

HANDBOOK OF BIOLOGICAL STATISTICS

T H I R D E D I T I O N

JOHN H. McDONALD

University of Delaware

SPARKY HOUSE PUBLISHING

Baltimore, Maryland, U.S.A.

©2014 by John H. McDonald

Non-commercial reproduction of this content, with attribution, is permitted;
for-profit reproduction without permission is prohibited.
See <http://www.biostathandbook.com/permissions.html> for details.

Contents

Basics

Introduction.....	1
Step-by-step analysis of biological data	3
Types of biological variables.....	6
Probability.....	14
Basic concepts of hypothesis testing	16
Confounding variables	24

Tests for nominal variables

Exact test of goodness-of-fit	29
Power analysis	40
Chi-square test of goodness-of-fit	45
G-test of goodness-of-fit.....	53
Chi-square test of independence	59
G-test of independence	68
Fisher's exact test of independence.....	77
Small numbers in chi-square and G-tests.....	86
Repeated G-tests of goodness-of-fit	90
Cochran–Mantel–Haenszel test for repeated tests of independence	94

Descriptive statistics

Statistics of central tendency	101
Statistics of dispersion.....	107
Standard error of the mean	111
Confidence limits.....	115

Tests for one measurement variable

Student's t-test for one sample	121
Student's t-test for two samples.....	126
Independence	131
Normality.....	133
Homoscedasticity and heteroscedasticity	137
Data transformations.....	140
One-way anova	145
Kruskal–Wallis test.....	157
Nested anova.....	165
Two-way anova.....	173
Paired t-test	180
Wilcoxon signed-rank test.....	186

Regressions

Correlation and linear regression.....	190
Spearman rank correlation.....	209
Curvilinear regression	213
Analysis of covariance	220
Multiple regression	229
Simple logistic regression.....	238
Multiple logistic regression.....	247

Multiple tests

Multiple comparisons	254
Meta-analysis	261

Miscellany

Using spreadsheets for statistics	266
Guide to fairly good graphs.....	274
Presenting data in tables.....	283
Getting started with SAS	285
Choosing a statistical test	293

Introduction

Welcome to the Third Edition of the *Handbook of Biological Statistics*! This 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. In my class and in this textbook, 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 biologist 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. But I hope that biologists in many fields will find this to be a useful introduction to statistics.

I have provided a spreadsheet to perform many of the statistical tests. Each comes with sample data already entered; just download the spreadsheet, 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 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 (www.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 (www.biostathandbook.com), you can also buy a spiral-bound, printed copy of the whole handbook for \$18 plus shipping at

www.lulu.com/content/paperback-book/handbook-of-biological-statistics/3862228
I've used this print-on-demand service as a convenience to you, not as a money-making scheme, so please don't feel obligated to buy one. You can also download a free pdf of the whole book from www.biostathandbook.com/HandbookBioStatThird.pdf, in case you'd like to print it yourself or view it on an e-reader.

If you use this handbook and want to cite it in a publication, please cite it as:

McDonald, J.H. 2014. Handbook of Biological Statistics, 3rd ed. Sparky House Publishing, Baltimore, Maryland.

It's better to cite the print version, rather than the web pages, so that people of the future can see exactly what were citing. If you just cite a web page, it might be quite different by the time someone looks at it a few years from now. If you need to see what someone has cited from an earlier edition, you can download pdfs of the first edition (www.biostathandbook.com/HandbookBioStatFirst.pdf) or the second edition (www.biostathandbook.com/HandbookBioStatSecond.pdf).

I am constantly trying to improve this textbook. If you find errors, broken links, typos, or have other 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 all areas of 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!).

Acknowledgments

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 to the students in my Biological Data Analysis class for helping me learn how to explain statistical concepts to biologists; to the many people from around the world who have e-mailed me with questions, comments and corrections about the previous editions of the Handbook; to my patient wife, Beverly Wolpert, for being so patient while I obsessed over writing this; and to my dad, Howard McDonald, for inspiring me to get away from the computer and go outside once in a while.

Step-by-step analysis of biological data

Here I describe how you should determine the best way to analyze your biological experiment.

How to determine the appropriate statistical test

I find that a systematic, step-by-step approach is the best way to decide how to analyze biological data. I recommend that you follow these steps:

1. Specify the biological question you are asking.
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 kinds 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 (primarily normality and homoscedasticity for tests of measurement variables). If it doesn't, choose a more appropriate test.
11. Apply the statistical test you chose, and interpret the results.
12. Communicate your results effectively, usually with a graph or table.

As you work your way through this textbook, you'll learn about the different parts of this process. One important point for you to remember: "do the experiment" is step 9, *not* step 1. You should do a lot of thinking, planning, and decision-making *before* you do an experiment. If you do this, you'll have an experiment that is easy to understand, easy to analyze and interpret, answers the questions you're trying to answer, and is neither too big nor too small. If you just slap together an experiment without thinking about how you're going to do the statistics, you may end up needing more complicated and obscure statistical tests, getting results that are difficult to interpret and explain to others, and

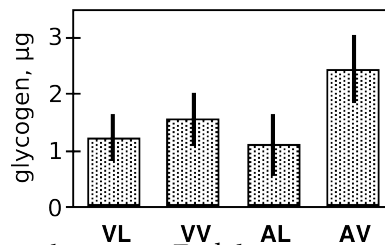
maybe using too many subjects (thus wasting your resources) or too few subjects (thus wasting the whole experiment).

Here's an example of how the procedure 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, often in the form "Does changing X cause a change in Y?" You might want to know whether a drug changes blood pressure; whether soil pH affects the growth of blueberry bushes; or whether protein Rab10 mediates membrane transport to cilia.
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." By thinking about the biological null and alternative hypotheses, you are making sure that your experiment will give different results for different answers to your biological question.
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. Testing your statistical null hypothesis is the main subject of this handbook, and it should give you a clear answer; you will either reject or accept that statistical null. Whether rejecting a statistical null hypothesis is enough evidence to answer your biological question can be a more difficult, more subjective decision; there may be other possible explanations for your results, and you as an expert in your specialized area of biology will have to consider how plausible they are.
4. The two relevant variables in the Verrelli and Eanes experiment are glycogen content and PGM sequence.
5. Glycogen content is a measurement variable, something that you record 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 you usually record 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). It also would have been possible to observe the confounding variables; for example, Verrelli and Eanes could have used flies of different ages, and then used a statistical technique that adjusted for the age. This would have made the analysis more complicated to perform and more difficult to explain, and while it might have turned up something interesting about age and glycogen content, it would not have helped address the main biological question about PGM genotype and glycogen content.
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 categories,

the appropriate statistical test is a one-way anova. Once you know what variables you're analyzing and what type they are, the number of possible statistical tests is usually limited to one or two (at least for tests I present in this handbook).

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 95% confidence intervals.

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 web page, 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 biological variables

There are three main types of variables: measurement variables, which are expressed as numbers (such as 3.7 mm); nominal variables, which are expressed as names (such as “female”); and ranked variables, which are expressed as positions (such as “third”). You need to identify the types of variables in an experiment in order to choose the correct method of analysis.

Introduction

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. I classify variables into three types: measurement variables, nominal variables, and ranked variables. You’ll see other names for these variable types and other ways of classifying variables in other statistics references, so try not to get confused.

You’ll analyze similar experiments, with similar null and alternative hypotheses, 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 compare head width in males and females with a two-sample t -test or a 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 in males and females with a Fisher’s exact test. 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 compare unrolling time in males and females 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. Other names for them include “numeric” or “quantitative” variables.

Some authors divide measurement variables into two types. One type is continuous variables, such as length of an isopod’s antenna, which in theory have an infinite number of possible values. The other is discrete (or meristic) variables, which only have whole number values; these are things you count, such as the number of spines on an isopod’s antenna. The mathematical theories underlying statistical tests involving measurement variables assume that the variables are continuous. Luckily, these statistical tests work well on discrete measurement variables, so you usually don’t need to worry about the

difference between continuous and discrete measurement variables. The only exception would be if you have a very small number of possible values of a discrete variable, in which case you might want to treat it as a nominal variable instead.

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, let's say your isopods have 20 to 55 spines on their left antenna, and you want to know whether the average number of spines on the left antenna is different between males and females. You should consider spine number to be a measurement variable and analyze the data using a two-sample *t*-test or a one-way anova. If there are only two different spine numbers—some isopods have 32 spines, and some have 33—you should treat spine number as a nominal variable, with the values "32" and "33," and compare the proportions of isopods with 32 or 33 spines in males and females using a Fisher's exact test of independence (or chi-square or *G*-test of independence, if your sample size is really big). The same is true for laboratory experiments; if you give your isopods food with 15 different mannose concentrations and then measure their growth rate, mannose concentration would be a measurement variable; if you give some isopods food with 5 mM mannose, and the rest of the isopods get 25 mM mannose, then mannose concentration would be a nominal variable.

But what if you design an experiment with 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 isopod 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, a one-way anova could give a significant result. If your alternative hypothesis is "isopods 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. Simulation studies show that analyzing such *dependent* variables with the methods used for measurement variables works well (Fagerland et al. 2011). I am not aware of any research on the effect of treating *independent* variables with small numbers of values as measurement or nominal. Your decision about how to treat your variable will depend in part on your biological question. You may be able to 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.

Something that could be measured is a measurement variable, even when you set the values. For example, if you grow isopods with one batch of food containing 10 mM mannose, another batch of food with 20 mM mannose, another batch with 30 mM mannose, etc. up to 100 mM mannose, the different mannose concentrations are a measurement variable, even though you made the food and set the mannose concentration yourself.

Be careful when you count something, as it is sometimes a nominal variable and sometimes a measurement 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. Now, 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.

Ratios

Sometimes you can simplify your statistical analysis by taking the ratio of two measurement variables. For example, if you want to know whether male isopods have bigger heads, relative to body size, than female isopods, you could take the ratio of head width to body length for each isopod, and compare the mean ratios of males and females using a two-sample *t*-test. However, this assumes that the ratio is the same for different body sizes. We know that’s not true for humans—the head size/body size ratio in babies is freakishly large, compared to adults—so you should look at the regression of head width on body length and make sure the regression line goes pretty close to the origin, as a straight regression line through the origin means the ratios stay the same for different values of the *X* variable. If the regression line doesn’t go near the origin, it would be better to keep the two variables separate instead of calculating a ratio, and compare the regression line of head width on body length in males to that in females using an analysis of covariance.

Circular variables

One 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 polynomial regression of bear attacks vs. time of the year in Yellowstone National Park, you could treat “month” as a measurement variable, with March as 1 and November as 9; you wouldn’t have to worry that February (month 12) is next to March, because bears are hibernating in December through February, and you would ignore those three months.

However, if your variable really is circular, there are special, very obscure statistical tests designed just for circular data; chapters 26 and 27 in Zar (1999) are a good place to start.

Nominal variables

Nominal variables classify observations into discrete categories. 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). A good rule of thumb is that an individual observation of a nominal variable can be expressed as a word, not a number. If you have just two values of what would normally be a measurement variable, it’s nominal instead: think of it as “present” vs. “absent” or “low” vs. “high.” Nominal variables are often used to divide individuals up into categories, so that other variables may be compared among the categories. In the comparison of head width in male vs. female isopods, the isopods are classified by sex, a nominal variable, and the measurement variable head width is compared between the sexes.

Nominal variables are also called categorical, discrete, qualitative, or attribute variables. “Categorical” is a more common name than “nominal,” but some authors use “categorical” to include both what I’m calling “nominal” and what I’m calling “ranked,” while other authors use “categorical” just for what I’m calling nominal variables. I’ll stick with “nominal” to avoid this ambiguity.

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

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

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 time, but your stopwatch was broken, 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. When I write an exam question involving ranked variables, it’s usually some ridiculous scenario like “Imagine you’re on a desert island with no ruler, and you want to do statistics on the size of coconuts. You line them up from smallest to largest....” For a homework assignment, I ask students to pick a paper from their favorite biological journal and identify all the variables, and anyone who finds a ranked variable gets a donut; I’ve had to buy four donuts in 13 years. The only common biological ranked variables I can think of are dominance hierarchies in behavioral biology (see the dog example on the Kruskal-Wallis page) and developmental stages, such as the different instars that molting insects pass through.

The main reason that ranked variables are important is that the statistical tests designed for ranked variables (called “non-parametric tests”) 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 *not* a ranked variable; it 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.

Just like with measurement variables, if there are a very small number of possible values for a ranked variable, it would be better to treat it as a nominal variable. For example, if you make a honeybee sting people on one arm and a yellowjacket sting people on the other arm, then ask them “Was the honeybee sting the most painful or the second most painful?”, you are asking them for the rank of each sting. But you should treat the data as a nominal variable, one which has three values (“honeybee is worse” or “yellowjacket is worse” or “subject is so mad at your stupid, painful experiment that they refuse to answer”).

