. For example, let's say you want to know whether moonlight-energized water cures headaches. You expose some water to moonlight, give little bottles of it to 20 of your friends, and say "Take this the next time you have a headache." You ask them to record the severity of their headache on a 10-point scale, drink the moonlight-energized water, then record the severity of their headache 30 minutes later. This study is crap—any reported improvement could be due to the placebo effect, or headaches naturally getting better with time, or moonlight-energized water curing dehydration just as well as regular water, or your friends lying because they knew you wanted to see improvement. If you include this crappy study in a big meta-analysis of the effects of moonlight-energized water on pain, no amount of sophisticated statistical analysis is going to make its crappiness go away.

You're probably thinking "moonlight-energized water" is another ridiculously absurd thing that I just made up, aren't you? That no one could be stupid enough to believe in such a thing? Unfortunately, there are people that stupid.

The hard work of a meta-analysis is finding all the studies and extracting the necessary information from them, so it's tempting to be impressed by a meta-analysis of a large number of studies. A meta-analysis of 50 studies sounds more impressive than a meta-analysis of 5 studies; it's 10 times as big and represents 10 times as much work, after all. However, you have to ask yourself, "Why do people keep studying the same thing over and over? What motivated someone to do that 50th experiment when it had already been done 49 times before?" Often, the reason for doing that 50th study is that the preceding 49 studies were flawed in some

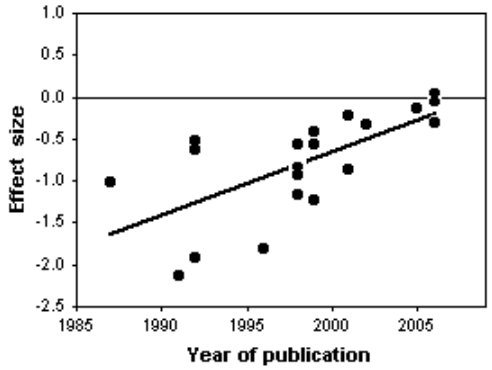
way. If you've got 50 studies, and 5 of them are better by some objective criteria than the other 45, you'd be better off using just the 5 best studies in your meta-analysis.

Example

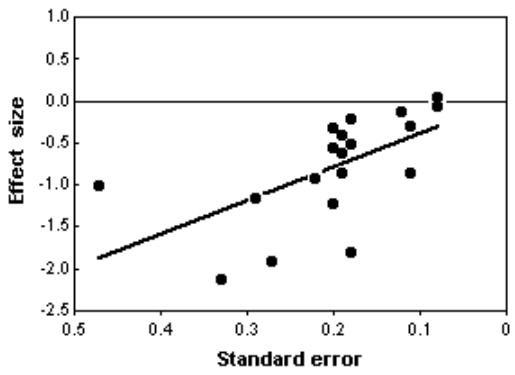
Chondroitin is a polysaccharide derived from cartilage. It is commonly used by people with arthritis in the belief that it will reduce pain, but clinical studies of its effectiveness have yielded conflicting results. Reichenbach et al. (2007) performed a meta-analysis of studies on chondroitin and arthritis pain of the knee and hip. They identified relevant studies by electronically searching literature databases and clinical trial registries, manual searching of conference proceedings and the reference lists of papers, and contacting various experts in the field. Only trials that involved comparing patients given chondroitin with control patients were used; the control could be either a placebo or no treatment. They obtained the necessary information about the amount of pain and the variation by measuring graphs in the papers, if necessary, or by contacting the authors.

The initial literature search yielded 291 potentially relevant reports, but after eliminating those that didn't use controls, those that didn't randomly assign patients to the treatment and control groups, those that used other substances in combination with chondroitin, those for which the necessary information wasn't available, etc., they were left with 20 trials.

The statistical analysis of all 20 trials showed a large, significant effect of chondroitin in reducing arthritis pain. However, the authors noted that earlier studies, published in 1987-2001, had large effects, while more recent studies (which you would hope are better) showed little or no effect of chondroitin. In addition, trials with smaller standard errors (due to larger sample sizes or less variation among patients) showed little or no effect. In the end, Reichenbach et al. (2007) analyzed just the three largest studies with what they considered the best designs, and they showed essentially zero effect of chondroitin. They concluded that there's no good evidence that chondroitin is effective for knee and hip arthritis pain. Other researchers disagree with their conclusion (Goldberg et al. 2007, Pelletier 2007); while a careful meta-analysis is a valuable way to summarize the available information, it is unlikely to provide the last word on a question that has been addressed with large numbers of poorly designed studies.



Effect of chondroitin vs. year of publication of the study. Negative numbers indicate less pain with condroitin than in the control group. The linear regression is significant (r^2=0.45, P=0.001).



Effect of chondroitin vs. standard error of the mean effect size. Negative numbers indicate less pain with condroitin than in the control group. The linear regression is significant (r^2=0.35, P=0.006).

Further reading

Berman, N.G., and R.A. Parker. 2002. (http://www.biomedcentral.com/1471-2288/2/10/) Meta-analysis: neither quick nor easy. BMC Med. Res. Meth. 2:10. [A good readable introduction to medical meta-analysis, with lots of useful references.]

Gurevitch, J., and L.V. Hedges. 2001. Meta-analysis: combining the results of independent experiments. pp. 347-369 in Design and Analysis of Ecological Experiments, S.M. Scheiner and J. Gurevitch, eds. Oxford University Press, New York. [Discusses the use of meta-analysis in ecology, a different perspective than the more common uses of meta-analysis in medical research and the social sciences.]

Hedges, L.V., and I. Olkin. 1985. Statistical methods for meta-analysis. Academic Press, London. [I haven't read this, but apparently this is the classic text on meta-analysis.]

References

Goldberg, H., A. Avins, and S. Bent. 2007. Chondroitin for osteoarthritis of the knee or hip. Ann. Intern. Med. 147: 883.

Pelletier, J.-P. 2007. Chondroitin for osteoarthritis of the knee or hip. Ann. Intern. Med. 147: 883-884.

Reichenbach, S., R. Sterchi, M. Scherer, S. Trelle, E. Bürgi, U. Bürgi, P.A. Dieppe, and P. Jüni. 2007. Meta-analysis: Chondroitin for osteoarthritis of the knee or hip. Ann. Intern. Med. 146: 580-590.

Using spreadsheets for statistics

Real statisticians may sneer, but if you're like most biologists, you can do all of your statistics with spreadsheets. You may spend months getting the most technologically sophisticated new biological techniques to work, but in the end your data can be analzyed with a simple chi-squared test, t-test or linear regression. The graphing abilities of spreadsheets make it easy to inspect data for errors and outliers, look for non-linear relationships and non-normal distributions, and display your final results. Even if you're going to use something like SAS or SPSS, there will be many times when it's easier to enter your data into a spreadsheet first, inspect it for errors, sort and arrange it, then export it into a format suitable for your fancy-schmancy statistics package.

The dominant spreadsheet program by far is Excel, part of the Microsoft Office package, available for Windows or Mac. If your computer already has Excel on it, there's no real advantage to trying anything else. Most of the spreadsheets on these web pages were written with a ten-year-old version of Excel, and the instructions on this web page were based on it; if you're using a newer version and notice that something doesn't work right, please drop me a line and tell me about it.

If your computer doesn't have Excel on it, you could use Calc, part of the free, open-source OpenOffice.org (http://openoffice.org) package. Calc does almost everything that Excel does, with just enough exceptions to be annoying. Calc will open Excel files and can save files in Excel format. The OpenOffice.org package is available for Windows, Mac, and Linux; Mac users may want to use the NeoOffice (http://neooffice.org) port, which looks and feels more like a regular Mac application (but is rather slow). OpenOffice.org also includes a word processor (like Word) and presentation software (like PowerPoint).

For Linux users, Gnumeric sounds like a good, free, open-source spreadsheet program. Because it's a separate program, rather than part of a large office suite, it should be faster than Calc. I haven't used it, so I don't know how well my spreadsheets will work with it.

The instructions on this web page apply to both Calc and Excel, unless otherwise noted.

Basic spreadsheet tasks

I'm going to assume you know how to enter data into a spreadsheet, copy and paste, insert and delete rows and columns, and other simple tasks. If you're a complete beginner, you may want to look at tutorials on using Excel here (http://www.wesleyan.edu/libr/tut/excel/index.html) , here (http://www.ischool.utexas.edu/technology/tutorials/office/excel/) , or here (http://www.usd.edu/trio/tut/excel/) . Here are a few other things that will be useful for handling data:

Separate text into columns

Excel: When you copy columns of data from a web page or text document, then paste them into an Excel spreadsheet, all the data will be in one column. To put the data into multiple columns, choose "Text to columns..." from the Data menu. If you choose "Delimited," you can tell it that the columns are separated by spaces, commas, or some other character. Check the "Treat consecutive delimiters as one" box if numbers may be separated by more than one space, more than one tab, etc. If you choose "Fixed width," you can do things like tell it that the first 10 characters go in column 1, the next 7 characters go in column 2, and so on. The data will be entered into the columns to the right of the original column, so make sure they're empty.

If you paste more text into the same spreadsheet, it will automatically be separated into columns using the same delimiters. If you want to turn this off, select the column where you want to paste the data, choose "Text to columns..." from the Data menu, and choose "Delimited." Then unclick all the boxes for delimiters (spaces, commas, etc.) and click "Finish." Now paste your data into the column.

Calc: In Calc, when you paste columns of data from a text document or web page into a spreadsheet, you'll get a Text Import window that asks you how to divide the text up. Click on the characters that separate the columns (usually spaces, tabs, or commas) and click on "Merge Delimiters" if columns may be separated by more than one space or tab.

Series fill

This is most often used for numbering a bunch of rows. Let's say you have data in cells B1 through E100, and you want to number the rows 1 through 100. Numbering them will help you keep track of which row is which, and it will be especially useful if you want to sort the data, then put them back in their original order. Put a "1" in cell A1, select cells A1-A100, choose "Fill: Series..." from the Edit menu, and you'll put the numbers 1 through 100 in the cells.

Sorting

To sort a bunch of data, select the cells and choose "Sort" from the Data menu. You can sort by up to three columns; for example, you could sort data on a bunch of chickens by "Breed" in column A, "Sex" in column B, and "Weight" in column C, and it would sort the data by breeds, then within each breed have all the females first and then all the males, and within each breed/sex combination the chickens would be listed from smallest to largest.

If you've entered a bunch of data, it's a good idea to sort each column of numbers and look at the smallest and largest values. This may help you spot numbers with misplaced decimal points and other egregious typing errors, as they'll be much larger or much smaller than the correct numbers.

Graphing

See the web pages on graphing with Excel or graphing with Calc. Drawing some quick graphs is another good way to check your data for weirdness. For example, if you've entered the height and leg length of a bunch of people, draw a quick graph with height on the X axis and leg length on the Y axis. The two variables should be pretty tightly correlated, so if you see some outlier who's 2.10 meters tall and has a leg that's only 0.65 meters long, you know to double-check the data for that person.

Absolute and relative cell references

In the formula "=B1+C1", B1 and C1 are relative cell references. If this formula is in cell D1, "B1" means "that cell that is two cells to the left." When you copy cell D1 into cell D2, the formula becomes "=B2+C2"; when you copy it into cell G1, it would become "=E1+F1". This is a great thing about spreadsheets; for example, if you have long columns of numbers in columns A and B and you want to know the sum of each pair, you don't need to type "=Bi+Ci" into each cell in column D, where *i* is the row number; you just type "=B1+C1" once into cell D1, then copy and paste it into all the cells in column D at once.

Sometimes you don't want the cell references to change when you copy a cell; in that case, you should use absolute cell references, indicated with a dollar sign. A dollar sign before the letter means the column won't change, while a dollar sign before the number means the row won't change. If the equation in cell E1 is "=$B1*C$1+D1^2" and you copy it into cell F2, the first term would change from $B1 to $B2 (because you've moved down one row), the second term would change from C$1 to D$1 (because you've moved right one column), and the third term, D1, wouldn't change (because it has dollar signs before both the letter and the number). So if you had 100 numbers in column B, you could enter "=B1-AVERAGE(B$1:B$100)" in cell C1, copy it into cells C2 through C100, and each value in column B would have the average of the 100 numbers subtracted from it.

Paste Special

When a cell has a formula in it (such as "=B1*C1+D1^2"), you see the numerical result of the formula (such as "7.15") in the spreadsheet. If you copy and paste that cell, the formula will be pasted into the new cell; unless the formula only has absolute cell references, it will show a different numerical result. Even if you use only abolute cell references, the result of the formula will change every time you change the values in B1, C1 or D1. When you want to copy and paste the number that results from a function in **Excel**, choose "Paste Special" from the Edit menu and then click the button that says "Values." The number (7.15, in this example) will be pasted into the cell.

In **Calc**, choose "Paste Special" from the Edit menu, uncheck the boxes labelled "Paste All" and "Formulas," and check the box labelled "Numbers."

Change number format

To change the number of decimal places that are displayed in a cell in **Excel**, choose "Cells" from the Format menu, then choose the "Number" tab. Under "Category," choose "Number" and tell it how many decimal places you want to display. Note that this only changes the way the number is displayed; all of the digits are still in the cell, they're just invisible.

The default format in **Excel** ("General" format) automatically uses scientific notation for very small or large numbers. If you've changed the format of a cell to "Number" format with a fixed number of decimal places, very small numbers will be rounded to 0. If you see a 0 in a spreadsheet where you expect a non-zero number (such as a P-value), change the format to General.

The default format in **Calc** is a fixed format with only two digits past the decimal point, and Calc doesn't have a format that automatically uses scientific notation for small numbers, which is annoying. One way to get around this is to create a user-defined format. Select the cells you want to fix, choose "Cells" from the Format menu, and paste the following into the box labelled "Format code":

```
[>0.00001]0.######;[<-0.00001]0.######;0.00E-00
```

The spreadsheets I've created for these web pages use this format for the cells containing P-values and other results. It will display 0 as 0.00E00, but otherwise it works pretty well.

If a column is too narrow to display a number in the specified format, digits to the right of the decimal point will be rounded. If there are too many digits to the left of the decimal point to display them all, the cell will contain "###". Make sure your columns are wide enough to display all your numbers.

Useful spreadsheet functions

There are hundreds of functions in Excel and Calc; here are the ones that I find most useful for statistics and general data handling. Note that where the argument (the part in parentheses) of a function is "*Y*", it means a single number or a single cell in the spreadsheet. Where the argument says "*Ys*", it means more than one number or cell. See AVERAGE(Ys) for an example.

All of the examples here are given in Excel format. Calc uses a semicolon instead of a comma to separate multiple parameters; for example, Excel would use "=ROUND(A1, 2)" to return the value in cell A1 rounded to 2 decimal places, while Calc would use "=ROUND(A1; 2)". If you import an Excel file into Calc or export a Calc file to Excel format, Calc automatically converts between commas and semicolons. However, if you type a formula into Calc with a comma instead of a semicolon, Calc acts like it has no idea what you're talking about; all it says is "#NAME?".

I've typed the function names in all capital letters to make them stand out, but you can use lower case letters.

Math functions

ABS(Y) Returns the absolute value of a number.

EXP(Y) Returns *e* to the *y*th power. This is the inverse of LN, meaning that EXP(LN(Y)) equals *Y*.

LN(Y) Returns the natural logarithm (logarithm to the base *e*) of *Y*.

LOG(Y) Returns the base-10 logarithm of *Y*. The inverse of LOG is raising 10 to the Yth power, meaning 10^(LOG(Y)) returns *Y*.

RAND() Returns a pseudorandom number, equal to or greater than zero and less than one. You must use empty parentheses so the spreadsheet knows that RAND is a function. For a pseudorandom number in some other range, just multiply; thus =RAND()*79 would give you a number greater than or equal to 0 and less than 79. The value will change every time you enter something in any cell. One use of random numbers is for randomly assigning individuals to different treatments; you could enter "=RAND()" next to each individual, Copy and Paste Special the random numbers, Sort the individuals based on the column of random numbers, then assign the first 10 individuals to the placebo, the next 10 individuals to 10 mg of the trial drug, etc.

A "pseudorandom" number is generated by a mathematical function; if you started with the same starting number (the "seed"), you'd get the same series of numbers. Excel's pseudorandom number generator bases its seed on the time given by the computer's internal clock, so you won't get the same seed twice. There are problems with Excel's pseudorandom number generator, but the numbers it produces are random enough for anything you're going to do as a biologist.

ROUND(Y,D) Returns *Y* rounded to *D* digits. For example, =ROUND(37.38, 1) returns 37.4, =ROUND(37.38, 0) returns 37, and =ROUND(37.38, -1) returns 40. Numbers ending in 5 are rounded up (away from zero), so =ROUND(37.35,1) returns 37.4 and =ROUND(-37.35) returns -37.4.

SQRT(Y) Returns the square root of *Y*.

SUM(Ys) Returns the sum of a set of numbers.

Logical functions

AND(logical_test1, logical_test2,...) Returns TRUE if logical_test1, logical_test2... are all true, otherwise returns FALSE. As an example, let's say that cells A1, B1 and C1 all contain numbers, and you want to know whether they're all greater than 100. One way to find out would be with the statement =AND(A1>100, B1>100, C1>100), which would return TRUE if all three were greater than 100 and FALSE if any one were not greater than 100.

IF(logical_test, A, B) Returns *A* if the logical test is true, *B* if it is false. As an example, let's say you have 1000 rows of data in columns A through E, with a unique ID number in column A, and you want to check for duplicates. Sort the data by column A, so if there are any duplicate ID numbers, they'll be adjacent. Then in cell F1, enter "=IF(A1=A2, "duplicate"," "). This will enter the word "duplicate" if the number in A1 equals the number in A2; otherwise, it will enter a blank space. Then copy this into cells F2 through F999. Now you can quickly scan through the rows and see where the duplicates are.

ISNUMBER(Y) Returns TRUE if *Y* is a number, otherwise returns FALSE. This can be useful for identifying cells with missing values. If you want to check the values in cells A1 to A1000 for missing data, you could enter "=IF(ISNUMBER(A1), "OK", "MISSING") into cell B1, copy it into cells B2 to B1000, and then every cell in A1 that didn't contain a number would have "MISSING" next to it in column B.

OR(logical_test1, logical_test2,...) Returns TRUE if one or more of logical_test1, logical_test2... are true, otherwise returns FALSE. As an example, let's say that cells A1, B1 and C1 all contain numbers, and you want to know whether any is greater than 100. One way to find out would be with the statement =OR(A1>100, B1>100, C1>100), which would return TRUE if one or more were greater than 100 and FALSE if all three were not greater than 100.

Statistical functions

AVERAGE(Ys) Returns the arithmetic mean of a set of numbers. For example, AVERAGE(B1..B17) would give the mean of the numbers in cells B1..B17, and AVERAGE(7, A1, B1..C17) would give the mean of 7, the number in cell A1, and the numbers in the cells B1..C17. Note that Excel only counts those cells that have numbers in them; you could enter AVERAGE(A1:A100), put numbers in cells A1 to A9, and Excel would correctly compute the arithmetic mean of those 9 numbers. This is true for other functions that operate on a range of cells.

BINOMDIST(S, K, P, cumulative_probability) Returns the binomial probability of getting *S* "successes" in *K* trials, under the hypothesis that the probability of a success is *P*. The argument "cumulative_probability" should be TRUE if you want the cumulative probability of getting *S* or fewer successes, while it should be FALSE if you want the probability of getting exactly *S* successes. (**Calc** uses 1 and 0 instead of TRUE and FALSE.)

CHIDIST(Y, df) Returns the probability associated with a variable, *Y*, that is chi-square distributed with *df* degrees of freedom. If you use SAS or some other program and it gives the result as "Chi-sq=78.34, 1 d.f., P<0.0001", you can use the CHIDIST function to figure out just how small your P-value is; in this case, "=CHIDIST(78.34, 1)" yields 8.67×10^{-19}.

CONFIDENCE(alpha, standard-deviation, sample-size) Returns the confidence interval of a mean, *assuming you know the population standard deviation.* Because you don't know the population standard deviation, **you should never use this function**; instead, see the web page on confidence intervals for instructions on how to calculate the confidence interval correctly.

COUNT(Ys) Counts the number of cells in a range that contain numbers; if you've entered data into cells A1 through A9, A11, and A17, "=count(A1:A100)" will yield 11.

DEVSQ(Ys) Returns the sum of squares of deviations of data points from the mean. This is what statisticians refer to as the "sum of squares."

FDIST(Y, df1, df2) Returns the probability value associated with a variable, *Y*, that is F-distributed with *df1* degrees of freedom in the numerator and *df2* degrees of freedom in the denominator. If you use SAS or some other program and it gives the result as "F=78.34, 1, 19 d.f., P<0.0001", you can use the FDIST function to figure out just how small your P-value is; in this case, "=FDIST(78.34, 1, 19)" yields 3.62×10^{-8}.

MEDIAN(Ys) Returns the median of a set of numbers. If the sample size is even, this returns the mean of the two middle numbers.

MIN(Ys) Returns the minimum of a set of numbers. Useful for finding the range, which is MAX(Ys)-MIN(Ys).

MAX(Ys) Returns the maximum of a set of numbers.

RANK(X, Ys, type) Returns the rank of *X* in the set of Ys. If *type* is set to 0, the largest number has a rank of 1; if *type* is set to 1, the smallest number has a rank of 0. For example, if cells A1:A8 contain the numbers 10, 12, 14, 14, 16, 17, 20, 21, "=RANK(A2, A1:A8, 0)" returns 7 (the number 12 is the 7th largest in that list), and "=RANK(A2, A1:A8, 1)" returns 2 (it's the 2nd smallest).

Spreadsheets give tied ranks the smallest rank; both of the 14's in the above list would get a rank of 5, as they are tied for 5th largest. The nonparametric tests used in statistics require that ties be given the average rank; both of the 14's in the above list should get a rank of 5.5, the average of 5 and 6, as they are the 5th and 6th largest. This formula shows how to get ranks with ties handled correctly:

```
=AVERAGE(RANK(A1, A1:A8, 0), 1+COUNT(A1:A8)-RANK(A1, A1:A8,
                                1))
```

STDEV(Ys) Returns an estimate of the standard deviation based on a population sample. This is the function you should use for standard deviation.

STDEVP(Ys) Returns the standard deviation of values from an entire population, not just a sample. **You should never use this function**.

SUM(Ys) Returns the sum of the Ys.

SUMSQ(Ys) Returns the sum of the squared values. Note that statisticians use "sum of squares" as a shorthand term for the sum of the squared deviations from the mean. SUMSQ does not give you the sum of squares in this statistical sense; for the statistical sum of squares, use DEVSQ. You will probably never use SUMSQ.

TDIST(Y, df, tails) Returns the probability value associated with a variable, *Y*, that is t-distributed with *df* degrees of freedom and *tails* equal to one or two (you'll almost always want the two-tailed test). If you use SAS or some other program and it gives the result as "t=78.34, 19 d.f., P<0.0001", you can use the TDIST function to figure out just how small your P-value is; in this case, "=TDIST(78.34, 19, 2)" yields 2.56×10^{-25}.

VAR(Ys) Returns an estimate of the variance based on a population sample. This is the function you should use for variance.

VARP(Ys) Returns the variance of values from an entire population, not just a sample. **You should never use this function**.