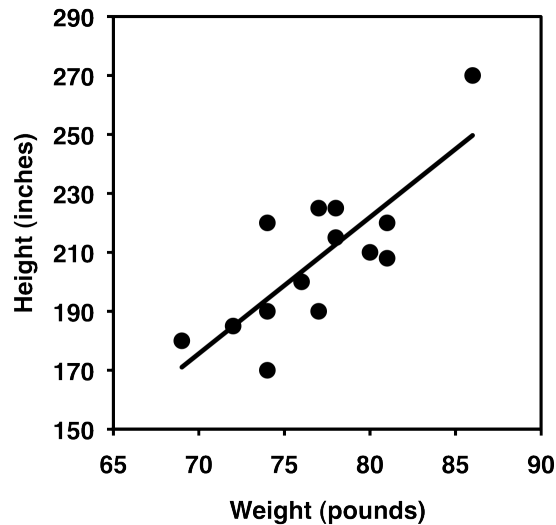
Categorizing

It is possible to convert a measurement variable to a nominal variable, dividing individuals up into a two or more classes based on ranges of the variable. For example, if you are studying the relationship between 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 two-sample *t*-test.

Converting measurement variables to nominal variables (“dichotomizing” if you split into two groups, “categorizing” in general) is common in epidemiology, psychology, and some other fields. However, there are several problems with categorizing measurement variables (MacCallum et al. 2002). One problem is that you’d be discarding a lot of information; in our blood pressure example, you’d be lumping together everyone with HDL from 0 to 39 mg/dl into one group. This reduces your statistical power, decreasing 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 (“cutpoint”) 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.

To illustrate the problem with categorizing, let’s say you wanted to know whether tall basketball players weigh more than short players. Here’s data for the 2012-2013 men’s basketball team at Morgan State University:

Height (inches)	Weight (pounds)
69	180
72	185
74	170
74	190
74	220
76	200
77	190
77	225
78	215
78	225
80	210
81	208
81	220
86	270



Height and weight of the Morgan State University men's basketball players.

If you keep both variables as measurement variables and analyze using linear regression, you get a P value of 0.0007; the relationship is highly significant. Tall basketball players really are heavier, as is obvious from the graph. However, if you divide the heights into two categories, "short" (77 inches or less) and "tall" (more than 77 inches) and compare the mean weights of the two groups using a two-sample t -test, the P value is 0.043, which is barely significant at the usual $P < 0.05$ level. And if you also divide the weights into two categories, "light" (210 pounds and less) and "heavy" (greater than 210 pounds), you get 6 who are short and light, 2 who are short and heavy, 2 who are tall and light, and 4 who are tall and heavy. The proportion of short people who are heavy is *not* significantly different from the proportion of tall people who are heavy, when analyzed using Fisher's exact test ($P = 0.28$). So by categorizing both measurement variables, you have made an obvious, highly significant relationship between height and weight become completely non-significant. This is not a good thing. I think it's better for most biological experiments if you don't categorize.

Likert items

Social scientists like to use Likert items: they'll present a statement like "It's important for all biologists to learn statistics" and ask people to choose 1=Strongly Disagree, 2=Disagree, 3=Neither Agree nor Disagree, 4=Agree, or 5=Strongly Agree. Sometimes they use seven values instead of five, by adding "Very Strongly Disagree" and "Very Strongly Agree"; and sometimes people are asked to rate their strength of agreement on a 9 or 11-point scale. Similar questions may have answers such as 1=Never, 2=Rarely, 3=Sometimes, 4=Often, 5=Always.

Strictly speaking, a Likert scale is the result of adding together the scores on several Likert items. Often, however, a single Likert item is called a Likert scale.

There is a lot of controversy about how to analyze a Likert item. One option is to treat it as a nominal variable with five (or seven, or however many) items. The data would then be summarized by the proportion of people giving each answer, and analyzed using chi-square or G -tests. However, this ignores the fact that the values go in order from least

agreement to most, which is pretty important information. The other options are to treat it as a ranked variable or a measurement variable.

Treating a Likert item as a measurement variable lets you summarize the data using a mean and standard deviation, and analyze the data using the familiar parametric tests such as anova and regression. One argument against treating a Likert item as a measurement variable is that the data have a small number of values that are unlikely to be normally distributed, but the statistical tests used on measurement variables are not very sensitive to deviations from normality, and simulations have shown that tests for measurement variables work well even with small numbers of values (Fagerland et al. 2011).

A bigger issue is that the answers on a Likert item are just crude subdivisions of some underlying measure of feeling, and the difference between “Strongly Disagree” and “Disagree” may not be the same size as the difference between “Disagree” and “Neither Agree nor Disagree”; in other words, the responses are not a true “interval” variable. As an analogy, imagine you asked a bunch of college students how much TV they watch in a typical week, and you give them the choices of 0=None, 1=A Little, 2=A Moderate Amount, 3=A Lot, and 4=Too Much. If the people who said “A Little” watch one or two hours a week, the people who said “A Moderate Amount” watch three to nine hours a week, and the people who said “A Lot” watch 10 to 20 hours a week, then the difference between “None” and “A Little” is a lot smaller than the difference between “A Moderate Amount” and “A Lot.” That would make your 0-4 point scale not be an interval variable. If your data actually were in hours, then the difference between 0 hours and 1 hour is the same size as the difference between 19 hours and 20 hours; “hours” would be an interval variable.

Personally, I don’t see how treating values of a Likert item as a measurement variable will cause any statistical problems. It is, in essence, a data transformation: applying a mathematical function to one variable to come up with a new variable. In chemistry, pH is the base-10 log of the reciprocal of the hydrogen activity, so the difference in hydrogen activity between a pH 5 and pH 6 solution is much bigger than the difference between pH 8 and pH 9. But I don’t think anyone would object to treating pH as a measurement variable. Converting 25-44 on some underlying “agreeicity index” to “2” and converting 45-54 to “3” doesn’t seem much different from converting hydrogen activity to pH, or micropascals of sound to decibels, or squaring a person’s height to calculate body mass index.

The impression I get, from briefly glancing at the literature, is that many of the people who use Likert items in their research treat them as measurement variables, while most statisticians think this is outrageously incorrect. I think treating them as measurement variables has several advantages, but you should carefully consider the practice in your particular field; it’s always better if you’re speaking the same statistical language as your peers. Because there is disagreement, you should include the number of people giving each response in your publications; this will provide all the information that other researchers need to analyze your data using the technique they prefer.

All of the above applies to statistics done on a single Likert item. The usual practice is to add together a bunch of Likert items into a Likert scale; a political scientist might add the scores on Likert questions about abortion, gun control, taxes, the environment, etc. and come up with a 100-point liberal vs. conservative scale. Once a number of Likert items are added together to make a Likert scale, there seems to be less objection to treating the sum as a measurement variable; even some statisticians are okay with that.

Independent and dependent variables

Another way to classify variables is as independent or dependent variables. An independent variable (also known as a predictor, explanatory, or exposure variable) is a

variable that you think may cause a change in a dependent variable (also known as an outcome or response variable). For example, if you grow isopods with 10 different mannose concentrations in their food and measure their growth rate, the mannose concentration is an independent variable and the growth rate is a dependent variable, because you think that different mannose concentrations may cause different growth rates. Any of the three variable types (measurement, nominal or ranked) can be either independent or dependent. For example, if you want to know whether sex affects body temperature in mice, sex would be an independent variable and temperature would be a dependent variable. If you wanted to know whether the incubation temperature of eggs affects sex in turtles, temperature would be the independent variable and sex would be the dependent variable.

As you'll see in the descriptions of particular statistical tests, sometimes it is important to decide which is the independent and which is the dependent variable; it will determine whether you should analyze your data with a two-sample *t*-test or simple logistic regression, for example. Other times you don't need to decide whether a variable is independent or dependent. For example, if you measure the nitrogen content of soil and the density of dandelion plants, you might think that nitrogen content is an independent variable and dandelion density is a dependent variable; you'd be thinking that nitrogen content might affect where dandelion plants live. But maybe dandelions use a lot of nitrogen from the soil, so it's dandelion density that should be the independent variable. Or maybe some third variable that you didn't measure, such as moisture content, affects both nitrogen content and dandelion density. For your initial experiment, which you would analyze using correlation, you wouldn't need to classify nitrogen content or dandelion density as independent or dependent. If you found an association between the two variables, you would probably want to follow up with experiments in which you manipulated nitrogen content (making it an independent variable) and observed dandelion density (making it a dependent variable), and other experiments in which you manipulated dandelion density (making it an independent variable) and observed the change in nitrogen content (making it the dependent variable).

References

- Fagerland, M. W., L. Sandvik, and P. Mowinckel. 2011. Parametric methods outperformed non-parametric methods in comparisons of discrete numerical variables. *BMC Medical Research Methodology* 11: 44.
- MacCallum, R. C., S. B. Zhang, K. J. Preacher, and D. D. Rucker. 2002. On the practice of dichotomization of quantitative variables. *Psychological Methods* 7: 19-40.
- Zar, J.H. 1999. *Biostatistical analysis*. 4th edition. Prentice Hall, Upper Saddle River, NJ.

Probability

Although estimating probabilities is a fundamental part of statistics, you will rarely have to do the calculations yourself. It's worth knowing a couple of simple rules about adding and multiplying probabilities.

Introduction

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. 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 2013 there were 22,166 students at the University of Delaware, and 3,679 of them were graduate students. If you sampled a single student at random, the probability that they would be a grad student would be $3,679 / 22,166$, or 0.166. In other words, 16.6% of students were grad students, so if you'd picked one student at random, the probability that they were a grad student would have been 16.6%.

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 *vestigial* locus, Mendel's theory predicts that the probability of an offspring individual being a recessive homozygote (having teeny-tiny wings) is one-fourth, or 0.25. This is equivalent to saying that one-fourth of a population of offspring will have tiny wings.

Multiplying probabilities

You could take a semester-long course on mathematical probability, but most biologists just need to know a few basic principles. You calculate the probability that an individual has one value of a nominal variable *and* another value of a second nominal variable by multiplying the probabilities of each value together. For example, if the probability that a *Drosophila* in a cross has vestigial wings is one-fourth, and the probability that it has legs where its antennae should be is three-fourths, the probability that it has vestigial wings *and* leg-antennae is one-fourth times three-fourths, or 0.25×0.75 , 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 vestigial wings *or* leg-antennae. You could calculate the probability for each of the four kinds of flies: normal wings/normal antennae ($0.75 \times 0.25 = 0.1875$), normal wings/leg-antennae ($0.75 \times 0.75 = 0.5625$), vestigial wings/normal antennae ($0.25 \times 0.25 = 0.0625$), and vestigial wings/leg-antennae ($0.25 \times 0.75 = 0.1875$). Then, since the last three kinds of flies are the ones with vestigial wings *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.

Basic concepts of hypothesis testing

One of the main goals of statistical hypothesis testing is to estimate the P value, which is the probability of obtaining the observed results, or something more extreme, if the null hypothesis were true. If the observed results are unlikely under the null hypothesis, you reject the null hypothesis. Alternatives to this “frequentist” approach to statistics include Bayesian statistics and estimation of effect sizes and confidence intervals.

Introduction

There are different ways of doing statistics. The technique used by the vast majority of biologists, and the technique that most of this handbook describes, is sometimes called “frequentist” or “classical” statistics. It involves testing a null hypothesis by comparing the data you observe in your experiment with the predictions of a null hypothesis. You estimate what the probability would be of obtaining the observed results, or something more extreme, if the null hypothesis were true. If this estimated probability (the P value) is small enough (below the significance value), then you conclude that it is unlikely that the null hypothesis is true; you reject the null hypothesis and accept an alternative hypothesis.

Many statisticians harshly criticize frequentist statistics, but their criticisms haven’t had much effect on the way most biologists do statistics. Here I will outline some of the key concepts used in frequentist statistics, then briefly describe some of the alternatives.

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 a 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. For example, let’s say you feed chocolate to a bunch of chickens, then look at the sex ratio in their offspring. If you get more females than males, it would be a tremendously exciting discovery: it would be a fundamental discovery about the mechanism of sex determination, female chickens are more valuable than male chickens in egg-laying

breeds, and you'd be able to publish your result in *Science* or *Nature*. Lots of people have spent a lot of time and money trying to change the sex ratio in chickens, and if you're successful, you'll be rich and famous. But if the chocolate doesn't change the sex ratio, it would be an extremely boring result, and you'd have a hard time getting it published in the *Eastern Delaware Journal of Chickenology*. It's therefore tempting to look for patterns in your data that support the exciting alternative hypothesis. For example, you might look at 48 offspring of chocolate-fed chickens and see 31 females and only 17 males. This looks promising, but 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 deviation from the null expectation that large, 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

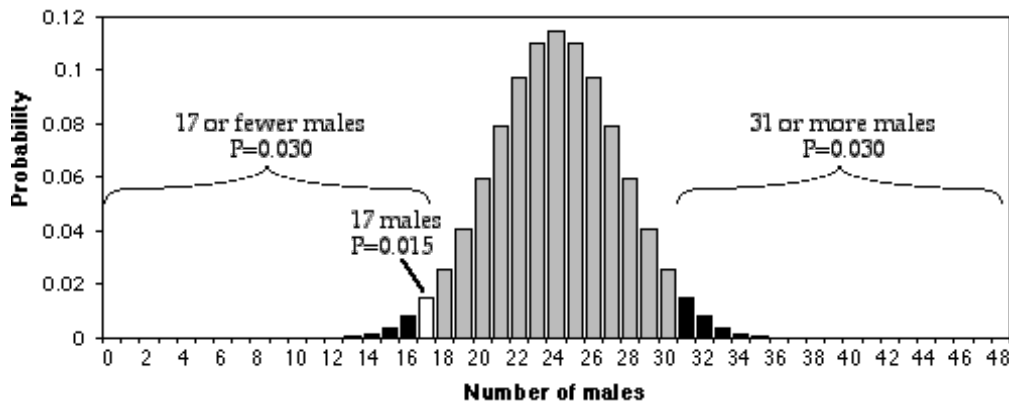
It is important to distinguish between *biological* null and alternative hypotheses and *statistical* null and alternative hypotheses. "Sexual selection by females has caused male chickens to evolve bigger feet than females" is a biological alternative hypothesis; it says something about biological processes, 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 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 fed chocolate to a bunch of female chickens (in birds, unlike mammals, the female parent determines the sex of the offspring), 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 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 chocolate 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, I used the BINOMDIST function of Excel 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% of chickens are male, is 0.030."