Guide to fairly good graphs with Excel

Drawing graphs is an important part of analyzing your data and presenting the results of your research. Here I describe the features of clear, effective graphs, and I outline techniques for generating graphs using Excel (there's a similar page on generating good graphs with Calc, part of the free OpenOffice.org (http://www.openoffice.org) suite of programs).

Some of the default conditions for Excel graphs are annoying, but with a little work, you can get it to produce graphs that are good enough for presentations and web pages.

General tips for all graphs

- 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. Both axes of your graph should be clearly labelled, including measurement units if appropriate. Symbols and patterns should be identified in a legend on the graph, or in the caption. If the graph has "error bars," the caption should explain whether they're 95 percent confidence interval, standard error, standard deviation, comparison interval, or something else.
- Don't use color in graphs for publication. If your paper is a success, many people will be reading photocopies or will print it on a black-and-white printer. If the caption of a graph says "Red bars are mean HDL levels for patients taking 2000 mg niacin/day, while blue bars are patients taking the placebo," some of your readers will just see gray bars and will be confused and angry. For bars, use solid black, empty, gray, cross-hatching, vertical stripes, horizontal stripes, etc. Don't use different shades of gray, they may be hard to distinguish in photocopies. There are enough different symbols that you shouldn't need to use colors.
- Do use color in graphs for presentations. It's pretty, and it makes it easier to distinguish different categories of bars or symbols. But don't use red type on a blue background (or vice-versa), as the eye has a hard time focusing

on both colors at once and it creates a distracting 3-D effect. And don't use both red and green bars or symbols on the same graph; from 5 to 10 percent of the males in your audience (and less than 1 percent of the females) have red-green colorblindness and can't distinguish red from green.

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. By far the most common graphs in scientific publications are scatter graphs and bar graphs.

A **scatter graph** (also known as an X-Y graph) is used for graphing data sets consisting of pairs of numbers. These could be measurement variables, or they could be nominal variables summarized as percentages. The independent variable is plotted on the X-axis (the horizontal axis), and the dependent variable is plotted 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, in oceanography it is traditional to 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.

A **bar graph** is used 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.

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. When you look at the points for January 15 and February 15, you connect them with a line (even if there isn't a line on the graph, you mentally connect them), and that implies that on days in between January 15 and February 15, the pollen content was intermediate between the values on those days. 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 the mental connect-the-dots of a scatter graph of these data would imply that the months in between January and February would have intermediate pollen levels, and of course there are no months between January and February.

Drawing scatter graphs with Excel

1. Put your independent variable in one column, with the dependent variable in the column to its right. You can have more than one dependent variable, each in its own column; each will be plotted with a different symbol. Note that if you have a blank cell or a cell containing anything that's not a number in the middle of a column, you'll get a line graph (which you don't want) instead of the scatter graph that you do want. This is one of the stupidest things about graphing with Excel. One way to get around this is to replace the blanks with "NA()".
2. If you are plotting 95 percent confidence intervals, standard errors, or some other kind of error bar, put the values in the next column. These should be confidence intervals, not confidence limits; thus if your first data point has an X-value of 7 and a Y-value of 4 ±1.5, you'd have 7 in the first column, 4 in the second column, and 1.5 in the third column. For confidence limits that are asymmetrical, such as the confidence limits on a binomial percentage, you'll need two columns, one for the difference between the percentage and the lower confidence limit, and one for the difference between the percentage and the upper confidence limit.

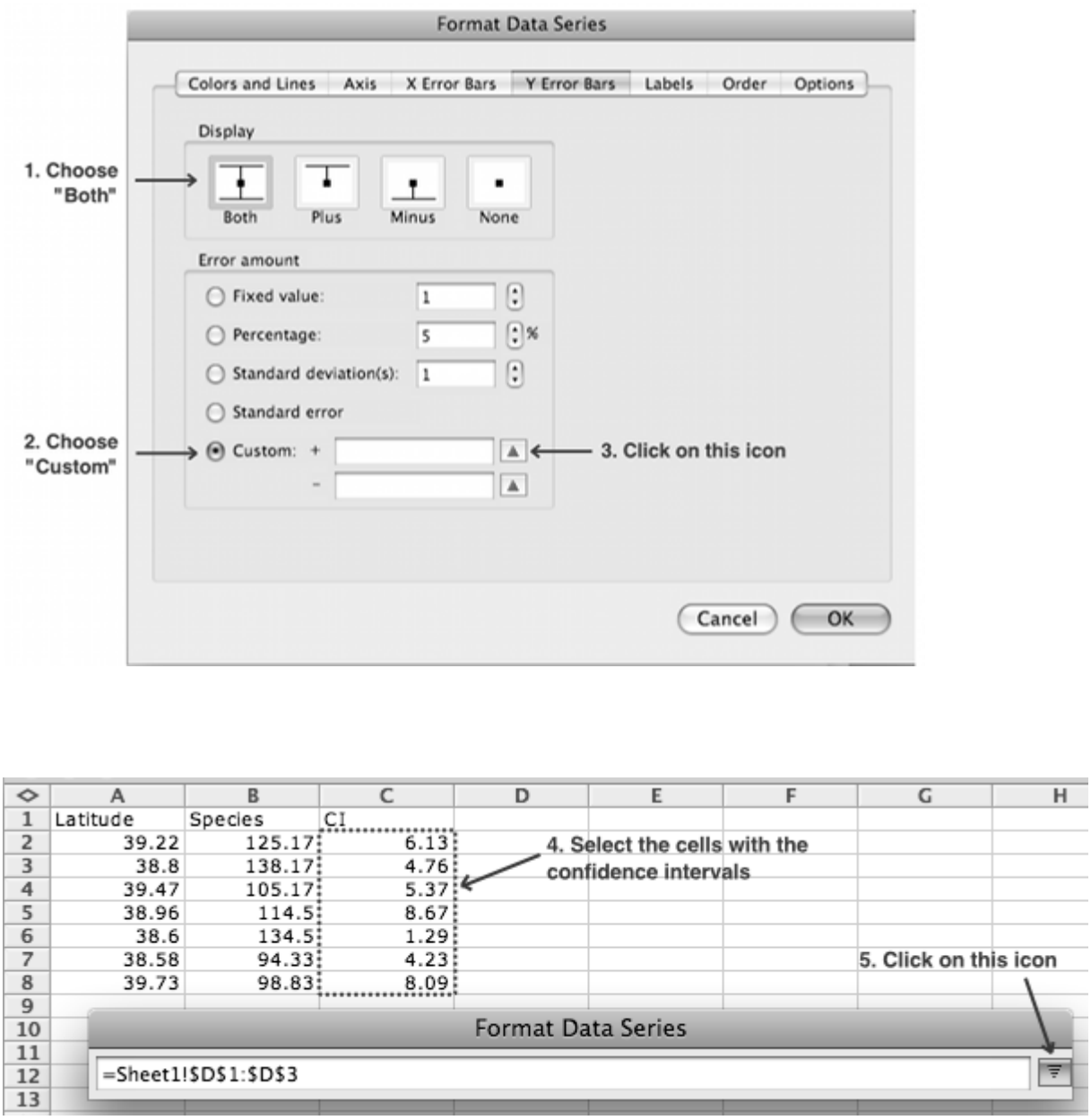
	A	B	C
1	Latitude	Species	CI
2	39.22	125.17	6.13
3	38.8	138.17	4.76
4	39.47	105.17	5.37
5	38.96	114.5	8.67
6	38.6	134.5	1.29
7	38.58	94.33	4.23
8	39.73	98.83	8.09

An Excel spreadsheet set up for a scatter graph including confidence intervals.

3. Select the cells that have the data in them. Don't select the cells that contain the confidence intervals.
4. From the "Insert" menu, choose "Chart" (or click on the little picture of a graph in the task bar). Choose "XY (Scatter)" as your chart type. Do *not* choose "Line"; the little picture may look like a scatter graph, but it isn't.
5. The next screen shows the "Data range," the cells that contain your data; you shouldn't need to change anything here.
6. On the "Titles" tab of the "Chart Options" screen, enter titles for the X axis and Y axis, including the units. A chart title is essential for a graph used in a presentation, but optional in a graph used for a publication (since it will have a detailed caption).
7. On the "Gridlines" tab of the "Chart Options" screen, get rid of the gridlines; they're ugly and unnecessary.
8. On the "Legend" tab of the "Chart Options" screen, get rid of the legend if you only have one set of Y values. If you have more than one set of Y values, get rid of the legend if you're going to explain the different symbols in the figure caption; leave the legend on if you think that's the most effective way to explain the symbols.
9. Click the "Finish" button, but you're far from done. Click on the white area outside the graph to select the whole image, then drag the sides or corners to make the graph the size you want.
10. Click in the gray area inside the graph, choose "Selected Plot Area" from the "Format" menu, and then choose "None" under "Area." This will get rid of the ugly gray background. Under "Border," make the color of the border black instead of gray.
11. Click on the Y-axis, choose "Selected Axis" from the "Format" menu, and make modifications to the tick marks, font and number format. Most publications recommend sans-serif fonts (such as Arial, Geneva, or Helvetica) for figures. On the "Font" tab, unselect "Auto scale," otherwise the font size will change when you change the graph size. On the "Scale"

tab, set the minimum and maximum values of Y. The maximum should be a nice round number, somewhat larger than the highest point on the graph. If you're plotting a binomial percentage, don't make the Y-scale greater than 100 percent. If you're going to be adding error bars, the maximum Y should be high enough to include them. The minimum value on the Y scale should usually be zero, unless your observed values vary over a fairly narrow range. A good rule of thumb (that I just made up, so don't take it too seriously) is that if your maximum observed Y is more than twice as large as your minimum observed Y, your Y scale should go down to zero. If you're plotting multiple graphs of similar data, they should all have the same scales for easier comparison.

12. Format your X-axis the same way you formatted your Y-axis.
13. Click on the Y-axis title, choose "Selected Axis Title" from the "Format" menu, and adjust the font. Unselect "Auto scale" so it won't change the font size if you adjust the size of the graph. Do the same for the X-axis title.
14. Pick one of the symbols, click on it, and choose "Selected Data Series" from the "Format" menu. On the "Patterns" tab, choose the symbol you want (choose "No Color" for the background to get a better view of the symbol choices). If you want to connect the dots with a line, choose that.
15. If you want to add error bars, go to the "Y Error Bars" tab and choose "Both" under "Display." *Ignore* the buttons for "standard deviation" and "standard error," even if they sound like what you want. Instead, choose "Custom." Click on the icon next to the box labelled "+" and then select the cells in your spreadsheet containing the upper confidence interval. Do the same for the box labelled "-" and the lower confidence interval (you'll be selecting the same cells, unless you have asymmetrical confidence intervals).

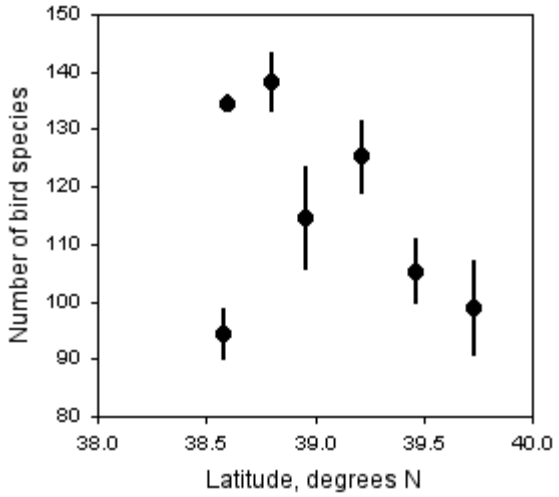


Adding error bars to a graph. Repeat steps 3, 4 and 5 for the box labelled "-".

16. Repeat the above for each set of symbols.
17. If you've added error bars, click on one of them and choose "Selected Error Bars" from the "Format" menu. On the "Patterns" tab, you can adjust the look of the error bars.
18. If you want to add a regression line, click on one of the symbols and choose "Add Trendline" from the "Chart" menu. Choose which kind you want (choose "Linear" unless you really know what you're doing).
19. Click in the graph area, *outside* the graph, to select the whole box that includes the graph and the labels. Choose "Selected Chart Area" from the

"Format" menu. On the "Patterns" tab, you'll probably want to make the border be "None." On the "Properties" tab, choose "Don't move or size with cells," so the graph won't change size if you adjust the column widths of the spreadsheet.

20. You should now have a beautiful, beautiful graph. You can click once on the graph area (in the blank area outside the actual graph), copy it, and paste it into a word processing document, graphics program or presentation.



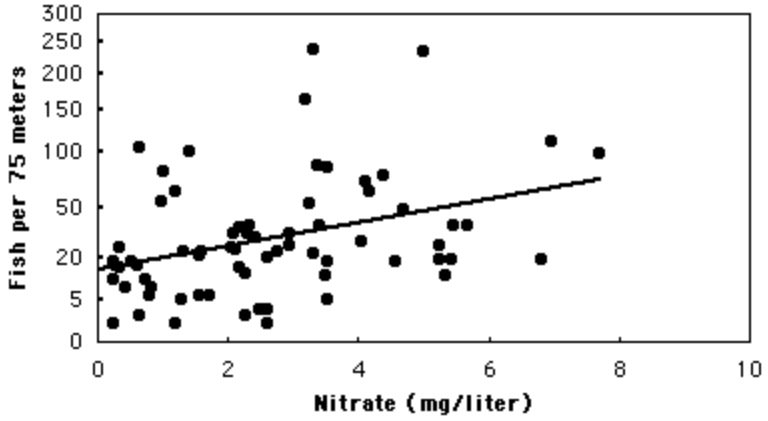
The number of bird species observed in the Christmas Bird Count vs. latitude at seven locations in Delaware. Data points are the mean number of species for the counts in 2001 through 2006, with 95 percent confidence intervals.

Back-transformed axis labels

If you have transformed your data, don't plot the untransformed data; instead, plot the transformed data. For example, if your Y-variable ranges from 1 to 1000 and you've log-transformed it, you would plot the logs on the Y-axis, which would range from 0 to 3 (if you're using base-10 logs). If you square-root transformed those data, you'd plot the square roots, which would range from 1 to about 32. However, you should put the back-transformed numbers (1 to 1000, in this case) on the axes, to keep your readers from having to do squaring or exponentiation in their heads.

I've put together three spreadsheets with graphs that you can use as templates: a spreadsheet graph with log-transformed or square-root transformed X values, a spreadsheet graph with log-transformed or square-root transformed Y values, or a spreadsheet graph with log-transformed or square-root transformed X and Y

values. While they're set up for log-transformed or square-root transformed data, it should be pretty obvious how to modify them for any other transformation.



Abundance of the longnose dace, in number of fish per 75 linear meters of stream, versus nitrate concentration. Fish abundance was square-root transformed for the linear regression.

Drawing bar graphs with Excel

1. Put the values of the independent variable (the nominal variable) in one column, with the dependent variable in the column to its right. The first column will be used to label the bars or clusters of bars. You can have more than one dependent variable, each in its own column; each will be plotted with a different pattern of bar.
2. If you are plotting 95 percent confidence intervals or some other kind of error bar, put the values in the next column. These should be confidence intervals, not confidence limits; thus if your first data point has an X-value of 7 and a Y-value of 4 ±1.5, you'd have 7 in the first column, 4 in the second column, and 1.5 in the third column. For confidence limits that are asymmetrical, such as the confidence limits on a binomial percentage, you'll need two columns, one for the difference between the percentage and the lower confidence limit, and one for the difference between the percentage and the upper confidence limit.

	A	B	C
1	Location	Species	CI
2	Bombay Hook	125.17	6.13
3	Cape Henlopen	138.17	4.76
4	Middletown	105.17	5.37
5	Milford	114.5	8.67
6	Rehoboth	134.5	1.29
7	Seaford-Nanticoke	94.33	4.23
8	Wilmington	98.83	8.09

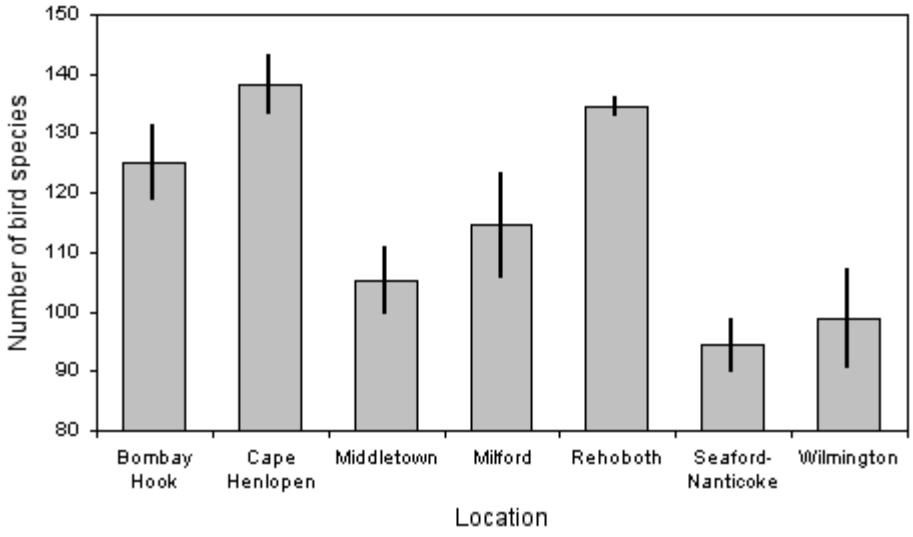
An Excel spreadsheet set up for a bar graph including confidence intervals.

3. Select the cells that have the data in them. Do include the first column, with the values of the nominal variable, but don't select cells that contain confidence intervals.
4. From the "Insert" menu, choose "Chart" (or click on the little picture of a graph in the task bar). Choose "Column" as your chart type, and the picture of bars next to each other (not on top of each other) as the "Chart sub-type." Do not choose the three-dimensional bars, as they just add a bunch of clutter to your graph without conveying any additional information.
5. The next screen shows the "Data range," the cells that contain your data; you shouldn't need to change anything here.
6. On the "Titles" tab of the "Chart Options" screen, enter titles for the X-axis and Y-axis, including the units for the Y-axis. A chart title is essential for a graph used in a presentation, but optional in a graph used for a publication (since it will have a detailed caption). Because each bar or cluster of bars will be labelled on the X-axis, you may not need an X-axis title.
7. On the "Gridlines" tab of the "Chart Options" screen, get rid of the gridlines; they're ugly and unnecessary.
8. On the "Legend" tab of the "Chart Options" screen, get rid of the legend if you only have one set of Y values. If you have more than one set of Y values, get rid of the legend if you're going to explain the different bar patterns in the figure caption; leave the legend on if you think that's the most effective way to explain the bar patterns.
9. Click the "Finish" button, but you're not done yet. Click on the white area outside the graph to select the whole image, then drag the sides or corners to make the graph the size you want.
10. Click in the gray area inside the graph, choose "Selected Plot Area" from the "Format" menu, and then choose "None" under "Area." This will get rid of the ugly gray background. Under "Border," make the color of the border black instead of gray.

11. Click on the Y-axis, choose "Selected Axis" from the "Format" menu, and make modifications to the tick marks, font, and number format. Most publications recommend sans-serif fonts (such as Arial, Geneva, or Helvetica) for figures. On the "Font" tab, unselect "Auto scale," otherwise the font size will change when you change the graph size. On the "Scale" tab, set the minimum and maximum values of Y. The maximum should be a nice round number, somewhat larger than the highest point on the graph. If you're plotting a binomial percentage, don't make the Y-scale greater than 100 percent. If you're going to be adding error bars, the maximum Y should be high enough to include them. The minimum value on the Y scale should usually be zero, unless your observed values vary over a fairly narrow range. A good rule of thumb (that I just made up, so don't take it too seriously) is that if your maximum observed Y is more than twice as large as your minimum observed Y, your Y scale should go down to zero. If you're plotting multiple graphs of similar data, they should all have the same scales for easier comparison.
12. Format your X-axis the same way you formatted your Y-axis. It doesn't have a scale, of course. You may want to get rid of the tick marks, they don't really serve a purpose.
13. Click on the Y-axis title, choose "Selected Axis Title" from the "Format" menu, and adjust the font. Unselect "Auto scale" so it won't change the font size if you adjust the size of the graph. Do the same for the X-axis title.
14. Pick one of the bars, click on it, and choose "Selected Data Series" from the "Format" menu. On the "Patterns" tab, choose the color you want. Click on "Fill effects," then the "Pattern" tab to get halftone grays (little black dots), hatching, and other patterns that work well in black-and-white. On the "Options" tab, you can adjust the width of the bars by changing the "Gap width."
15. If you want to add error bars, go to the "Y Error Bars" tab and choose "Both" under "Display." *Ignore* the buttons for "standard deviation" and "standard error," even if they sound like what you want. Instead, choose "Custom." Click on the icon next to the box labelled "+" and then select the cells in your spreadsheet containing the upper confidence interval. Do the same for the box labelled "-" and the lower confidence interval (you'll be selecting the same cells, unless you have asymmetrical confidence intervals).
16. Repeat the above for each set of bars.
17. If you've added error bars, click on one of them and choose "Selected Error Bars" from the "Format" menu. On the "Patterns" tab, you can adjust the look of the error bars.
18. Click in the graph area, *outside* the graph, to select the whole box that includes the graph and the labels. Choose "Selected Chart Area" from the "Format" menu. On the "Patterns" tab, you'll probably want to make the

border be "None." On the "Properties" tab, choose "Don't move or size with cells," so the graph won't change size if you adjust the column widths of the spreadsheet.

19. You should now have a beautiful, beautiful graph. You can click once on the graph area (in the blank area outside the actual graph), copy it, and paste it into a word processing document, graphics program or presentation.



The number of bird species observed in the Christmas Bird Count at seven locations in Delaware. Data points are the mean number of species for the counts in 2001 through 2006, with 95 percent confidence intervals.

Back-transformed axis labels in bar graphs

If you have transformed your data, don't plot the untransformed data; instead, plot the transformed data. For example, if you've done an anova of log-transformed data, the bars should represent the means of the log-transformed values. Excel has an option to make the X-axis of a bar graph on a log scale, but it's pretty useless, as it only labels the tick marks at 1, 10, 100, 1000…. The only way I know of to get the labels at the right place is to format the axis to not have labels or tick marks, then use the drawing and text tools to put tick marks and labels at the right positions. Get the graph formatted and sized the way you want it, then put in dummy values for the first bar to help you position the tick marks. For example, if you've log-transformed the data and want to have 10, 20, 50, 100, 200, on the Y-axis, give the first bar a value of LOG(10), then use the drawing tools to draw a tick mark even with the top of the bar, then use the text tool to label it "10". Change the dummy value to LOG(20), draw another tick mark, and so on.

Exporting Excel graphs to other formats

Once you've produced a graph, you'll probably want to export it to another program. You may want to put the graph in a presentation (Powerpoint, Keynote, Impress, etc.) or a word processing document. You should be able to click in the graph to select the whole thing, copy it, then paste it into your presentation or word processing document. Sometimes, this will be good enough quality for your purposes.

You'll often want to put the graph in a graphics program, so you can refine the graphics in ways that aren't possible in Excel, or so you can export the graph as a separate graphics file. This is particularly important for publications, where you need each figure to be a separate graphics file in the format and high resolution demanded by the publisher. With earlier versions of Excel, this was easy to do; you could copy a graph, paste it into a drawing program, then "ungroup" the elements and change fonts, symbols, colors, and more. The stupidheads at Microsoft have made newer versions of Excel much worse, so it is much more difficult to make publication-quality graphs with Excel.