False positives vs. false negatives

After you do a statistical test, you are either going to reject or accept the null hypothesis. Rejecting the null hypothesis means that you conclude that the null hypothesis is not true; in our chicken sex example, you would conclude that the true proportion of male chicks, if you gave chocolate to an infinite number of chicken mothers, would be less than 50%.

When you reject a null hypothesis, there's a chance that you're making a mistake. The null hypothesis might really be true, and it may be that your experimental results deviate from the null hypothesis purely as a result of chance. In a sample of 48 chickens, it's possible to get 17 male chickens purely by chance; it's even possible (although extremely unlikely) to get 0 male and 48 female chickens purely by chance, even though the true

proportion is 50% males. This is why we never say we “prove” something in science; there’s always a chance, however miniscule, that our data are fooling us and deviate from the null hypothesis purely due to chance. When your data fool you into rejecting the null hypothesis even though it’s true, it’s called a “false positive,” or a “Type I error.” So another way of defining the P value is the probability of getting a false positive like the one you’ve observed, *if* the null hypothesis is true.

Another way your data can fool you is when you don’t reject the null hypothesis, even though it’s not true. If the true proportion of female chicks is 51%, the null hypothesis of a 50% proportion is not true, but you’re unlikely to get a significant difference from the null hypothesis unless you have a huge sample size. Failing to reject the null hypothesis, even though it’s not true, is a “false negative” or “Type II error.” This is why we never say that our data shows the null hypothesis to be true; all we can say is that we haven’t rejected the null hypothesis.

Significance levels

Does a probability of 0.030 mean that you should reject the null hypothesis, and conclude that chocolate 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 P value 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, it was chosen rather arbitrarily during the early days of statistics; people could have agreed upon 0.04, or 0.025, or 0.071 as the conventional significance level.

The significance level (also known as the “critical value” or “alpha”) you should use depends on the costs of different kinds of errors. With a significance level of 0.05, you have a 5% chance of rejecting the null hypothesis, even if it is true. If you try 100 different treatments on your chickens, and none of them really change the sex ratio, 5% of your experiments will give you data that are significantly different from a 1:1 sex ratio, just by chance. In other words, 5% of your experiments will give you a false positive. 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.

The relative costs of false positives and false negatives, and thus the best P value to use, will be different for different experiments. If you are screening a bunch of potential sex-ratio-changing treatments and get a false positive, it wouldn’t be a big deal; you’d just run a few more tests on that treatment until you were convinced the initial result was 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 for your initial tests. On the other hand, once your sex-ratio-changing treatment is undergoing final trials before being sold to farmers, a false positive could be very expensive; you’d want to be very confident that it really worked. 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, for your final tests.

The significance level you choose should also depend on how likely you think it is that your alternative hypothesis will be true, a prediction that you make *before* you do the experiment. This is the foundation of Bayesian statistics, as explained below.

You must choose your significance level before you collect the data, of course. If you choose to use a different significance level than the conventional 0.05, people will be

skeptical; you must be able to justify your choice. **Throughout this handbook, I will always use $P < 0.05$ as the significance level.** If you are doing an experiment where the cost of a false positive is a lot greater or smaller than the cost of a false negative, or an experiment where you think it is unlikely that the alternative hypothesis will be true, you should consider using a different 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% male chickens, would you really publish the result as “The treatment did not cause a significant decrease in the proportion of male chickens”? I hope 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, it 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 our chocolate-eating 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, which would be cheating. 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.

Computer statistics programs can give somewhat inaccurate P values when they are very small. Once your P values get very small, you can just say " $P < 0.00001$ " or some other impressively small number. You should also give either your raw data, or the test statistic and degrees of freedom, in case anyone wants to calculate your exact P value.

Effect sizes and confidence intervals

A fairly common criticism of the hypothesis-testing approach to statistics is that the null hypothesis will always be false, if you have a big enough sample size. In the chicken-feet example, critics would argue that if you had an infinite sample size, it is impossible that male chickens would have *exactly* the same average foot size as female chickens. Therefore, since you know before doing the experiment that the null hypothesis is false, there's no point in testing it.

This criticism only applies to two-tailed tests, where the null hypothesis is "Things are exactly the same" and the alternative is "Things are different." Presumably these critics think it would be okay to do a one-tailed test with a null hypothesis like "Foot length of male chickens is the same as, or less than, that of females," because the null hypothesis that male chickens have smaller feet than females could be true. So if you're worried about this issue, you could think of a two-tailed test, where the null hypothesis is that things are the same, as shorthand for doing two one-tailed tests. A significant rejection of the null hypothesis in a two-tailed test would then be the equivalent of rejecting one of the two one-tailed null hypotheses.

A related criticism is that a significant rejection of a null hypothesis might not be biologically meaningful, if the difference is too small to matter. For example, in the chicken-sex experiment, having a treatment that produced 49.9% male chicks might be significantly different from 50%, but it wouldn't be enough to make farmers want to buy your treatment. These critics say you should estimate the effect size and put a confidence interval on it, not estimate a P value. So the goal of your chicken-sex experiment should not be to say "Chocolate gives a proportion of males that is significantly less than 50% ($P = 0.015$)" but to say "Chocolate produced 36.1% males with a 95% confidence interval of 25.9 to 47.4%." For the chicken-feet experiment, you would say something like "The difference between males and females in mean foot size is 2.45 mm, with a confidence interval on the difference of ± 1.98 mm."

Estimating effect sizes and confidence intervals is a useful way to summarize your results, and it should usually be part of your data analysis; you'll often want to include confidence intervals in a graph. However, there are a lot of experiments where the goal is to decide a yes/no question, not estimate a number. In the initial tests of chocolate on chicken sex ratio, the goal would be to decide between "It changed the sex ratio" and "It didn't seem to change the sex ratio." *Any* change in sex ratio that is large enough that you could detect it would be interesting and worth follow-up experiments. While it's true that the difference between 49.9% and 50% might not be worth pursuing, you wouldn't do an experiment on enough chickens to detect a difference that small.

Often, the people who claim to avoid hypothesis testing will say something like "the 95% confidence interval of 25.9 to 47.4% does not include 50%, so we the plant extract significantly changed the sex ratio." This is a clumsy and roundabout form of hypothesis testing, and they might as well admit it and report the P value.

Bayesian statistics

Another alternative to frequentist statistics is Bayesian statistics. A key difference is that Bayesian statistics requires specifying your best guess of the probability of each

possible value of the parameter to be estimated, before the experiment is done. This is known as the “prior probability.” So for your chicken-sex experiment, you’re trying to estimate the “true” proportion of male chickens that would be born, if you had an infinite number of chickens. You would have to specify how likely you thought it was that the true proportion of male chickens was 50%, or 51%, or 52%, or 47.3%, etc. You would then look at the results of your experiment and use the information to calculate new probabilities that the true proportion of male chickens was 50%, or 51%, or 52%, or 47.3%, etc. (the posterior distribution).

I’ll confess that I don’t really understand Bayesian statistics, and I apologize for not explaining it well. In particular, I don’t understand how people are supposed to come up with a prior distribution for the kinds of experiments that most biologists do. With the exception of systematics, where Bayesian estimation of phylogenies is quite popular and seems to make sense, I haven’t seen many research biologists using Bayesian statistics for routine data analysis of simple laboratory experiments. This means that even if the cult-like adherents of Bayesian statistics convinced you that they were right, you would have a difficult time explaining your results to your biologist peers. Statistics is a method of conveying information, and if you’re speaking a different language than the people you’re talking to, you won’t convey much information. So I’ll stick with traditional frequentist statistics for this handbook.

Having said that, there’s one key concept from Bayesian statistics that is important for all users of statistics to understand. To illustrate it, imagine that you are testing extracts from 1000 different tropical plants, trying to find something that will kill beetle larvae. The reality (which you don’t know) is that 500 of the extracts kill beetle larvae, and 500 don’t. You do the 1000 experiments and do the 1000 frequentist statistical tests, and you use the traditional significance level of $P < 0.05$. The 500 plant extracts that really work all give you $P < 0.05$; these are the true positives. Of the 500 extracts that don’t work, 5% of them give you $P < 0.05$ by chance (this is the meaning of the P value, after all), so you have 25 false negatives. So you end up with 525 plant extracts that gave you a P value less than 0.05. You’ll have to do further experiments to figure out which are the 25 false positives and which are the 500 true positives, but that’s not so bad, since you know that most of them will turn out to be true positives.

Now imagine that you are testing those extracts from 1000 different tropical plants to try to find one that will make hair grow. The reality (which you don’t know) is that one of the extracts makes hair grow, and the other 999 don’t. You do the 1000 experiments and do the 1000 frequentist statistical tests, and you use the traditional significance level of $P < 0.05$. The one plant extract that really works gives you $P < 0.05$; this is the true positive. But of the 999 extracts that don’t work, 5% of them give you $P < 0.05$ by chance, so you have about 50 false negatives. You end up with 51 P values less than 0.05, but almost all of them are false positives.

Now instead of testing 1000 plant extracts, imagine that you are testing just one. If you are testing it to see if it kills beetle larvae, you know (based on everything you know about plant and beetle biology) there’s a pretty good chance it will work, so you can be pretty sure that a P value less than 0.05 is a true positive. But if you are testing that one plant extract to see if it grows hair, which you know is very unlikely (based on everything you know about plants and hair), a P value less than 0.05 is almost certainly a false positive. In other words, **if you expect that the null hypothesis is probably true, a statistically significant result is probably a false positive.** This is sad; the most exciting, amazing, unexpected results in your experiments are probably just your data trying to make you jump to ridiculous conclusions. You should require a much lower P value to reject a null hypothesis that you think is probably true.

A Bayesian would insist that you put in numbers just how likely you think the null hypothesis and various values of the alternative hypothesis are, before you do the experiment, and I’m not sure how that is supposed to work in practice for most

experimental biology. But the general concept is a valuable one: as Carl Sagan summarized it, “Extraordinary claims require extraordinary evidence.”

Recommendations

Here are three experiments to illustrate when the different approaches to statistics are appropriate. In the first experiment, you are testing a plant extract on rabbits to see if it will lower their blood pressure. You already know that the plant extract is a diuretic (makes the rabbits pee more) and you already know that diuretics tend to lower blood pressure, so you think there’s a good chance it will work. If it does work, you’ll do more low-cost animal tests on it before you do expensive, potentially risky human trials. Your prior expectation is that the null hypothesis (that the plant extract has no effect) has a good chance of being false, and the cost of a false positive is fairly low. So you should do frequentist hypothesis testing, with a significance level of 0.05.

In the second experiment, you are going to put human volunteers with high blood pressure on a strict low-salt diet and see how much their blood pressure goes down. Everyone will be confined to a hospital for a month and fed either a normal diet, or the same foods with half as much salt. For this experiment, you wouldn’t be very interested in the P value, as based on prior research in animals and humans, you are already quite certain that reducing salt intake will lower blood pressure; you’re pretty sure that the null hypothesis that “Salt intake has no effect on blood pressure” is false. Instead, you are very interested to know how *much* the blood pressure goes down. Reducing salt intake in half is a big deal, and if it only reduces blood pressure by 1 mm Hg, the tiny gain in life expectancy wouldn’t be worth a lifetime of bland food and obsessive label-reading. If it reduces blood pressure by 20 mm with a confidence interval of ± 5 mm, it might be worth it. So you should estimate the effect size (the difference in blood pressure between the diets) and the confidence interval on the difference.

In the third experiment, you are going to put magnetic hats on guinea pigs and see if their blood pressure goes down (relative to guinea pigs wearing the kind of non-magnetic hats that guinea pigs usually wear). This is a really goofy experiment, and you know that it is very unlikely that the magnets will have any effect (it’s not impossible—magnets affect the sense of direction of homing pigeons, and maybe guinea pigs have something similar in their brains and maybe it will somehow affect their blood pressure—it just seems really unlikely). You might analyze your results using Bayesian statistics, which will require specifying in numerical terms just how unlikely you think it is that the magnetic hats will work. Or you might use frequentist statistics, but require a P value much, much lower than 0.05 to convince yourself that the effect is real.

Confounding variables

A confounding variable is a variable other than the independent variable that you're interested in, that may affect the dependent variable. This can lead to erroneous conclusions about the relationship between the independent and dependent variables. You deal with confounding variables by controlling them; by matching; by randomizing; or by statistical control.

Introduction

Due to a variety of genetic, developmental, and environmental factors, no two organisms, no two tissue samples, no two cells are exactly alike. This means that when you design an experiment with samples that differ in independent variable X , your samples will also differ in other variables that you may or may not be aware of. If these confounding variables affect the dependent variable Y that you're interested in, they may trick you into thinking there's a relationship between X and Y when there really isn't. Or, the confounding variables may cause so much variation in Y that it's hard to detect a real relationship between X and Y when there is one.

As an example of confounding variables, imagine that you want to know whether the genetic differences between American elms (which are susceptible to Dutch elm disease) and Princeton elms (a strain of American elms that is resistant to Dutch elm disease) cause a difference in the amount of insect damage to their leaves. You look around your area, 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, which is what you were looking for. 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. If you conclude that Princeton elms have more insect damage because of the genetic difference between the strains, when in reality it's because the Princeton elms in your sample were younger, you will look like an idiot to all of your fellow elm scientists as soon as they figure out your mistake.

On the other hand, let's say you're not *that* much of an idiot, and you make sure your sample of Princeton elms has the same average age as your sample of American elms. There's still a lot of variation in ages among the individual trees in each sample, and if that

affects insect damage, there will be a lot of variation among individual trees in the amount of insect damage. This will make it harder to find a statistically significant difference in insect damage between the two strains of elms, and you might miss out on finding a small but exciting difference in insect damage between the strains.

Controlling confounding variables

Designing an experiment to eliminate differences due to confounding variables is critically important. One way is to control a possible confounding variable, meaning you keep it identical for all the individuals. For example, you could plant a bunch of American elms and a bunch of Princeton elms all at the same time, so they'd be the same age. You could plant them in the same field, and give them all the same amount of water and fertilizer.

It is easy to control many of the possible confounding variables in laboratory experiments on model organisms. All of your mice, or rats, or *Drosophila* will be the same age, the same sex, and the same inbred genetic strain. They will grow up in the same kind of containers, eating the same food and drinking the same water. But there are always some possible confounding variables that you can't control. Your organisms may all be from the same genetic strain, but new mutations will mean that there are still some genetic differences among them. You may give them all the same food and water, but some may eat or drink a little more than others. After controlling all of the variables that you can, it is important to deal with any other confounding variables by randomizing, matching or statistical control.

Controlling confounding variables is harder with organisms that live outside the laboratory. Those elm trees that you planted in the same field? Different parts of the field may have different soil types, different water percolation rates, different proximity to roads, houses and other woods, and different wind patterns. And if your experimental organisms are humans, there are a lot of confounding variables that are impossible to control.

Randomizing

Once you've designed your experiment to control as many confounding variables as possible, you need to randomize your samples to make sure that they don't differ in the confounding variables that you can't control. For example, let's say you're going to make 20 mice wear sunglasses and leave 20 mice without glasses, to see if sunglasses help prevent cataracts. You shouldn't reach into a bucket of 40 mice, grab the first 20 you catch and put sunglasses on them. The first 20 mice you catch might be easier to catch because they're the slowest, the tamest, or the ones with the longest tails; or you might subconsciously pick out the fattest mice or the cutest mice. I don't know whether having your sunglass-wearing mice be slower, tamer, with longer tails, fatter, or cuter would make them more or less susceptible to cataracts, but you don't know either. You don't want to find a difference in cataracts between the sunglass-wearing and non-sunglass-wearing mice, then have to worry that maybe it's the extra fat or longer tails, not the sunglasses, that caused the difference. So you should randomly assign the mice to the different treatment groups. You could give each mouse an ID number and have a computer randomly assign them to the two groups, or you could just flip a coin each time you pull a mouse out of your bucket of mice.

In the mouse example, you used all 40 of your mice for the experiment. Often, you will sample a small number of observations from a much larger population, and it's important that it be a random sample. In a random sample, each individual has an equal probability of being sampled. To get a random sample of 50 elm trees from a forest with 700 elm trees, you could figure out where each of the 700 elm trees is, give each one an ID number, write

the numbers on 700 slips of paper, put the slips of paper in a hat, and randomly draw out 50 (or have a computer randomly choose 50, if you're too lazy to fill out 700 slips of paper or don't own a hat).

You need to be careful to make sure that your sample is truly random. I started to write "Or an easier way to randomly sample 50 elm trees would be to randomly pick 50 locations in the forest by having a computer randomly choose GPS coordinates, then sample the elm tree nearest each random location." However, this would have been a mistake; an elm tree that was far away from other elm trees would almost certainly be the closest to one of your random locations, but you'd be unlikely to sample an elm tree in the middle of a dense bunch of elm trees. It's pretty easy to imagine that proximity to other elm trees would affect insect damage (or just about anything else you'd want to measure on elm trees), so I almost designed a stupid experiment for you.

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.

Matching

Sometimes there's a lot of variation in confounding variables that you can't control; even if you randomize, the large variation in confounding variables may cause so much variation in your dependent variable that it would be hard to detect a difference caused by the independent variable that you're interested in. This is particularly true for humans. Let's say you want to test catnip oil as a mosquito repellent. If you were testing it on rats, you would get a bunch of rats of the same age and sex and inbred genetic strain, apply catnip oil to half of them, then put them in a mosquito-filled room for a set period of time and count the number of mosquito bites. This would be a nice, well-controlled experiment, and with a moderate number of rats you could see whether the catnip oil caused even a small change in the number of mosquito bites. But if you wanted to test the catnip oil on humans going about their everyday life, you couldn't get a bunch of humans of the same "inbred genetic strain," it would be hard to get a bunch of people all of the same age and sex, and the people would differ greatly in where they lived, how much time they spent outside, the scented perfumes, soaps, deodorants, and laundry detergents they used, and whatever else it is that makes mosquitoes ignore some people and eat others up. The very large variation in number of mosquito bites among people would mean that if the catnip oil had a small effect, you'd need a huge number of people for the difference to be statistically significant.

One way to reduce the noise due to confounding variables is by matching. You generally do this when the independent variable is a nominal variable with two values, such as "drug" vs. "placebo." You make observations in pairs, one for each value of the

independent variable, that are as similar as possible in the confounding variables. The pairs could be different parts of the same people. For example, you could test your catnip oil by having people put catnip oil on one arm and placebo oil on the other arm. The variation in the size of the *difference* between the two arms on each person could be a lot smaller than the variation among different people, so you won't need nearly as big a sample size to detect a small difference in mosquito bites between catnip oil and placebo oil. Of course, you'd have to randomly choose which arm to put the catnip oil on.

Other ways of pairing include before-and-after experiments. You could count the number of mosquito bites in one week, then have people use catnip oil and see if the number of mosquito bites for each person went down. With this kind of experiment, it's important to make sure that the dependent variable wouldn't have changed by itself (maybe the weather changed and the mosquitoes stopped biting), so it would be better to use placebo oil one week and catnip oil another week, and randomly choose for each person whether the catnip oil or placebo oil was first.

For many human experiments, you'll need to match two different people, because you can't test both the treatment and the control on the same person. For example, let's say you've given up on catnip oil as a mosquito repellent and are going to test it on humans as a cataract preventer. You're going to get a bunch of people, have half of them take a catnip-oil pill and half take a placebo pill for five years, then compare the lens opacity in the two groups. Here the goal is to make each pair of people be as similar as possible in confounding variables that you think might be important. If you're studying cataracts, you'd want to match people based on known risk factors for cataracts: age, amount of time outdoors, use of sunglasses, blood pressure. Of course, once you have a matched pair of individuals, you'd want to randomly choose which one gets the catnip oil and which one gets the placebo. You wouldn't be able to find perfectly matching pairs of individuals, but the better the match, the easier it will be to detect a difference due to the catnip-oil pills.

One kind of matching that is often used in epidemiology is the case-control study. "Cases" are people with some disease or condition, and each is matched with one or more controls. Each control is generally the same sex and as similar in other factors (age, ethnicity, occupation, income) as practical. The cases and controls are then compared to see whether there are consistent differences between them. For example, if you wanted to know whether smoking marijuana caused or prevented cataracts, you could find a bunch of people with cataracts. You'd then find a control for each person who was similar in the known risk factors for cataracts (age, time outdoors, blood pressure, diabetes, steroid use). Then you would ask the cataract cases and the non-cataract controls how much weed they'd smoked.

If it's hard to find cases and easy to find controls, a case-control study may include two or more controls for each case. This gives somewhat more statistical power.

Statistical control

When it isn't practical to keep all the possible confounding variables constant, another solution is to statistically control them. Sometimes you can do this with a simple ratio. If you're interested in the effect of weight on cataracts, height would be a confounding variable, because taller people tend to weigh more. Using the body mass index (BMI), which is the ratio of weight in kilograms over the squared height in meters, would remove much of the confounding effects of height in your study. If you need to remove the effects of multiple confounding variables, there are multivariate statistical techniques you can use. However, the analysis, interpretation, and presentation of complicated multivariate analyses are not easy.

Observer or subject bias as a confounding variable

In many studies, the possible bias of the researchers is one of the most important confounding variables. Finding a statistically significant result is almost always more interesting than not finding a difference, so you need to constantly be on guard to control the effects of this bias. The best way to do this is by blinding yourself, so that you don't know which individuals got the treatment and which got the control. Going back to our catnip oil and mosquito experiment, if you know that Alice got catnip oil and Bob didn't, your subconscious body language and tone of voice when you talk to Alice might imply "You didn't get very many mosquito bites, did you? That would mean that the world will finally know what a genius I am for inventing this," and you might carefully scrutinize each red bump and decide that some of them were spider bites or poison ivy, not mosquito bites. With Bob, who got the placebo, you might subconsciously imply "Poor Bob—I'll bet you got a ton of mosquito bites, didn't you? The more you got, the more of a genius I am" and you might be more likely to count every hint of a bump on Bob's skin as a mosquito bite. Ideally, the subjects shouldn't know whether they got the treatment or placebo, either, so that they can't give you the result you want; this is especially important for subjective variables like pain. Of course, keeping the subjects of this particular imaginary experiment blind to whether they're rubbing catnip oil on their skin is going to be hard, because Alice's cat keeps licking Alice's arm and then acting stoned.

Exact test of goodness-of-fit

You use the exact test of goodness-of-fit when you have one nominal variable, you want to see whether the number of observations in each category fits a theoretical expectation, and the sample size is small.

Introduction

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 of goodness-of-fit, are different. There is no test statistic; instead, you directly calculate the probability of obtaining the observed data under the null hypothesis. 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 test of goodness-of-fit when you have one nominal variable. The most common use is a nominal variable with only two values (such as male or female, left or right, green or yellow), in which case the test may be called the exact binomial test. You compare the observed data 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 you determined before you collected the data. 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 you should use a *G*-test or chi-square test of goodness-of-fit instead (and they 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 most of 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 0.5×0.5 , 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 of the binomial distribution 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 a 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, 10, is the total number of trials. The third number is the expected proportion of whichever event there were fewer than expected of, if the null hypothesis were true; here the null hypothesis predicts that half of all ribbon-battings will be with the left paw. 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

EXACT TEST OF GOODNESS-OF-FIT

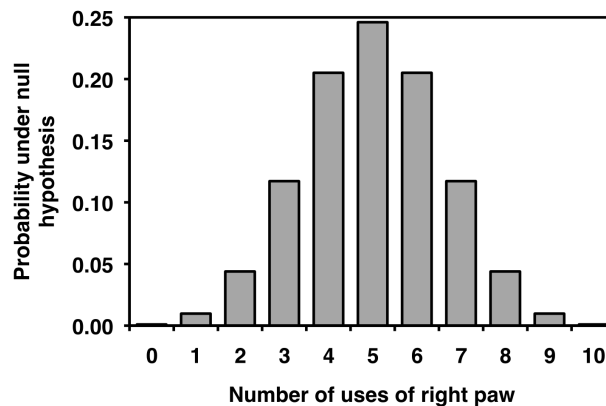
0 times out of ten. Adding these probabilities together gives $P=0.055$, which is not quite significant at the $P<0.05$ level. You do this in a spreadsheet by entering

`=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 you calculate the two-tailed P value 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, statisticians disagree about the meaning of a two-tailed exact binomial test, and different statistical programs will give slightly different results. The simplest method is to use the binomial equation, as described above, to calculate the probability of whichever event is less common than expected, then multiply it by two. For example, let’s say you’ve crossed a number of cats that are heterozygous at the hair-length gene; because short hair is dominant, you expect 75% of the kittens to have short hair and 25% to have long hair. You end up with 7 short haired and 5 long haired cats. There are 7 short haired cats when you expected 9, so you use the binomial equation to calculate the probability of 7 or fewer short-haired cats; this adds up to 0.158. Doubling this would give you a two-tailed P value of 0.315. This is what SAS and Richard Lowry’s online calculator (faculty.vassar.edu/lowry/binomialX.html) do.