The reason it is hard to use Excel for publication-quality graphs is because it produces bitmap images with low (72 pixels per inch) resolution. This looks fine on a computer monitor, but lousy in print. For publications, you need either a bitmap image with much higher resolution (such as 600 dots per inch), or a vector image.

A vector image file stores the drawing as a set of descriptions of different elements. A black dot, 1 mm in diameter, would be described in a vector image file as "draw a circle with the center at 40 mm right and 50 mm up from the lower left corner of the drawing; make the circle have a diameter of 1 mm; and fill in the circle with black." Common examples of vector graphics formats include SVG (Scaleable Vector Graphics), ODG (OpenDocument Graphics), and WMF/EMF (Windows Metafile/Extended Metafile). Commercial drawing programs such as Adobe Illustrator and CorelDraw, and free programs such as Inkscape (http://www.inkscape.org/) and OpenOffice.org Draw, work with vector graphics.

Bitmap image files store an image as a description of each pixel. A black dot might be described as "Put 7 white pixels, then a black pixel, then 7 white pixels; on the next line, put 6 white, 3 black, and 6 white pixels; ..." Common bitmap formats include JPEG, TIFF, GIF, and BMP; commercial programs such as Adobe Photoshop and free programs such as GIMP (http://www.gimp.org) work with bitmap images.

Excel stores graphs in a vector format, but with the new "improved" versions of Excel, it isn't possible to copy the graph in vector format and paste it into a drawing program (at least it doesn't work for me—if you know more about it than I do, please e-mail me). It's also not possible to increase the resolution of the graph within Excel. With older versions of Excel, it was possible to copy a graph, paste it

into a drawing program, ungroup the individual elements and have them in vector format. I don't know exactly when this ability was removed, but if you have an older version of Excel, try copying and pasting a graph from Excel into a drawing program, then zoom in and see if it is a vector or bitmap image. If it is a vector image, don't ever upgrade your version of Excel, as the more recent versions are worse than the one you have.

A 12-point Helvetica letter "e," enlarged 8 times. The letter on the left is from an Excel graph, showing the poor resolution. The letter on the right is from a drawing program (Inkscape).

To make publication-quality graphs with Excel, you have the following options:

- Copy the Excel graph and paste it into a drawing program. Then use the tools in the drawing program to "trace" the graph; make letters, words, symbols, and lines, and put them in the appropriate places over the Excel graph. Once you've redrawn your graph using the drawing program, delete the Excel graph. This is slow and clumsy, but it will give you a useable graph.
- Make the Excel graph much bigger than you want it to be, with bigger fonts and symbols. Copy the graph and paste it into a graphics program. Then shrink the drawing to the size you want. I think this ought to work, but I haven't had much luck with it.
- Create your graph in Calc and copy it into Draw, both programs that are part of OpenOffice.org, (http://www.openoffice.org) a free, open-source suite of programs. It's not as easy to draw good graphs with Calc, but at least you can import them into the drawing program.
- Use a specialized graphing program. Commercial graphing programs include DeltaGraph, KaleidaGraph, and SigmaPlot; I don't know of any good free ones.

Guide to good graphs with Calc

Drawing graphs is an important part of presenting the results of your research. Here I describe the features of clear, effective graphs, and I outline techniques for generating good graphs using Calc, part of the free OpenOffice.org (http://www.openoffice.org) suite of programs. (I've also got a page on good graphs with Excel). Calc can produce graphs suitable for presentations and publication; the biggest deficiency is a very limited selection of symbols. Make sure you're using the latest version of OpenOffice.org, as earlier versions of Calc did not have a way to add error bars to a graph.

General tips for all graphs

- 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. Both axes of your graph should be clearly labelled, including measurement units if appropriate. Symbols and patterns should be identified in a legend on the graph, or in the caption. If the graph has "error bars," the caption should explain whether they're 95 percent confidence interval, standard error, standard deviation, or something else.
- Don't use color in graphs for publication. If your paper is a success, many people will be reading photocopies or will print it on a black-and-white printer. If the caption of a graph says "Red bars are mean HDL levels for patients taking 2000 mg niacin/day, while blue bars are patients taking the placebo," some of your readers will just see gray bars and will be confused and angry. For bars, use solid black, empty, gray, cross-hatching, vertical stripes, horizontal stripes, etc. Don't use different shades of gray, they may be hard to distinguish in photocopies. There are enough different symbols that you shouldn't need to use colors.
- Do use color in graphs for presentations. It's pretty, and it makes it easier to distinguish different categories of bars or symbols. But don't use red type on a blue background (or vice-versa), as the eye has a hard time focusing

on both colors at once and it creates a distracting 3-D effect. And don't use both red and green bars or symbols on the same graph; from 5 to 10 percent of the males in your audience (and less than 1 percent of the females) have red-green colorblindness and can't distinguish red from green.

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. By far the most common graphs in scientific publications are scatter graphs and bar graphs.

A **scatter graph** (also known as an X-Y graph) is used for graphing data sets consisting of pairs of numbers. These could be measurement variables, or they could be nominal variables summarized as percentages. The independent variable is plotted on the x-axis (the horizontal axis), and the dependent variable is plotted 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 makes sense to put the independent variable on the Y-axis. For example, in oceanography it is traditional to 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.

A **bar graph** is used 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.

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 scattergraph 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. When you look at the points for January 15 and February 15, you connect them with a line (even if there isn't a line on the graph, you mentally connect them), and that implies that on days in between January 15 and February 15, the pollen content was intermediate between the values on those days. 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 the mental connect-the-dots of a scatter graph of these data would imply that the months in between January and February would have intermediate pollen levels, and of course there are no months between January and February.

Drawing scatter graphs with Calc

1. Put your independent variable in one column, with the dependent variable in the column to its right. You can have more than one dependent variable, each in its own column; each will be plotted with a different symbol.
2. If you are plotting 95 percent confidence intervals, standard error, or some other kind of error bar, put the values in the next column. These should be confidence intervals, not confidence limits; thus if your first data point has an X-value of 7 and a Y-value of 4±1.5, you'd have 7 in the first column, 4 in the second column, and 1.5 in the third column. For confidence limits that are asymmetrical, such as the confidence limits on a binomial percentage, you'll need two columns, one for the difference between the percentage and the lower confidence limit, and one for the difference between the percentage and the upper confidence limit.

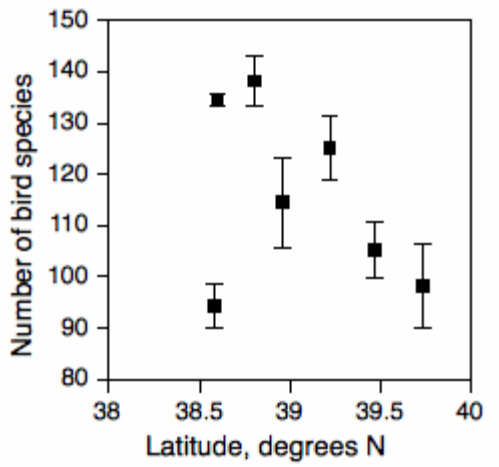
	A	B	C
1	Latitude	Species	CI
2	39.22	125.17	6.13
3	38.8	138.17	4.76
4	39.47	105.17	5.37
5	38.96	114.5	8.67
6	38.6	134.5	1.29
7	38.58	94.33	4.23
8	39.73	98.33	8.09
9			

A Calc spreadsheet set up for a scatter graph including confidence intervals.

3. Select the cells that have the data in them. Don't select the cells that contain the confidence intervals.
4. From the "Insert" menu, choose "Chart" (or click on the little picture of a graph in the task bar). Choose "XY (Scatter)" (the picture of a graph with dots on it) as your chart type. Do *not* choose "Line"; the little picture with lines may look like an XY graph, but it isn't.
5. Click the "Next" button a couple of times. On the "Chart Elements" screen, enter titles for the X axis and Y axis, including the units. A chart title is essential for a graph used in a presentation, but optional in a graph used for a publication (since it will have a detailed caption). Get rid of the legend if you only have one set of Y values. If you have more than one set of Y values, get rid of the legend if you're going to explain the different symbols in the figure caption; leave the legend on if you think that's the most effective way to explain the symbols.
6. Click the "Finish" button, but you're not done yet. Click on the white area outside the graph to select the whole image, then drag the sides or corners to make the graph the size you want.
7. Choose "Chart Wall" from the "Format" menu, and then choose "White" on the "Area" tab. This will get rid of the ugly gray background. Under "Lines," make style "Continuous" and the color "Black," to give you a border around the graph.
8. Choose "Axis" from the "Format" menu, then "Y axis", and make modifications to the tick marks, font and number format. Most publications recommend sans-serif fonts (such as Arial, Geneva, or Helvetica) for figures. On the "Scale" tab, set the minimum and maximum values of Y. The maximum should be a nice round number, somewhat larger than the highest point on the graph. If you're plotting a binomial percentage, don't make the Y-scale greater than 100 percent. If you're adding error bars, the maximum Y should be high enough to include them. The minimum value on the Y scale should usually be zero, unless your observed values vary over a fairly narrow range. A good rule of thumb (that I just made up, so don't take it too seriously) is that if your maximum observed Y is more than twice as large as your minimum observed Y, your Y scale should go down to zero. If you're plotting multiple graphs of similar data, they should all have the same scales for easier comparison.
9. Format your X-axis the same way you formatted your Y-axis.
10. Choose "Title" from the "Format" menu, then "Y axis title", and adjust the font. Do the same for the X-axis title.
11. Pick one of the symbols, click on it, and choose "Object properties" from the "Format" menu. On the "Line" tab, choose the kind of line you want connecting the points, if any. Then choose the symbol under "Icon." Unfortunately, Calc has a very limited number of symbols; there is no circle, for example. (There is a "Gallery" of cartoonish symbols that are

useless for scientific graphs.) As near as I can tell, you can't make the connect-the-dot line a different color from the symbol background (such as a black line connecting open symbols), either (if you know of a way to do this, please let me know, as this is really stupid).

12. If you want a regression line, select a symbol from the data series, then choose "Trend Lines" from the "Insert" menu.
13. Repeat the above for each set of symbols.
14. To add error bars, select a symbol from the data series, then choose "Y error bars" from the "Insert" menu. Under "Error Category", choose "Cell Range." In the box next to "Positive(+)", enter the range of cells containing the confidence intervals. Click the "Same value for both" box if the confidence intervals are symmetrical; if the lower confidence interval is different from the upper, enter its range of cells in the "Negative(-)" box.
    You can change the color and width of the error bars, but unfortunately, the only style you can use is bars with a "T" at each end.
15. Choose "Chart Area" from the "Format" menu. On the "Lines" tab, you'll probably want to make the border be "Invisible."
16. You should now have a fairly good-looking graph. You can click once on the graph area, copy it, and paste it into a word processing document, graphics program or presentation.

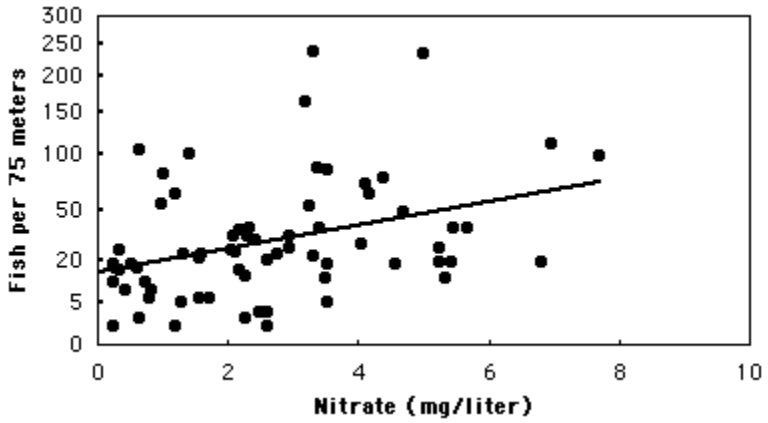


The number of bird species observed in the Christmas Bird Count vs. latitude at seven locations in Delaware. Data points are the mean number of species for the counts in 2001 through 2006, with 95 percent confidence intervals.

Back-transformed axis labels

If you have transformed your data, don't plot the untransformed data; instead, plot the transformed data. For example, if your Y-variable ranges from 1 to 1000 and you've log-transformed it, you would plot the logs on the Y-axis, which would range from 0 to 3 (if you're using base-10 logs). If you square-root transformed those data, you'd plot the square roots, which would range from 1 to about 32. However, you should put the back-transformed numbers (1 to 1000, in this case) on the axes, to keep your readers from having to do squaring or exponentiation in their heads.

I've put together three spreadsheets with graphs that you can use as templates: a spreadsheet graph with log-transformed or square-root transformed X values, a spreadsheet graph with log-transformed or square-root transformed Y values, or a spreadsheet graph with log-transformed or square-root transformed X and Y values. While they're set up for log-transformed or square-root transformed data, it should be pretty obvious how to modify them for any other transformation. Although these graphs put the tick marks in the right places, I couldn't figure out how to label the tick marks automatically in Calc; you'll have to copy the graph to a graphics program and add the tick mark labels there.



Abundance of the longnose dace, in number of fish per 75 linear meters of stream, versus nitrate concentration. Fish abundance was square-root transformed for the linear regression.

Drawing bar graphs using Calc

1. Put the values of the independent variable (the nominal variable) in one column, with the dependent variable in the column to its right. The first column will be used to label the bars or clusters of bars. You can have more than one dependent variable, each in its own column; each will be plotted with a different pattern of bar.

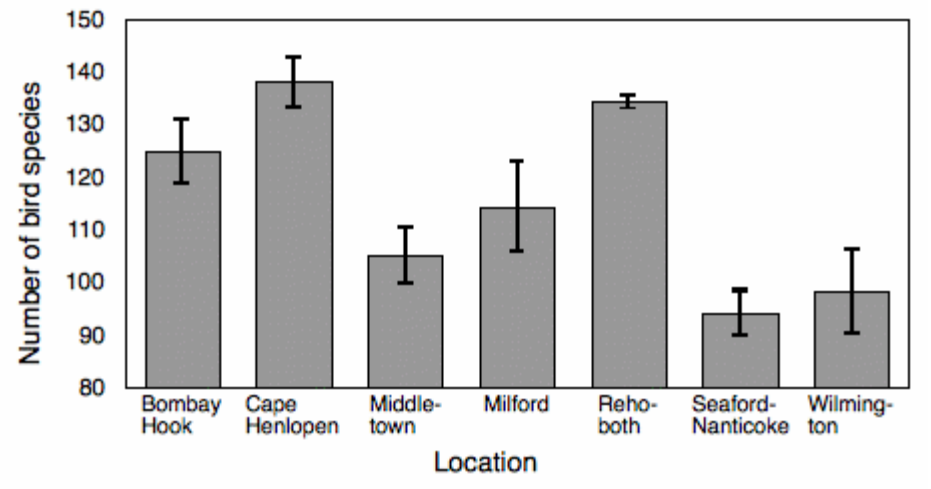
	A	B	C
1	Location	Species	CI
2	Bombay Hook	125.17	6.13
3	Cape Henlopen	138.17	4.76
4	Middle-town	105.17	5.37
5	Milford	114.5	8.67
6	Reho-both	134.5	1.29
7	Seaford-Nanticoke	94.33	4.23
8	Wilmington	98.33	8.09

A Calc spreadsheet set up for a bar graph including confidence intervals.

2. Select the cells that have the data in them, including the first column, with the values of the nominal variable.
3. From the "Insert" menu, choose "Chart" (or click on the little picture of a graph in the task bar). On the "Choose a chart type" screen, choose "Column" and "Normal," the one with columns next to each other.
4. Click the "Next" button a couple of times. On the "Chart Elements" screen, enter titles for the X axis and Y axis, including the units. A chart title is essential for a graph used in a presentation, but optional in a graph used for a publication (since it will have a detailed caption). Get rid of the legend if you only have one set of Y values. If you have more than one set of Y values, get rid of the legend if you're going to explain the different bar patterns in the figure caption; leave the legend on if you think that's the most effective way to explain the patterns.
5. Click the "Finish" button, but you're not done yet. Click on the white area outside the graph to select the whole image, then drag the sides or corners to make the graph the size you want.
6. Choose "Chart Wall" from the "Format" menu, and then choose "White" on the "Area" tab. This will get rid of the ugly gray background. Under "Lines," make style "Continuous" and the color "Black," to give you a border around the graph.
7. Choose "Axis" from the "Format" menu, then "Y axis", and make modifications to the tick marks, font and number format. Most publications recommend sans-serif fonts (such as Arial, Geneva, or Helvetica) for figures. On the "Scale" tab, set the minimum and maximum values of Y. The maximum should be a nice round number, somewhat larger than the highest point on the graph. If you're plotting a binomial percentage, don't make the Y-scale greater than 100 percent. If you're adding error bars, the maximum Y should be high enough to include them. The minimum value on the Y scale should usually be zero, unless your observed values vary

over a fairly narrow range. A good rule of thumb (that I just made up, so don't take it too seriously) is that if your maximum observed Y is more than twice as large as your minimum observed Y, your Y scale should go down to zero. If you're plotting multiple graphs of similar data, they should all have the same scales for easier comparison.

8. Format your X-axis the same way you formatted your Y-axis.
9. Choose "Title" from the "Format" menu, then "Y axis title", and adjust the font. Do the same for the X-axis title.
10. Pick one of the bars, click on it, and choose "Object properties" from the "Format" menu. On the "Borders" tab, choose the kind of border you want for the bars, then choose the pattern inside the bar on the "Area" tab. On the "Options" tab, adjust the width of the bars.
11. Repeat the above for each set of bars.
12. To add error bars, select a symbol from the data series, then choose "Y error bars" from the "Insert" menu. Under "Error Category", choose "Cell Range." In the box next to "Positive(+)", enter the range of cells containing the confidence intervals. Click the "Same value for both" box if the confidence intervals are symmetrical; if the lower confidence interval is different from the upper, enter its range of cells in the "Negative(-)" box.
    You can change the color and width of the error bars, but unfortunately, the only style you can use is bars with a "T" at each end.
13. Choose "Chart Area" from the "Format" menu. On the "Lines" tab, you'll probably want to make the border be "Invisible."
14. You should now have a fairly good looking graph. You can click once on the graph area (in the blank area outside the actual graph), copy it, and paste it into a word processing document, graphics program or presentation.



The number of bird species observed in the Christmas Bird Count at seven locations in Delaware. Data points are the mean number of species for the counts in 2001 through 2006, with 95 percent confidence intervals.

Back-transformed axis labels in bar graphs

If you have transformed your data, don't plot the untransformed data; instead, plot the transformed data. For example, if you've done an anova of log-transformed data, the bars should represent the means of the log-transformed values. Calc has an option to make the X-axis of a bar graph on a log scale, but it's pretty useless, as it only labels the tick marks at 1, 10, 100, 1000…. The only way I know of to get the labels at the right place is to format the axis to not have labels or tick marks, then use the drawing and text tools to put tick marks and labels at the right positions. Get the graph formatted and sized the way you want it, then put in dummy values for the first bar to help you position the tick marks. For example, if you've log-transformed the data and want to have 10, 20, 50, 100, 200, on the Y-axis, give the first bar a value of LOG(10), then use the drawing tools to draw a tick mark even with the top of the bar, then use the text tool to label it "10". Change the dummy value to LOG(20), draw another tick mark, and so on.

Exporting Calc graphs to other formats

Once you've produced a graph, you'll probably want to export it to another program. You may want to put the graph in a presentation (Powerpoint, Keynote, Impress, etc.) or a word processing document. You should be able to click in the graph to select the whole thing, copy it, then paste it into your presentation or word processing document. Sometimes, this will be good enough quality for your purposes.

You'll often want to put the graph in a graphics program, so you can refine the graphics in ways that aren't possible in Calc, or so you can export the graph as a

separate graphics file. This is particularly important for publications, where you need each figure to be a separate graphics file in the format and high resolution demanded by the publisher. This is actually easier to do with Calc than with Excel.

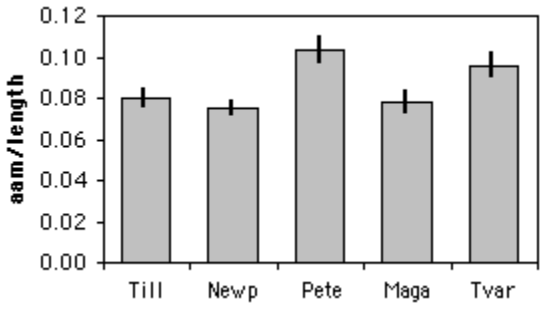
To make publication-quality graphs with Calc, get your graph the way you like it, then copy it and paste it into Draw. In Draw, choose "Break" from the "Modify" menu. You may have to choose "Break" a second time; keep choosing it until the word is grayed out in the menu. At that point, you've broken the graph into its individual elements, and you can modify them individually. You can change the font or color of text, change line thicknesses, change the pattern or color of filled-in rectangles, etc.

Once you have the graph the way you want it, save it. Then export a copy to the file format you want: probably .eps or .tif for publications, .gif for web pages.

Presenting data in tables

Graph or table?

For a presentation, you should almost always use a graph, rather than a table, to present your data. It's easier to compare numbers to each other if they're represented by bars or symbols on a graph, rather than numbers. Here's data from the one-way anova page presented in both a graph and a table:



Length of the anterior adductor muscle scar divided by total length in *Mytilus trossulus*. Means ±one standard error are shown for five locations.

Length of the anterior adductor muscle scar divided by total length in *Mytilus trossulus*. SE: standard error. N: sample size.

Location	Mean AAM/ length	SE	N
Tillamook	0.080	0.0038	10
Newport	0.075	0.0030	8
Petersburg	0.103	0.0061	7
Magadan	0.078	0.0046	8
Tvarminne	0.096	0.0053	6

It's a lot easier to look at the graph and quickly see that the AAM/length ratio is highest at Petersburg and Tvarminne, while the other three locations are lower and about the same as each other. If you put this table in a presentation, you would have to point your laser frantically at one of the 15 numbers and say, "Here! Look at this number!" as your audience's attention slowly drifted away from your science and towards the refreshments table. "Would it be piggish to take a couple of cookies on the way out of the seminar, to eat later?" they'd be thinking. "Mmmmm, cookies...."

In a publication, the choice between a graph and a table is trickier. A graph is still easier to read and understand, but a table provides more detail. Most of your readers will probably be happy with a graph, but a few people who are deeply interested in your results may want more detail than can be shown in a graph. If

anyone is going to do a meta-analysis of your data, for example, they'll want means, sample sizes, and some measure of variation (standard error, standard deviation, or confidence limits). If you've done a bunch of statistical tests and someone wants to reanalyze your data using a correction for multiple comparisons, they'll need the exact P-values, not just stars on a graph indicating significance. Someone who is planning a similar experiment to yours who is doing power analysis will need some measure of variation, as well.

Editors generally won't let you show a graph with the exact same information that you're also presenting in a table. What you can do for many journals, however, is put graphs in the main body of the paper, then put tables as supplemental material. Because these supplemental tables are online-only, you can put as much detail in them as you want; you could even have the individual measurements, not just means, if you thought it might be useful to someone.