The alternative approach is called the method of small P values (see <http://www.quantitativeskills.com/sisa/papers/paper5.htm>), and I think most statisticians prefer it. For our example, you use the binomial equation to calculate the probability of obtaining exactly 7 out of 12 short-haired cats; it is 0.103. Then you calculate the probabilities for every other possible number of short-haired cats, and you add together those that are less than 0.103. That is the probabilities for 6, 5, 4 ... 0 short-haired cats, and in the other tail, only the probability of 12 out of 12 short-haired cats. Adding these probabilities gives a P value of 0.189. This is what my exact binomial spreadsheet does. I think the arguments in favor of the method of small P values make sense. If you are using the exact binomial test with expected proportions other than 50:50, make sure you specify which method you use (remember that it doesn't matter when the expected proportions are 50:50).

Sign test

One common application of the exact binomial test is known as the sign test. You use the sign 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" or "left" and "right," and the other nominal variable identifies the pairs of observations. In a study of a hair-growth ointment, "amount of hair" would be the measurement variable, "before" and "after" would be the values of one nominal variable, and "Arnold," "Bob," "Charles" would be values of the second nominal variable.

The data for a sign test usually 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. However, sometimes you're not interested in the size of the difference, just the direction. In the hair-growth example, you might have decided that you didn't care how much hair the men grew or lost, you just wanted to know whether more than half of the men grew hair. In that case, you count the number of differences in one direction, count the number of differences in the opposite direction, and use the exact binomial test to see whether the numbers are different from a 1:1 ratio.

You should decide that a sign test is the test you want before you look at the data. If you analyze your data with a paired t -test and it's not significant, then you notice it would be significant with a sign test, it would be very unethical to just report the result of the sign test as if you'd planned that from the beginning.

Exact multinomial test

While the most common use of exact tests of goodness-of-fit is the exact binomial test, it is also possible to perform exact multinomial tests when there are more than two values of the nominal variable. The most common example in biology would be the results of genetic crosses, where one might expect a 1:2:1 ratio from a cross of two heterozygotes at one codominant locus, a 9:3:3:1 ratio from a cross of individuals heterozygous at two dominant loci, etc. The basic procedure is the same as for the exact binomial test: you calculate the probabilities of the observed result and all more extreme possible results and add them together. The underlying computations are more complicated, and if you have a lot of categories, your computer may have problems even if the total sample size is less than 1000. If you have a small sample size but so many categories that your computer program won't do an exact test, you can use a G -test or chi-square test of goodness-of-fit, but understand that the results may be somewhat inaccurate.

Post-hoc test

If you perform the exact multinomial test (with more than two categories) and get a significant result, you may want to follow up by testing whether each category deviates significantly from the expected number. It's a little odd to talk about just one category deviating significantly from expected; if there are more observations than expected in one category, there have to be fewer than expected in at least one other category. But looking at each category might help you understand better what's going on.

For example, let's say you do a genetic cross in which you expect a 9:3:3:1 ratio of purple, red, blue, and white flowers, and your observed numbers are 72 purple, 38 red, 20 blue, and 18 white. You do the exact test and get a P value of 0.0016, so you reject the null hypothesis. There are fewer purple and blue and more red and white than expected, but is there an individual color that deviates significantly from expected?

To answer this, do an exact binomial test for each category vs. the sum of all the other categories. For purple, compare the 72 purple and 76 non-purple to the expected 9:7 ratio. The P value is 0.07, so you can't say there are significantly fewer purple flowers than expected (although it's worth noting that it's close). There are 38 red and 110 non-red flowers; when compared to the expected 3:13 ratio, the P value is 0.035. This is below the significance level of 0.05, but because you're doing four tests at the same time, you need to correct for the multiple comparisons. Applying the Bonferroni correction, you divide the significance level (0.05) by the number of comparisons (4) and get a new significance level of 0.0125; since 0.035 is greater than this, you can't say there are significantly more red flowers than expected. Comparing the 18 white and 130 non-white to the expected ratio of 1:15, the P value is 0.006, so you can say that there are significantly more white flowers than expected.

It is possible that an overall significant P value could result from moderate-sized deviations in all of the categories, and none of the post-hoc tests will be significant. This would be frustrating; you'd know that something interesting was going on, but you couldn't say with statistical confidence exactly what it was.

I doubt that the procedure for post-hoc tests in a goodness-of-fit test that I've suggested here is original, but I can't find a reference to it; if you know who really invented this, e-mail me with a reference. And it seems likely that there's a better method that takes into account the non-independence of the numbers in the different categories (as the numbers in one category go up, the number in some other category must go down), but I have no idea what it might be.

Intrinsic hypothesis

You use exact test of goodness-of-fit that I've described here when testing fit to an extrinsic hypothesis, a hypothesis that you knew before you collected the data. For example, even before the kittens are born, you can predict that the ratio of short-haired to long-haired cats will be 3:1 in a genetic cross of two heterozygotes. Sometimes you want to test the fit to an intrinsic null hypothesis: one that is based on the data you collect, where you can't predict the results from the null hypothesis until after you collect the data. The only example I can think of in biology is Hardy-Weinberg proportions, where the number of each genotype in a sample from a wild population is expected to be p^2 or $2pq$ or q^2 (with more possibilities when there are more than two alleles); you don't know the allele frequencies (p and q) until after you collect the data. Exact tests of fit to Hardy-Weinberg raise a number of statistical issues and have received a lot of attention from population geneticists; if you need to do this, see Engels (2009) and the older references he cites. If you have biological data that you want to do an exact test of goodness-of-fit with an intrinsic hypothesis on, and it doesn't involve Hardy-Weinberg, e-mail me; I'd be very curious to see what kind of biological data requires this, and I will try to help you as best as I can.

Assumptions

Goodness-of-fit tests assume that the individual observations are independent, meaning that the value of one observation does not influence the value of other observations. To give an example, let's say you want to know what color of flowers that bees like. You plant four plots of flowers: one purple, one red, one blue, and one white. You get a bee, put it in a dark jar, carry it to a point equidistant from the four plots of flowers, and release it. You record which color flower it goes to first, then re-capture it and hold it prisoner until the experiment is done. You do this again and again for 100 bees. In this case, the observations are independent; the fact that bee #1 went to a blue flower has no influence on where bee #2 goes. This is a good experiment; if significantly more than $1/4$ of the bees go to the blue flowers, it would be good evidence that the bees prefer blue flowers.

Now let's say that you put a beehive at the point equidistant from the four plots of flowers, and you record where the first 100 bees go. If the first bee happens to go to the plot of blue flowers, it will go back to the hive and do its bee-butt-wiggling dance that tells the other bees, "Go 15 meters southwest, there's a bunch of yummy nectar there!" Then some more bees will fly to the blue flowers, and when they return to the hive, they'll do the same bee-butt-wiggling dance. The observations are NOT independent; where bee #2 goes is strongly influenced by where bee #1 happened to go. If "significantly" more than $1/4$ of the bees go to the blue flowers, it could easily be that the first bee just happened to go there by chance, and bees may not really care about flower color.

Examples

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 ($P=9 \times 10^{-6}$).

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.

As an example of the sign test, 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 (Corthyliina 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.	Gymnosperm-feeding	Spp.
Corthylinea	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 + Spondyliinae	78
Megalopodinae	400	Palophaginae	3
Higher Chrysomelidae	33400	Aulocoscelinae + Orsodacninae	26

Mendel (1865) 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 using the method of small P values, as implemented in my spreadsheet, 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. (SAS uses a different method that gives a P value of 0.530. With a smaller sample size, the difference between the “method of small P values” that I and most statisticians prefer, and the cruder method that SAS uses, could be large enough to be important.)

Mendel (1865) also crossed peas that were heterozygous at two genes: one for yellow vs. green, the other for round vs. wrinkled; yellow was dominant over green, and round was dominant over wrinkled. The expected and observed results were:

	expected ratio	expected number	observed number
yellow+round	9	312.75	315
green+round	3	104.25	108
yellow+wrinkled	3	104.25	101
round+wrinkled	1	34.75	32

This is an example of the exact multinomial test, since there are four categories, not two. The P value is 0.93, so the difference between observed and expected is nowhere near significance.

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 chapter on small sample sizes for further discussion.

If you try to do an exact test with a large number of categories, your computer may not be able to do the calculations even if your total sample size is less than 1000. In that case, you can cautiously use the G -test or chi-square goodness-of-fit test, knowing that the results may be somewhat inaccurate.

The exact test of goodness-of-fit is not the same as Fisher's exact test of independence. You use a test of independence 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 (www.biostathandbook.com/exactbin.xls). It is self-explanatory. It uses the method of small P values when the expected proportions are different from 50:50.

Web page

Richard Lowry has set up a web page that does the exact binomial test (faculty.vassar.edu/lowry/binomialX.html). It does not use the method of small P values, so I do not recommend it if your expected proportions are different from 50:50. I'm not aware of any web pages that will do the exact binomial test using the method of small P values, and 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%, so the technique SAS uses isn't as good as the method of small P values. I

EXACT TEST OF GOODNESS-OF-FIT

don't recommend doing the exact binomial test in SAS when the expected proportion is anything other than 50%.

```
DATA gus;
  INPUT paw $;
  DATALINES;
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;
  DATALINES;
right 10
left 2
;
PROC FREQ DATA=gus;
  WEIGHT count / ZEROS;
  TABLES paw / BINOMIAL(P=0.5);
  EXACT BINOMIAL;
RUN;
```

This example shows how to do 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, wrnkyl, 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;
  DATALINES;
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 labeled “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

Before you do an experiment, you should do a power analysis to estimate the sample size you’ll need. To do this for an exact binomial test using G*Power, choose “Exact” under “Test Family” and choose “Proportion: Difference from constant” under “Statistical test.” Under “Type of power analysis”, choose “A priori: Compute required sample size”. For “Input parameters,” enter the number of tails (you’ll almost always want two), alpha (usually 0.05), and Power (often 0.5, 0.8, or 0.9). The “Effect size” is the difference in proportions between observed and expected that you hope to see, and the “Constant proportion” is the expected proportion for one of the two categories (whichever is smaller). Hit “Calculate” and you’ll get the Total Sample Size.

As an example, let’s say you 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, so enter that in “Constant proportion”. You decide that if the probability of him using his left paw is 0.40, you want your experiment to have an 80% probability of getting a significant ($P < 0.05$) result, so enter 0.10 for Effect Size, 0.05 for Alpha, and 0.80 for Power. If he uses his left paw 60% of the time, you’ll accept that as a significant result too, so it’s a two-tailed test. The result is 199. This means that if Gus really is using his left paw 40% (or 60%) of the time, a sample size of 199 observations will have an 80% probability of giving you a significant ($P < 0.05$) exact binomial test.

Many power calculations for the exact binomial test, like G*Power, find the smallest sample size that will give the desired power, but there is a “sawtooth effect” in which increasing the sample size can actually *reduce* the power. Chernick and Liu (2002) suggest finding the smallest sample size that will give the desired power, even if the sample size is increased. For the Gus example, the method of Chernick and Liu gives a sample size of 210, rather than the 199 given by G*Power. Because both power and effect size are usually just arbitrary round numbers, where it would be easy to justify other values that would change the required sample size, the small differences in the method used to calculate desired sample size are probably not very important. The only reason I mention this is so

that you won't be alarmed if different power analysis programs for the exact binomial test give slightly different results for the same parameters.

G*Power does not do a power analysis for the exact test with more than two categories. If you have to do a power analysis and your nominal variable has more than two values, use the power analysis for chi-square tests in G*Power instead. The results will be pretty close to a true power analysis for the exact multinomial test, and given the arbitrariness of parameters like power and effect size, the results should be close enough.

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. *American Statistician* 56: 149-155.
- Engels, W.R. 2009. Exact tests for Hardy-Weinberg proportions. *Genetics* 183: 1431-1441.
- 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.
- Mendel, G. 1865. Experiments in plant hybridization. available at www.mendelweb.org/Mendel.html
- Sullivan, B.T., E.M. Pettersson, K.C. Seltsmann, and C.W. Berisford. 2000. Attraction of the bark beetle parasitoid *Roptrocercus xylophagorum* (Hymenoptera: Pteromalidae) to host-associated olfactory cues. *Environmental Entomology* 29: 1138-1151.
- Yukilevich, R., and J.R. True. 2008. Incipient sexual isolation among cosmopolitan *Drosophila melanogaster* populations. *Evolution* 62: 2112-2121.

Power analysis

Before you do an experiment, you should perform a power analysis to estimate the number of observations you need to have a good chance of detecting the effect you're looking for.

Introduction

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 when 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 large, expensive study would not be worthwhile.