Making a good table

Whatever word processor you're using probably has the ability to make good tables. Here are some tips:

- Each column should have a heading. It should include the units, if applicable.
- Don't separate columns with vertical lines. In the olden days of lead type, it was difficult for printers to make good-looking vertical lines; it would be easy now, but most journals still prohibit them.
- When you have a column of numbers, make sure the decimal points are aligned vertically with each other.
- Use a reasonable number of digits. For nominal variables summarized as proportions, use two digits for N less than 101, three digits for N from 101 to 1000, etc. This way, someone can use the proportion and the N and calculate your original numbers. For example, if N is 143 and you give the proportion as 0.22, it could be 31/143 or 32/143; reporting it as 0.217 lets anyone who's interested calculate that it was 31/143. For measurement variables, you should usually report the mean using one more digit than the individual measurement has; for example, if you've measured hip extension to the nearest degree, report the mean to the nearest tenth of a degree. The standard error or other measure of variation should have two or three digits. P-values are usually reported with two digits (P=0.44, P=0.032, $P=2.7\times10^{-5}$, etc.).
- Don't use excessive numbers of horizontal lines. You'll want horizontal lines at the top and bottom of the table, and a line separating the heading from the main body, but that's probably about it. The exception is when you have multiple lines that should be grouped together. If the table of AAM/length ratios above had separate numbers for male and female

mussels at each location, it might be acceptable to separate the locations with horizontal lines.

- Table formats sometimes don't translate well from one computer program to another; if you prepare a beautiful table using a Brand X word processor, then save it in Microsoft Word format or as a pdf to send to your collaborators or submit to a journal, it may not look so beautiful. So don't wait until the last minute; try out any format conversions you'll need, well before your deadline.

Getting started with SAS

SAS, SPSS and Stata are some of the most popular software packages for doing serious statistics. I have a little experience with SAS, so I've prepared this web page to get you started on the basics. UCLA's Academic Technology Services department has prepared very useful guides to SAS (http://www.ats.ucla.edu/stat/sas/) , SPSS (http://www.ats.ucla.edu/stat/spss/) and Stata (http://www.ats.ucla.edu/stat/stata/) .

SAS may seem intimidating and old-fashioned; accomplishing anything with it requires writing what is, in essence, a computer program, one where a misplaced semicolon can have disastrous results. But I think that if you take a deep breath and work your way patiently through the examples, you'll soon be able to do some pretty cool statistics.

The instructions here are for the University of Delaware, but most of it should apply anywhere that SAS is installed. There are three ways of using SAS:

- in batch mode. This is what I recommend, and this is what I'll describe below.
- interactively in line mode. I don't recommend this.
- interactively with the Display Manager System. From what I've seen, this isn't very easy. If you really want to try it, here are instructions (http://www.udel.edu/topics/software/special/statmath/sas/) . Keep in mind that "interactive" doesn't mean "user friendly graphical interface like you're used to"; you still have to write the same SAS programs.

SAS runs on a mainframe computer, not your personal computer. You'll have to connect your personal computer to Strauss, one of the University of Delaware's mainframes. The operating system for Strauss is Unix; in order to run SAS on Strauss in batch mode, you'll have to learn a few Unix commands.

Getting connected to Strauss from a Mac

On a Mac, find the program Terminal; it should be in the Utilities folder, inside your Applications folder. You'll probably want to drag it to your taskbar for easy access in the future. The first time you run Terminal, go to Preferences in the Terminal menu, choose Settings, then choose Advanced. Set "Declare terminal as:"

to "vt100". Then check the box that says "Delete sends Ctrl-H". (Some versions of Terminal may have the preferences arranged somewhat differently, and you may need to look for a box to check that says "Delete key sends backspace.") Then quit and restart Terminal. You won't need to change these settings again.

When you start up Terminal, you'll get a prompt that looks like this:

```
Your-Names-Computer:~ yourname$
```

After the dollar sign, type `ssh userid@strauss.udel.edu`, where `userid` is your UDelNet ID, and hit return. It will ask you for your password; enter it. You'll then be connected to Strauss, and you'll get this prompt:

```
strauss.udel.edu%
```

You're now ready to start typing Unix commands.

Getting connected to Strauss from Windows

On a Windows computer, see if the program SSH Secure Shell is on your computer, and if it isn't, download it from UDeploy (http://udeploy.udel.edu/) . (If you're not at Delaware, ask your site administrator which "terminal emulator" they recommend for Windows). Start up the program, then click on "quick connect" and enter `strauss.udel.edu` for the host name and your UDelNet ID for the username. It will ask you for your password; if you enter it successfully, you'll get this prompt:

```
strauss.udel.edu%
```

You're now ready to start typing Unix commands.

Getting connected to Strauss from Linux

If you're running Linux, you're already enough of a geek that you don't need my help getting connected to the mainframe.

A little bit of Unix

The operating system on Strauss is Unix, so you've got to learn a few Unix commands. Unix was apparently written by people for whom typing is very painful, as most of the commands are a small number of cryptic letters. Case does matter; don't enter `CD` and think it means the same thing as `cd`. Here is all the Unix you need to know to run SAS. Commands are in **bold** and file and directory names, which you choose, are in *italics*.

ls	Lists all of the file names in your current directory.
pico *filename*	pico is a text editor; you'll use it for writing SAS programs. Enter **pico** *practice.sas* to open an existing file named *practice.sas*, or create it if it doesn't exist. To exit pico, enter the **control** and **X** keys. You have to use the arrow keys, not the mouse, to move around the text once you're in a file. For this reason, I prefer to create and edit SAS programs in a text editor on my computer (TextEdit on a Mac, NotePad on Windows), then copy and paste them into a file I've created with pico. I then use pico for minor tweaking of the program. Note that there are other popular text editors, such as vi and emacs, and one of the defining characters of a serious computer geek is a strong opinion about the superiority of their favorite text editor and total loserness of all other text editors. To avoid becoming one of them, try not to get emotional about pico. Unix filenames should be made of letters and numbers, dashes (-), underscores (_), and periods. Don't use spaces or other punctuation (slashes, parentheses, exclamation marks, etc.), as they have special meanings in Unix and may confuse the computer. It is common to use an extension after a period, such as *.sas* to indicate a SAS program, but that it for your convenience in recognizing what kind of file it is; it isn't required by Unix.
cat *filename*	Opens a file for viewing and printing, but not editing. It will automatically take you to the end of the file, so you'll have to scroll up. To print, you may want to copy what you want, then paste it into a word processor document for easier formatting.
mv *oldname newname*	Changes the name of a file from *oldname* to *newname*. When you run SAS on the file *practice.sas*, the output will be in a file called *practice.lst*. Before you make changes to *practice.sas* and run it again, you may want to change the name of *practice.lst* to something else, so it won't be overwritten.
cp *oldname newname*	Makes a copy of file *oldname* with the name *newname*.
rm *filename*	Deletes a file.
logout	Logs you out of Strauss.
mkdir *directoryname*	Creates a new directory. You don't need to do this, but if you end up creating a lot of files, you may find it helpful to keep them organized into different directories.
cd *directoryname*	Changes from one directory to another. For example, if you have a directory named *sasfiles* in your home directory, enter **cd** *sasfiles*. To go from within a directory up to your home directory, just enter **cd**.

rmdir *directoryname*	Deletes a directory, if it doesn't have any files in it. If you want to delete a directory and the files in it, first go into the directory, delete all the files in it using **rm**, then delete the directory using **rmdir**.
sas *filename*	Runs SAS. Be sure to enter **sas** *filename.sas*. If you just enter **sas** and then hit return, you'll be in interactive SAS mode, which is scary; enter **;endsas;** if that happens and you need to get out of it.

Writing a SAS program

To use SAS, you first use pico to create an empty file; you can call the first one *practice.sas*. Then you type in the SAS program that you've written and save the file by hitting the control and X keys. Once you've exited pico, you enter `sas practice.sas`; the word `sas` is the command that tells Unix to run the SAS program, and `practice.sas` is the file it is going to run SAS on. SAS then creates a file named *practice.log*, which reports any errors. If there are no fatal errors, SAS also creates a file named *practice.lst*, which contains the results of the analysis.

The SAS program (which you write using pico) consists of a series of commands. Each command is one or more words, followed by a semicolon. You can put comments into your program to remind you of what you're trying to do; these comments have a slash and asterisk on each side, like this:

```
/*This is a comment. It is not read by the SAS program.*/
```

The SAS program has two basic parts, the DATA step and the PROC step. (Note--I'll capitalize all SAS commands to make them stand out, but you don't have to when you write your programs.) The DATA step reads in data, either from another file or from within the program.

In a DATA step, you first say "DATA dataset;" where *dataset* is an arbitrary name you give the dataset. Then you say "INPUT variable1 variable2...;" giving an arbitrary name to each of the variables that is on a line in your data. So if you have a data set consisting of the length and width of mussels from two different species, you could start the program by writing:

```
data mussels;
   input species $ length width;
```

A variable name for a nominal variable (a name or character) has a space and a dollar sign ($) after it. In our practice data set, "species" is a nominal variable. If you want to treat a number as a nominal variable, such as an ID number, remember to put a dollar sign after the name of the variable. Don't use spaces within variable names; use `Medulis` or `M_edulis`, not `M. edulis` (there are ways of handling variables containing spaces, but they're complicated).

If you are putting the data directly in the program, the next step is a line that says "CARDS;", followed by the data. A semicolon on a line by itself tells SAS it's done reading the data. Each observation is on a separate line, with the variables separated by one or more spaces:

```
data mussel;
   input species $ length width;
   cards;
edulis 49.0 11.0
trossulus 51.2 9.1
trossulus 45.9 9.4
edulis 56.2 13.2
edulis 52.7 10.7
edulis 48.4 10.4
trossulus 47.6 9.5
trossulus 46.2 8.9
trossulus 37.2 7.1
;
```

If you have a large data set, it will be more convenient to keep it in a separate file from your program. To read in data from another file, use the INFILE statement, with the name of the data file in single quotes. In this example, I use the FIRSTOBS option to tell SAS that the first observation is on line 2 of the data file, because line 1 has column headings that remind me what the variables are. You don't have to do this, but I find it's a good idea to have one or more lines of explanatory information at the start of a data file; otherwise, it's too easy to forget what all the numbers are.

```
data mussel;
   infile 'shells.dat' firstobs=2;
   input species $ length width;
```

The DATA statement can create new variables from mathematical operations on the original variables. Here I make two new variables, "loglength," which is just the base-10 log of length, and "shellratio," the width divided by the length. SAS can do statistics on these variables just as it does on the original variables.

```
data mussel;
   infile 'shells.dat' firstobs=2;
   input species $ length width;
   loglength=log10(length);
   shellratio=width/length;
```

The PROC step

Once you've entered in the data, it's time to analyze it. This is done with one or more PROC commands. For example, to calculate the mean and standard

deviation of the lengths, widths, and log-transformed lengths, you would use PROC MEANS:

```
proc means data=mussel mean std;
   var length width loglength;
   run;
```

PROC MEANS tells SAS which procedure to run. It is followed by certain options. DATA=MUSSEL tells it which data set to analyze. MEAN and STD are options that tell PROC MEANS to calculate the mean and standard deviation. On the next line, VAR LENGTH WIDTH LOGLENGTH tells PROC MEANS which variables to analyze. RUN tells it to run.

Now put it all together and run a SAS program. Connect to Strauss and use pico to create a file named "practice.sas". Copy and paste the following into the file:

```
data mussel;
   input species $ length width;
   loglength=log10(length);
   shellratio=width/length;
   cards;
edulis 49.0 11.0
tross 51.2 9.1
tross 45.9 9.4
edulis 56.2 13.2
edulis 52.7 10.7
edulis 48.4 10.4
tross 47.6 9.5
tross 46.2 8.9
tross 37.2 7.1
;
proc means data=mussel mean std;
   var length width loglength;
   run;
```

Then exit pico (hit control-X). At the dollar sign prompt, enter `sas practice.sas`. Then enter `ls` to list the file names; you should see new files named `practice.log` and `practice.lst`. First, enter `cat test.log` to look at the log file. This will tell you whether there are any errors in your SAS program. Then enter `cat practice.lst` to look at the output from your program. You should see something like this:

```
                The SAS System

              The MEANS Procedure

     Variable             Mean         Std Dev
     ----------------------------------------
     length         48.2666667       5.2978769
```

```
width           9.9222222       1.6909892
loglength       1.6811625       0.0501703
----------------------------------------
```

If you do, you've successfully run SAS. Yay!