Parameters

There are four or five numbers involved in a power analysis. You must choose the values for each one before you do the analysis. If you don't have a good reason for using a particular value, you can try different values and look at the effect on sample size.

Effect size

The 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%. You would then say that your effect size is 10%. If you're testing something to make the hens lay more eggs, the effect size might be 2 eggs per month.

Occasionally, you'll have a good economic or clinical reason for choosing a particular effect size. If you're testing a chicken feed supplement that costs \$1.50 per month, you're only interested in finding out whether it will produce more than \$1.50 worth of extra eggs each month; knowing that a supplement produces an extra 0.1 egg a month is not useful information to you, and you don't need to design your experiment to find that out. But for most basic biological research, the effect size is just a nice round number that you pulled out of your butt. Let's say you're doing a power analysis for a study of a mutation in a promoter region, to see if it affects gene expression. How big a change in gene expression are you looking for: 10%? 20%? 50%? It's a pretty arbitrary number, but it will have a huge effect on the number of transgenic mice who will give their expensive little lives for your science. If you don't have a good reason to look for a particular effect size, you might as well admit that and draw a graph with sample size on the X-axis and effect size on the Y-axis. G*Power will do this for you.

Alpha

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 or power

Beta, in a power analysis, is the probability 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 (getting a significant result) when the real difference is equal to the minimum effect size. Power is $1-\beta$. There is no clear consensus on the value to use, so this is another number you pull out of your butt; a power of 80% (equivalent to a beta of 20%) is probably the most common, while some people use 50% or 90%. 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.

Standard deviation

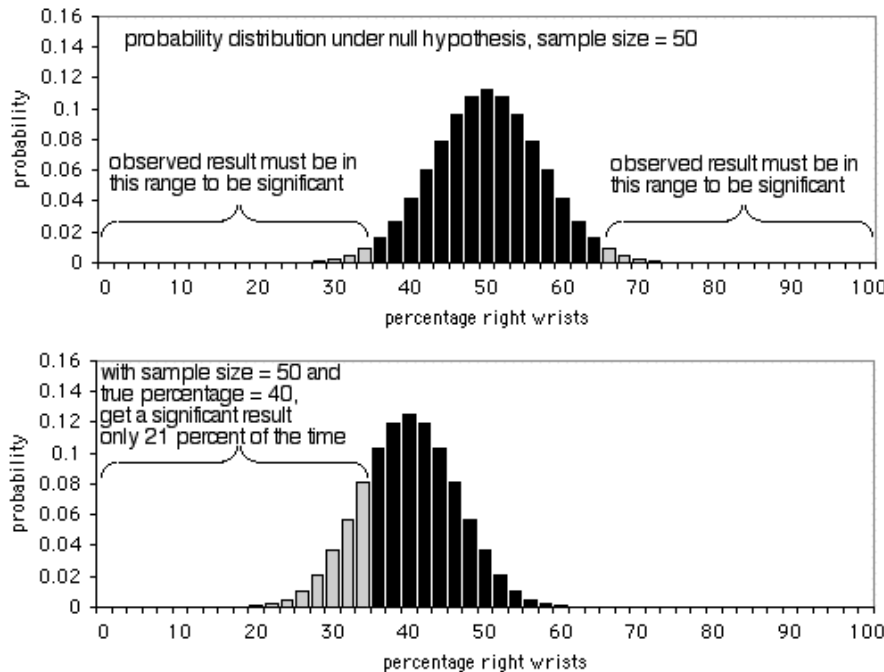
For measurement variables, you also need an estimate of the standard deviation. As standard deviation gets bigger, it gets harder to detect a significant difference, so you'll need a bigger sample size. Your estimate of the standard deviation can come from pilot experiments or from similar experiments 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.

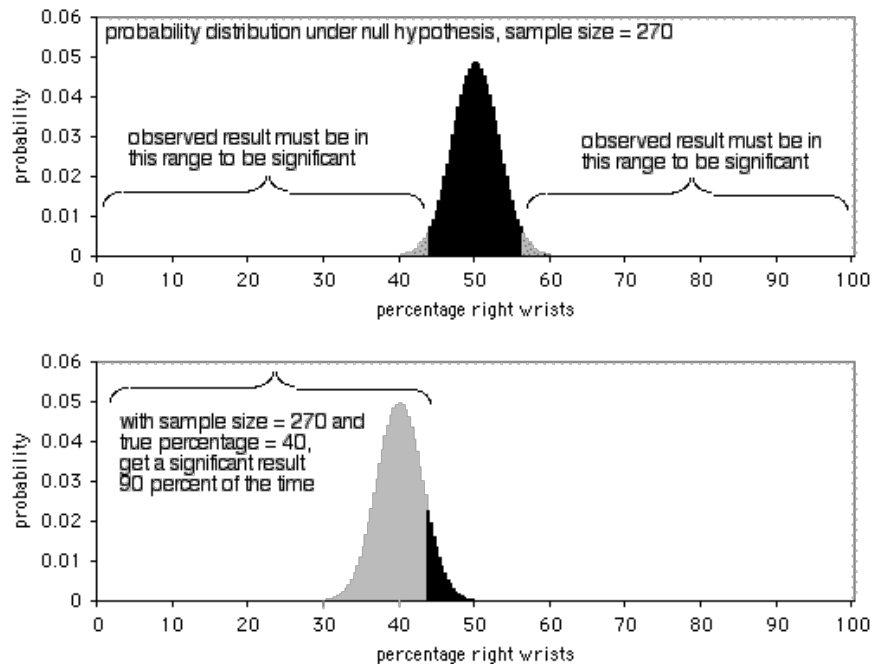
How it works

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%; if the percentage of people who break their right wrist is 60% or more, or 40% or less, you want to have a significant result from the exact binomial test. I have no idea why you picked 10%, but that's what you'll use. Alpha is 5%, as usual. You want power to be 90%, which means that if the percentage of broken right wrists really is 40% or 60%, you want a sample size that will yield a significant ($P < 0.05$) result 90% of the time, and a non-significant result (which would be a false negative in this case) only 10% of the time.



The first graph shows the probability distribution under the null hypothesis, with a sample size of 50 individuals. If the null hypothesis is true, you'll see less than 36% or more than 64% of people breaking their right wrists (a false positive) about 5% of the time. As the second graph shows, if the true percentage is 40%, the sample data will be less than 36 or more than 64% only 21% of the time; you'd get a true positive only 21% of the time, and a false negative 79% of the time. Obviously, a sample size of 50 is too small for this experiment; it would only yield a significant result 21% of the time, even if there's a 40:60 ratio of broken right wrists to left wrists.

POWER ANALYSIS



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% or more than 56.3% of people breaking their right wrists. As the second graph shows, if the true percentage is 40%, the sample data will be this extreme 90% of the time. A sample size of 270 is pretty good for this experiment; it would yield a significant result 90% of the time if there's a 40:60 ratio of broken right wrists to left wrists. If the ratio of broken right to left wrists is further away from 50:50, you'll have an even higher probability of getting a significant result.

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% more or fewer yellow peas than expected, with a power of 90%. You will test the data using the exact binomial test of goodness-of-fit if the sample size is small enough, or a G-test of goodness-of-fit if the sample size is larger. The power analysis is the same for both tests.

Using G*Power as described for the exact test of goodness-of-fit, the result is that it would take 2190 pea plants if you want to get a significant ($P < 0.05$) result 90% of the time, if the true proportion of yellow peas is 78 or 72%. That's a lot of peas, but you're reassured to see that it's not a ridiculous number. If you want to detect a difference of 0.1% between the expected and observed numbers of yellow peas, you can calculate that you'll need 1,970,142 peas; if that's what you need to detect, the sample size analysis tells you that you're going to have to include a pea-sorting robot in your budget.

The example data for the two-sample 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% chance of a difference this large being significant. Using G*Power as described on the two-sample t -test 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% of the time, if the true difference really is 2.0 inches.

How to do power analyses

G*Power

G*Power (www.gpower.hhu.de/) is an excellent free program, available for Mac and Windows, that will do power analyses for a large variety of tests. I will explain how to use G*Power for power analyses for each of the tests in this handbook.

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.

Cochran–Mantel–Haenszel test for repeated tests of independence

Use the Cochran–Mantel–Haenszel test when you have data from 2×2 tables that you’ve repeated at different times or locations. It will tell you whether you have a consistent difference in proportions across the repeats.

When to use it

Use the Cochran–Mantel–Haenszel test (which is sometimes called the Mantel–Haenszel test) for repeated tests of independence. The most common situation is that you have multiple 2×2 tables of independence; you’re analyzing the kind of experiment that you’d analyze with a test of independence, and you’ve done the experiment multiple times or at multiple locations. There are three nominal variables: the two variables of the 2×2 test of independence, and the third nominal variable that identifies the repeats (such as different times, different locations, or different studies). There are versions of the Cochran–Mantel–Haenszel test for any number of rows and columns in the individual tests of independence, but they’re rarely used and 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 legwarmers work in the winter but not in the summer, or maybe your first set of volunteers had worse arthritis than your second and third sets. In addition, pooling different studies together can show a “significant” difference in proportions when there isn’t one, or even show the opposite of a true difference. This is known as Simpson’s paradox. For these reasons, it’s better to analyze repeated tests of independence using the Cochran–Mantel–Haenszel test.

Null hypothesis

The null hypothesis is that the relative proportions of one variable are independent of the other variable within the repeats; in other words, there is no consistent difference in proportions in the 2×2 tables. For our 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 understand than odds ratios, so I'll put everything in terms of proportions. But if you're in a field such as epidemiology where this kind of analysis is common, you're probably going to have to think in terms of odds ratios.

How the test works

If you label the four numbers in a 2×2 test of independence like this:

$$\begin{array}{cc} a & b \\ c & d \end{array}$$

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

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

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 . You subtract the 0.5 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), which should make the P value more accurate. Some programs do the Cochran–Mantel–Haenszel test without the continuity correction, so be sure to specify whether you used it when reporting your results.

Assumptions

In addition to testing the null hypothesis, the Cochran-Mantel-Haenszel test also produces an estimate of the common odds ratio, a way of summarizing how big the effect is when pooled across the different repeats of the experiment. This requires assuming that the odds ratio is the same in the different repeats. You can test this assumption using the Breslow-Day test, which I'm not going to explain in detail; its null hypothesis is that the odds ratios are equal across the different repeats.

If some repeats have a big difference in proportion in one direction, and other repeats have a big difference in proportions but in the opposite direction, the Cochran-Mantel-Haenszel test may give a non-significant result. So when you get a non-significant Cochran-Mantel-Haenszel test, you should perform a test of independence on each 2×2 table separately and inspect the individual *P* values and the direction of difference to see whether something like this is going on. In our legwarmer example, if the proportion of people with ankle pain was much smaller for legwarmer-wearers in the winter, but much higher in the summer, and the Cochran-Mantel-Haenszel test gave a non-significant result, it would be erroneous to conclude that legwarmers had no effect. Instead, you could conclude that legwarmers had an effect, it just was different in the different seasons.

Examples

When you look at the back of someone's head, the hair either whorls clockwise or counterclockwise. Lauterbach and Knight (1927) compared the proportion of clockwise whorls in right-handed and left-handed children. With just this one set of people, you'd have two nominal variables (right-handed vs. left-handed, clockwise vs. counterclockwise), each with two values, so you'd analyze the data with Fisher's exact test.

However, several other groups have done similar studies of hair whorl and handedness (McDonald 2011):

Study group	Handedness	Right	Left
white children	Clockwise	708	50
	Counterclockwise	169	13
	percent CCW	19.3%	20.6%
British adults	Clockwise	136	24
	Counterclockwise	73	14
	percent CCW	34.9%	38.0%
Pennsylvania whites	Clockwise	106	32
	Counterclockwise	17	4
	percent CCW	13.8%	11.1%
Welsh men	Clockwise	109	22
	Counterclockwise	16	26
	percent CCW	12.8%	54.2%
German soldiers	Clockwise	801	102
	Counterclockwise	180	25

German children	percent CCW	18.3%	19.7%
	Clockwise	159	27
	Counterclockwise	18	13
	percent CCW	10.2%	32.5%
New York	Clockwise	151	51
	Counterclockwise	28	15
	percent CCW	15.6%	22.7%
American men	Clockwise	950	173
	Counterclockwise	218	33
	percent CCW	18.7%	16.0%

You could just add all the data together and do a test of independence on the 4463 total people, but it would be better to keep each of the 8 experiments separate. Some of the studies were done on children, while others were on adults; some were just men, while others were male and female; and the studies were done on people of different ethnic backgrounds. Pooling all these studies together might obscure important differences between them.

Analyzing the data using the Cochran-Mantel-Haenszel test, the result is $\chi^2_{MH}=6.07$, 1 d.f., $P=0.014$. Overall, left-handed people have a significantly higher proportion of counterclockwise whorls than right-handed people.

McDonald and Siebenaller (1989) surveyed allele frequencies at the *Lap* locus in the mussel *Mytilus trossulus* on the Oregon coast. At four estuaries, we collected mussels 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 we pooled all the other alleles 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.