PROC SORT and PROC PRINT

Specific statistical procedures are described on the web page for each test. Two that are of general use are PROC SORT and PROC PRINT. PROC SORT sorts the data by one or more variables. For some procedures, you need to sort the data first. PROC PRINT writes the data set, including any new variables you've created (like loglength and shellratio in our example) to the output file. You can use it to make sure that SAS has read the data correctly, and your transformations, sorting, etc. have worked properly. You can sort the data by more than one variable; this example sorts the mussel data, first by species, then by length.

```
proc sort data=mussel;
   by species length;
   run;
proc print data=mussel;
   run;
```

Adding PROC SORT and PROC PRINT to the SAS file produces the following output:

```
                    The SAS System

Obs    species    length   width   loglength   shellratio

 1     edulis      48.4     10.4    1.68485      0.21488
 2     edulis      49.0     11.0    1.69020      0.22449
 3     edulis      52.7     10.7    1.72181      0.20304
 4     edulis      56.2     13.2    1.74974      0.23488
 5     trossulus   37.2      7.1    1.57054      0.19086
 6     trossulus   45.9      9.4    1.66181      0.20479
 7     trossulus   46.2      8.9    1.66464      0.19264
 8     trossulus   47.6      9.5    1.67761      0.19958
 9     trossulus   51.2      9.1    1.70927      0.17773
```

As you can see, the data were sorted first by species, then within each species, they were sorted by length.

Graphs in SAS

It's possible to draw graphs with SAS, but I don't find it to be very easy. I recommend you take whatever numbers you need from SAS, put them into a spreadsheet or specialized graphing program, and use that to draw your graphs.

Getting data from a spreadsheet into SAS

I find it easiest to enter my data into a spreadsheet first, even if I'm going to analyze it using SAS. If you try to copy data directly from a spreadsheet into a SAS file, the numbers will be separated by tabs, which SAS will choke on; your log file will say "NOTE: Invalid data in line...". One way to fix this is to copy the data from the spreadsheet into a text editor (TextEdit on a Mac, Notepad on Windows), then do a search-and-replace to change all the tabs to spaces. You can then copy from the text editor and paste into the file you've opened on Strauss with Pico. Another way to get rid of the tabs is to use the Save As... command in the spreadsheet program and save the spreadsheet as Space-delimited Text. After that, you open the file with TextEdit or Notepad, copy it, and paste it into your file on Strauss.

If you're going to keep your data in a separate file from the SAS program and read it using an INFILE statement, you can use the DELIMITER command to tell it that the values are separated by tabs. Here I've made a file named SHELLS.DAT using a spreadsheet, in which the values are separated by tabs (represented as '09'x in SAS):

```
data mussel;
   infile 'shells.dat' delimiter='09'x;
   input species $ length width;
```

If you have data separated by some other character, just put it in single quotation marks, such as DELIMITER='!' for data separated by exclamation marks.

More information about SAS

The user manuals for SAS (http://support.sas.com/onlinedoc/913/docMainpage.jsp) are available online for free, which is nice. Unfortunately, they're in "frames" format, which makes it impossible to link to specific pages, so you won't see links to the appropriate topics in the manual in this handbook.

The UCLA Academic Technology Services has put together an excellent set of examples of how to do the most common statistical tests in SAS, SPSS or Stata (http://www.ats.ucla.edu/stat/mult_pkg/whatstat/default.htm) ; it's a good place to start if you're looking for more information about a particular test.

Choosing a statistical test

This table is designed to help you decide which statistical test or descriptive statistic is appropriate for your experiment. In order to use it, you must be able to identify all the variables in the data set and tell what kind of variables they are.

The "hidden" nominal variable in a regression is the nominal variable that groups together two or more observations; for example, in a regression of height and weight, the hidden nominal variable is the name of each person. Most texts don't count this as a variable, and you don't need to write it down (you could just group the the height and weight numbers by putting them on the same line), so that's why I'm calling it "hidden."

test	nom.	cont.	rank	purpose	notes	example
exact test for goodness-of-fit	1	-	-	test fit of observed frequencies to expected frequencies	used for small sample sizes (less than 1000)	count the number of males and females in a small sample, test fit to expected 1:1 ratio
G-test for goodness-of-fit	1	-	-	test fit of observed frequencies to expected frequencies	used for large sample sizes (greater than 1000)	count the number of red, pink and white flowers in a genetic cross, test fit to expected 1:2:1 ratio
Chi-square test for goodness-of-fit	1	-	-	test fit of observed frequencies to expected frequencies	used for large sample sizes (greater than 1000)	count the number of red, pink and white flowers in a genetic cross, test fit to expected 1:2:1 ratio
Randomization test for goodness-of-fit	1	-	-	test fit of observed frequencies to expected frequencies	used for small sample sizes (less than 1000) with a large number of categories	count the number of offspring in a trihybrid genetic cross, test fit to expected 27:9:9:9:3:3:3:1 ratio

test	nom.	cont.	rank	purpose	notes	example
G-test of independence	2+	-	-	test hypothesis that proportions are the same in different groups	large sample sizes (greater than 1000)	count the number of apoptotic vs. non-apoptotic cells in liver tissue of organic chemists, molecular biologists, and regular people, test the hypothesis that the proportions are the same
Chi-square test of independence	2+	-	-	test hypothesis that proportions are the same in different groups	large sample sizes (greater than 1000)	count the number of apoptotic vs. non-apoptotic cells in liver tissue of organic chemists, molecular biologists, and regular people, test the hypothesis that the proportions are the same
Fisher's exact test	2	-	-	test hypothesis that proportions are the same in different groups	used for small sample sizes (less than 1000)	count the number of left-handed vs. right-handed grad students in Biology and Animal Science, test the hypothesis that the proportions are the same
Randomization test of independence	2	-	-	test hypothesis that proportions are the same in different groups	used for small sample sizes (less than 1000) and large numbers of categories	count the number of cells in each stage of the cell cycle in two different tissues, test the hypothesis that the proportions are the same
Mantel-Haenzel test	3	-	-	test hypothesis that proportions are the same in repeated pairings of two groups	-	count the number of left-handed vs. right-handed grad students in Biology and Animal Science at several universities, test the hypothesis that the proportions are the same; alternate hypothesis is a consistent direction of difference

test	nom.	cont.	rank	purpose	notes	example
arithmetic mean	-	1	-	description of central tendency of data	-	-
median	-	1	-	description of central tendency of data	more useful than mean for very skewed data	median height of trees in forest, if most trees are short seedlings and the mean would be skewed by the few very tall trees
range	-	1	-	description of dispersion of data	used more in everyday life than in scientific statistics	-
variance	-	1	-	description of dispersion of data	forms the basis of many statistical tests; in squared units, so not very understandable	-
standard deviation	-	1	-	description of dispersion of data	in same units as original data, so more understandable than variance	-
standard error of the mean	-	1	-	description of accuracy of an estimate of a mean	-	-
confidence interval	-	1	-	description of accuracy of an estimate of a mean	-	-

test	nom.	cont.	rank	purpose	notes	example
one-way anova, model I	1	1	-	test the hypothesis that the mean values of the continuous variable are the same in different groups	model I: the nominal variable is meaningful, differences among groups are interesting	compare mean heavy metal content in mussels from Nova Scotia, Maine, Massachusetts, Connecticut, New York and New Jersey, to see whether there is variation in the level of pollution
one-way anova, model II	1	1	-	estimate the proportion of variance in the continuous variable "explained" by the nominal variable	model II: the nominal variable is somewhat arbitrary, partitioning variance is more interesting than determining which groups are different	compare mean heavy metal content in mussels from five different families raised under common conditions, to see if there is heritable variation in heavy metal uptake
sequential Dunn-Sidak method	1	1	-	after a significant one-way model I anova, test the homogeneity of means of planned, non-orthogonal comparisons of groups	-	compare mean heavy metal content in mussels from Nova Scotia+Maine vs. Massachusetts+Connecticut, also Nova Scotia vs. Massachusetts+Connecticut+ New York
Gabriel's comparison intervals	1	1	-	after a significant one-way model I anova, test for significant differences between all pairs of groups	-	compare mean heavy metal content in mussels from Nova Scotia vs. Maine, Nova Scotia vs. Massachusetts, Maine vs. Massachusetts, etc.
Tukey-Kramer method	1	1	-	after a significant one-way model I anova, test for significant differences between all pairs of groups	-	compare mean heavy metal content in mussels from Nova Scotia vs. Maine, Nova Scotia vs. Massachusetts, Maine vs. Massachusetts, etc.
Bartlett's test	1	1	-	test the hypothesis that the variance of a continous variable is the same in different groups	usually used to see whether data fit one of the assumptions of an anova	-

test	nom.	cont.	rank	purpose	notes	example
nested anova	2+	1	-	test hypothesis that the mean values of the continous variable are the same in different groups, when each group is divided into subgroups	subgroups must be arbitrary (model II)	compare mean heavy metal content in mussels from Nova Scotia, Maine, Massachusetts, Connecticut, New York and New Jersey; several mussels from each location, with several metal measurements from each mussel
two-way anova	2	1	-	test the hypothesis that different groups, classified two ways, have the same means of the continuous variable	-	compare cholesterol levels in blood of male vegetarians, female vegetarians, male carnivores, and female carnivores
paired t-test	2	1	-	test the hypothesis that the means of the continuous variable are the same in paired data	-	compare the cholesterol level in blood of people before vs. after switching to a vegetarian diet
linear regression	-	2	-	see whether variation in an independent variable causes some of the variation in a dependent variable; estimate the value of one unmeasured variable corresponding to a measured variable	-	measure chirping speed in crickets at different temperatures, test whether variation in temperature causes variation in chirping speed; or use the estimated relationship to estimate temperature from chirping speed when no thermometer is available
correlation	-	2	-	see whether two variables covary	-	measure salt intake and fat intake in different people's diets, to see if people who eat a lot of fat also eat a lot of salt
multiple regression	-	3+	-	fit an equation relating several X variables to a single Y variable	-	measure air temperature, humidity, body mass, leg length, see how they relate to chirping speed in crickets
polynomial regression	-	2	-	test the hypothesis that an equation with X^2, X^3, etc. fits the Y variable significantly better than a linear regression	-	measure running speed in humans aged 5 to 95
analysis of covariance	1	2	-	test the hypothesis that different groups have the same regression lines	first step is to test the homogeneity of slopes; if they are not significantly different, the homogeneity of the Y-intercepts is tested	measure chirping speed vs. temperature in four species of crickets, see if there is significant variation among the species in the slope or y-intercept of the relationships

test	nom.	cont.	rank	purpose	notes	example
sign test	2	-	1	test randomness of direction of difference in paired data	often used as a non-parametric alternative to a paired t-test	compare the cholesterol level in blood of people before vs. after switching to a vegetarian diet, only record whether it is higher or lower after the switch
Kruskal–Wallis test	1	-	1	test the hypothesis that rankings are the same in different groups	often used as a non-parametric alternative to one-way anova	40 ears of corn (8 from each of 5 varieties) are judged for tastiness, and the mean rank is compared among varieties
Spearman rank correlation	-	-	2	see whether the ranks of two variables covary	often used as a non-parametric alternative to regression or correlation	ears of corn are ranked for tastiness and prettiness, see whether prettier corn is also tastier