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
	percent 94	58.3%	47.3%
Yaquina	94	61	257
	non-94	57	301
	percent 94	51.7%	46.1%
Alsea	94	73	65
	non-94	71	79
	percent 94	50.7%	45.1%
Umpqua	94	71	48
	non-94	55	48

percent 94	56.3%	50.0%
------------	-------	-------

The result is $\chi^2_{\text{MH}}=5.05$, 1 d.f., $P=0.025$. We can reject the null hypothesis that the proportion of *Lap^m* alleles is the same in the marine and estuarine locations.

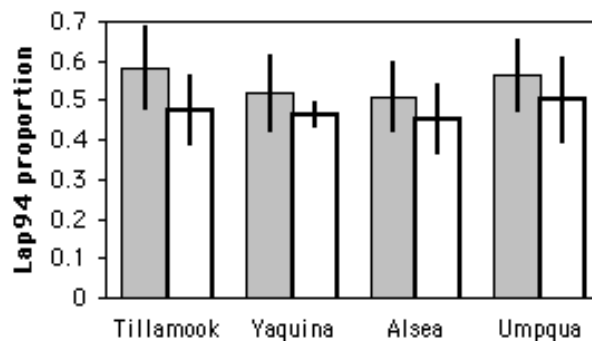
Duggal et al. (2010) did a meta-analysis of placebo-controlled studies of niacin and heart disease. They found 5 studies that met their criteria and looked for coronary artery revascularization in patients given either niacin or placebo:

Study		Revascularization	No revasc.	Percent revasc.
FATS	Niacin	2	46	4.2%
	Placebo	11	41	21.2%
AFREGS	Niacin	4	67	5.6%
	Placebo	12	60	16.7%
ARBITER 2	Niacin	1	86	1.1%
	Placebo	4	76	5.0%
HATS	Niacin	1	37	2.6%
	Placebo	6	32	15.8%
CLAS 1	Niacin	2	92	2.1%
	Placebo	1	93	1.1%

There are three nominal variables: niacin vs. placebo, revascularization vs. no revascularization, and the name of the study. The null hypothesis is that the rate of revascularization is the same in patients given niacin or placebo. The different studies have different overall rates of revascularization, probably because they used different patient populations and looked for revascularization after different lengths of time, so it would be unwise to just add up the numbers and do a single 2×2 test. The result of the Cochran-Mantel-Haenszel test is $\chi^2_{\text{MH}}=12.75$, 1 d.f., $P=0.00036$. Significantly fewer patients on niacin developed coronary artery revascularization.

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^m allele proportions (with 95% confidence intervals) in the mussel *Mytilus trossulus* at four bays in Oregon. Gray bars are marine samples and empty bars are estuarine samples.

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.

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 (www.biostat handbook.com/cmh.xls). It handles up to 50 2×2 tables. It gives you the choice of using or not using the continuity correction; the results are probably a little more accurate with the continuity correction. It does not do the Breslow-Day test.

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 must be listed first; in this case it is “location”.

```
DATA lap;
  INPUT location $ habitat $ allele $ count;
  DATALINES;
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
```

COCHRAN-MANTEL-HAENSZEL TEST

```
;
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 2×2 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.

The results also include the Breslow-Day test of homogeneity of odds ratios:

```
      Breslow-Day Test for
Homogeneity of the Odds Ratios
-----
Chi-Square          0.5295
DF                  3
Pr > ChiSq          0.9124
```

The Breslow-Day test for the example data shows no significant evidence for heterogeneity of odds ratios ($\chi^2=0.53$, 3 d.f., $P=0.91$).

References

- Cochran, W.G. 1954. Some methods for strengthening the common chi² tests. *Biometrics* 10: 417-451.
- Duggal, J.K., M. Singh, N. Attri, P.P. Singh, N. Ahmed, S. Pahwa, J. Molnar, S. Singh, S. Khosla and R. Arora. 2010. Effect of niacin therapy on cardiovascular outcomes in patients with coronary artery disease. *Journal of Cardiovascular Pharmacology and Therapeutics* 15: 158-166.
- Lauterbach, C.E., and J.B. Knight. 1927. Variation in whorl of the head hair. *Journal of Heredity* 18: 107-115.
- Mantel, N., and W. Haenszel. 1959. Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute* 22: 719-748.
- McDonald, J.H. 2011. *Myths of human genetics*. Sparky House Press, Baltimore.
- 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

A statistic of central tendency tells you where the middle of a set of measurements is. The arithmetic mean is by far the most common, but the median, geometric mean, and harmonic mean are sometimes useful.

Introduction

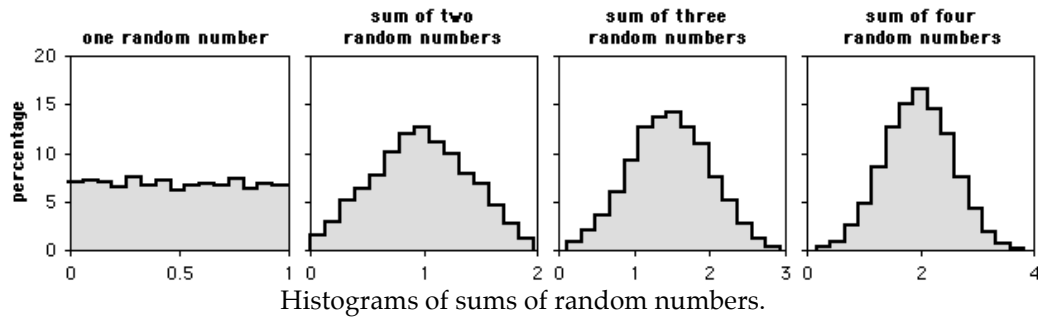
All of the tests in the first part of this handbook have analyzed nominal variables. You summarize data from a nominal variable as a percentage or a proportion. For example, 76.1% (or 0.761) of the peas in one of Mendel's genetic crosses were smooth, and 23.9% were wrinkled. If you have the percentage 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 described by just these two parameters, and the results of a parametric test may be inaccurate. In that case, the data can be converted to ranks and analyzed using a non-parametric test, which is less sensitive to deviations from normality.

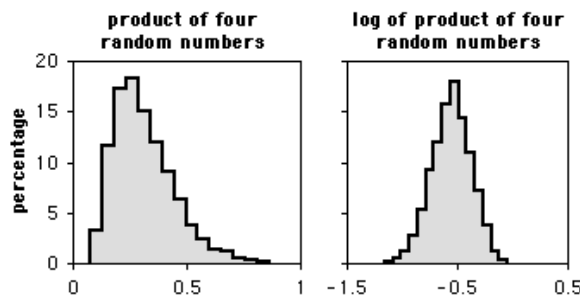
The normal distribution

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 with this same distribution.



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-slowness 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-slowness disease is not to subtract 10 cm/sec from the average speed, but instead to reduce the speed by 10% (in other words, multiply the 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, \dots is \bar{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% of trees are young trees, about 1 meter tall, that grew after a fire, and 1% 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 *N*th 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% one year, 8% the next year, and 4% the third year, its final height would be the initial height multiplied by $1.60 \times 1.08 \times 1.04 = 1.80$. 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 involving interest rates, etc., 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_i and Y_{i+1} . The median is given by the spreadsheet function MEDIAN(Y_i).

The median is useful when you are 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 not many people use them because of their lack of power, and I don't discuss them 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 measurements 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(Y_i).

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:

Stream	fish/75m
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 descriptive statistics spreadsheet that calculates the arithmetic, geometric and harmonic means, the median, and the mode, for up to 1000 observations (www.biostathandbook.com/descriptive.xls).

Web pages

This web page (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.

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;
  DATALINES;
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 mussels;
  INPUT species $ length width;
  DATALINES;
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 DATA=mussels;
  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.

References

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

Statistics of dispersion

A statistic of dispersion tells you how spread out a set of measurements is. Standard deviation is the most common, but there are others.

Introduction

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”). You use a statistic of dispersion to give a single number that describes how compact or spread out a set 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; if you were telling your Uncle Cletus about your research on the giant deep-sea isopod *Bathynomus giganteus*, you wouldn’t blather about means and standard deviations, you’d say they ranged from 4.4 to 36.5 cm long (Biornes-Fourzán and Lozano-Alvarez 1991). Then you’d explain that isopods are roly-polies, and 36.5 cm is about 14 American inches, and Uncle Cletus would finally be impressed, because a roly-poly that’s over a foot long is pretty impressive.

Range 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, or two samples from the same population could have very different ranges, purely by chance. In addition, the range increases as the sample size increases; the more observations you make, the greater the chance that you’ll sample a very large or very small value. There is no range function in spreadsheets; you can calculate the range by using $\text{=MAX}(Ys) - \text{MIN}(Ys)$, where Ys represents a set of cells.

Sum of squares: This is not really a statistic of dispersion by itself, but I mention it here because it forms the basis of the variance and standard deviation. Subtract the 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 $\text{DEVSQ}(Ys)$ (not by the function SUMSQ). You’ll probably never have a reason to calculate the sum of squares, but it’s an important concept.

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 parametric variance. You can only calculate the parametric variance of a population 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; I only mention it because there's a spreadsheet function for it *that you should never use*, VARP(Ys).

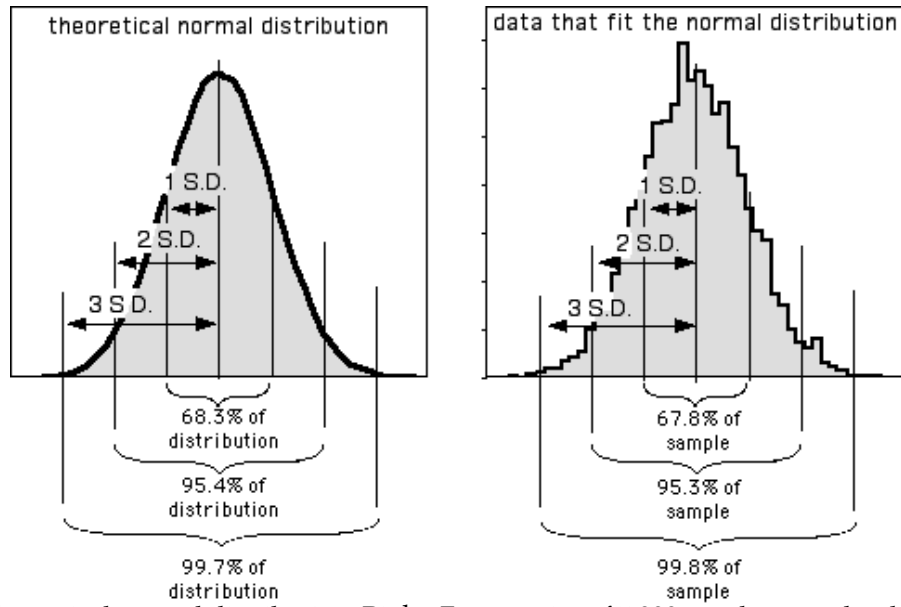
Sample variance: You almost always have a sample of observations that you are using to estimate a population parameter. To get an unbiased estimate of the population variance, divide the sum of squares by $n-1$, not by n . This sample variance, which is the one you will always use, is given by the spreadsheet 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, for most purposes 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; a set of volumes measured in cm^3 would have a variance expressed in cm^6 , which is even weirder. 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 never use; is given by the spreadsheet function STDEVP(Ys). The square root of the sample variance is given by the spreadsheet function STDEV(Ys). You should always use the sample standard deviation; from here on, when you see "standard deviation," it means the sample standard deviation.

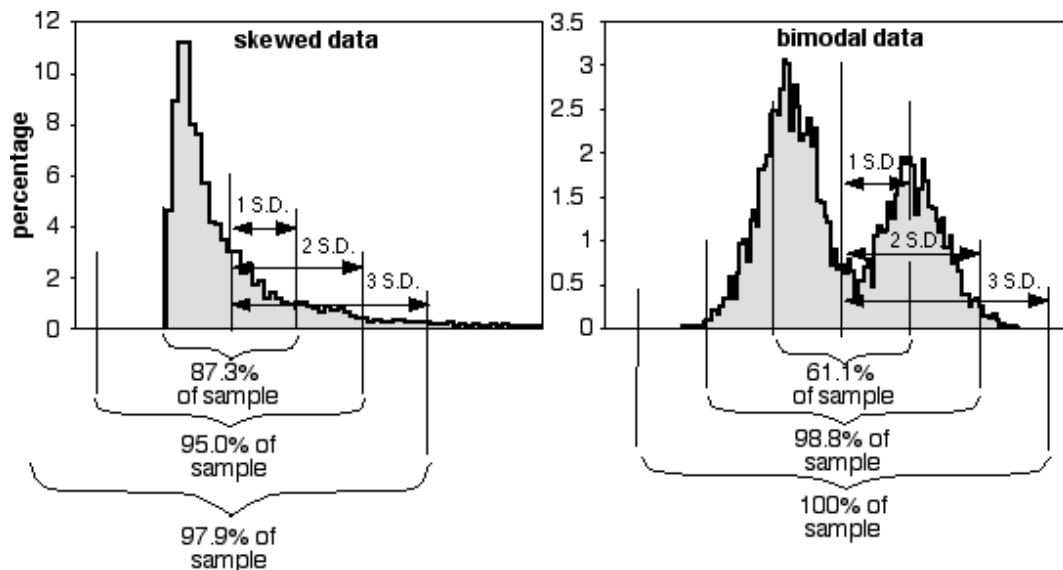
The square root of the sample variance actually underestimates the sample standard deviation by a little bit. Gurland and Tripathi (1971) came up with a correction factor that gives a more accurate estimate of the standard deviation, but very few people use it. Their correction factor makes the standard deviation about 3% bigger with a sample size of 9, and about 1% bigger with a sample size of 25, for example, and most people just don't need to estimate standard deviation that accurately. Neither SAS nor Excel uses the Gurland and Tripathi correction; I've included it as an option in my descriptive statistics spreadsheet. If you use the standard deviation with the Gurland and Tripathi correction, be sure to say this when you write up your results.

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 an understandable way. Many variables in biology fit the normal probability distribution fairly well. If a variable fits the normal distribution, 68.3% (or roughly two-thirds) of the values are within one standard deviation of the mean, 95.4% are within two standard deviations of the mean, and 99.7% (or almost all) are within 3 standard deviations of the mean. Thus if someone says the mean length of men's feet is 270 mm with a standard deviation of 13 mm, you know that about two-thirds of men's feet are between 257 and 283 mm long, and about 95% of men's feet are between 244 and 296 mm long. 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 for one variable among groups with different means, or among different measurement variables. For example, the United States military measured foot length and foot width in 1774 American men. The standard deviation of foot length was 13.1 mm and the standard deviation for foot width was 5.26 mm, which makes it seem as if foot length is more variable than foot width. However, feet

are longer than they are wide. Dividing by the means (269.7 mm for length, 100.6 mm for width), the coefficients of variation is actually slightly smaller for length (4.9%) than for width (5.2%), which for most purposes would be a more useful measure of variation.

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 (www.biostathandbook.com/descriptive.xls) that calculates the range, sample variance, sample standard deviation (with or without the Gurland and Tripathi correction), and coefficient of variation, for up to 1000 observations.

Web pages

This web page (graphpad.com/quickcalcs/CImean1.cfm) calculates standard deviation and other descriptive statistics for up to 10000 observations.

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

SAS

PROC UNIVARIATE will calculate the range, variance, standard deviation (without the Gurland and Tripathi correction), and coefficient of variation. It calculates the sample variance and sample standard deviation. For examples, see the central tendency web page.

Reference

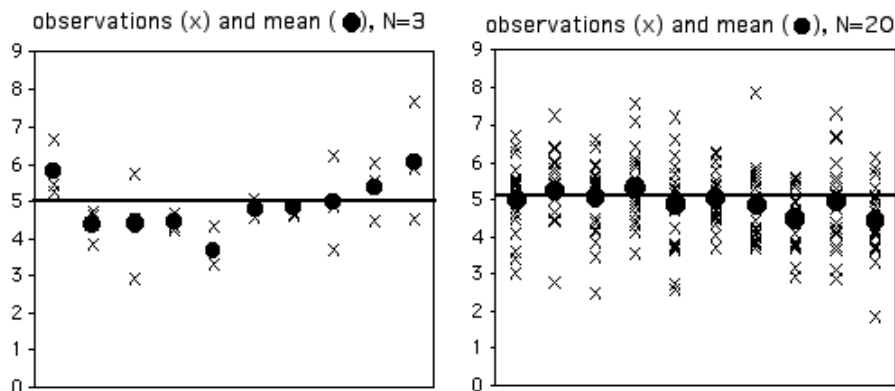
- Briones-Fourzán, P., and E. Lozano-Alvarez. 1991. Aspects of the biology of the giant isopod *Bathynomus giganteus* A. Milne Edwards, 1879 (Flabellifera: Cirolanidae), off the Yucatan Peninsula. *Journal of Crustacean Biology* 11: 375-385.
- Gurland, J., and R.C. Tripathi. 1971. A simple approximation for unbiased estimation of the standard deviation. *American Statistician* 25: 30-32.

Standard error of the mean

Standard error of the mean tells you how accurate your estimate of the mean is likely to be.

Introduction

When you take a sample of observations from a population and calculate the sample mean, you are estimating of the parametric mean, or mean of all of the individuals in the population. Your sample mean won't be exactly equal to the parametric mean that you're trying to estimate, and you'd like to have an idea of how close your sample mean is likely to be. 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 line is the parametric mean.



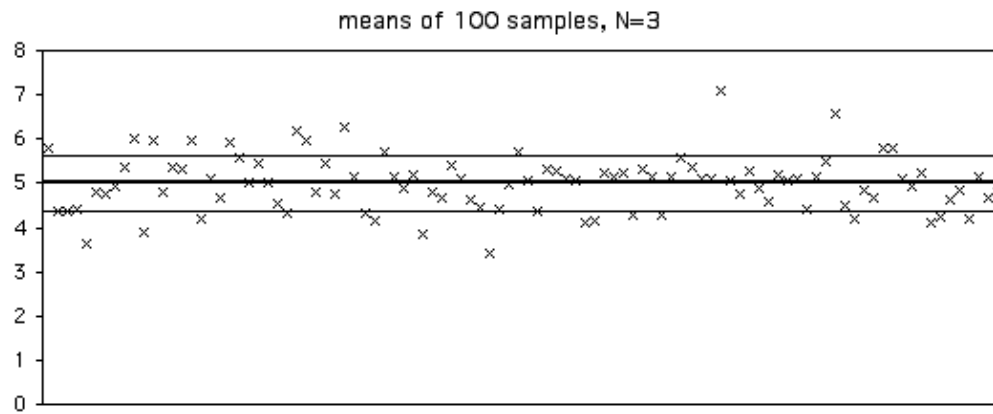
Individual observations (X's) and means (dots) for random samples from a population with a parametric mean of 5 (horizontal line).

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.

Once you've calculated the mean of a sample, you should let people know 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

STANDARD ERROR OF THE MEAN

parametric mean, 95.4% would be within two standard errors, and almost all (99.7%) would be within three standard errors.

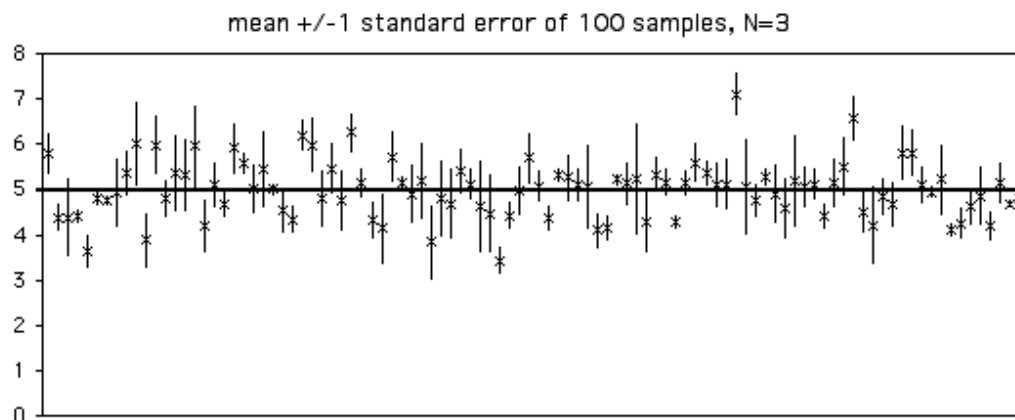


Means of 100 random samples (N=3) from a population with a parametric mean of 5 (horizontal line).

Here's a figure illustrating this. I took 100 samples of 3 from a population with a parametric mean of 5 (shown by the 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 \pm one standard error).

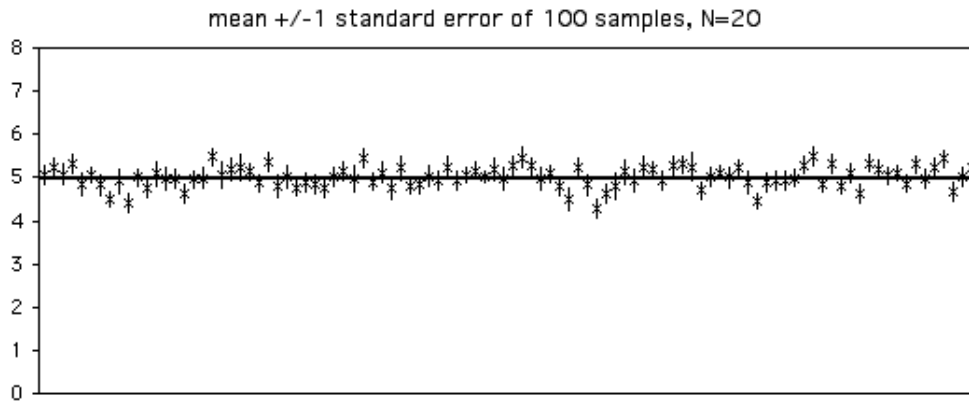
Usually you won't have multiple samples to use in making multiple estimates of the mean. Fortunately, you can estimate the standard error of the mean using the sample size and standard deviation of a single sample of observations. The standard error of the mean is estimated by the standard deviation 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, the standard error of the mean will become smaller. 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. Note that it's a function of the square root of the sample size; for example, to make the standard error half as big, you'll need four times as many observations.

"Standard error of the mean" and "standard deviation of the mean" are equivalent terms. People almost always say "standard error of the mean" 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, standard error of a regression coefficient, etc., you should specify standard error of the mean.

There is a myth that when two means have standard error bars that don't overlap, the means are significantly different (at the $P < 0.05$ level). This is not true (Browne 1979, Payton et al. 2003); it is easy for two sets of numbers to have standard error bars that don't overlap, yet not be significantly different by a two-sample t -test. Don't try to do statistical tests by visually comparing standard error bars, just use the correct statistical test.

Similar statistics

Confidence intervals and standard error of the mean serve the same purpose, to express the reliability of an estimate of the mean. When you look at scientific papers, sometimes the "error bars" on graphs or the \pm number after means in tables represent the standard error of the mean, while in other papers they represent 95% confidence intervals. I prefer 95% confidence intervals. When I see a graph with a bunch of points and error bars representing means and confidence intervals, I know that most (95%) of the error bars include the parametric means. When the error bars are standard errors of the mean, only about two-thirds of the error 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.

You use standard deviation and coefficient of variation to show how much variation there is among individual observations, while you use standard error or confidence intervals 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 \pm 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 (www.biostathandbook.com/descriptive.xls) calculates the standard error of the mean for up to 1000 observations, using the function `=STDEV(Ys)/SQRT(COUNT(Ys))`.

Web pages

This web page (graphpad.com/quickcalcs/CImean1.cfm) calculates standard error of the mean and other descriptive statistics for up to 10000 observations.

This web page (www.ruf.rice.edu/~lane/stat_analysis/descriptive.html) calculates standard error of the mean, along with other descriptive statistics. I don't know the maximum number of observations it can handle.

SAS

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

References

- Browne, R. H. 1979. On visual assessment of the significance of a mean difference. *Biometrics* 35: 657-665.
- Payton, M. E., M. H. Greenstone, and N. Schenker. 2003. Overlapping confidence intervals or standard error intervals: what do they mean in terms of statistical significance? *Journal of Insect Science* 3: 34.

Confidence limits

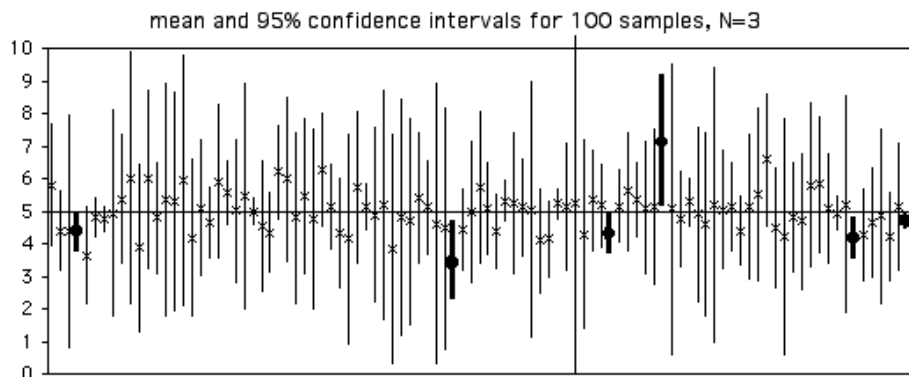
Confidence limits tell you how accurate your estimate of the mean is likely to be.

Introduction

After you've calculated the mean of a set of observations, you should give some indication of how close your estimate is likely to be to the parametric ("true") mean. One way to do this is with confidence limits. Confidence limits are the numbers at the upper and lower end of a confidence interval; for example, if your mean is 7.4 with confidence limits of 5.4 and 9.4, your confidence interval is 5.4 to 9.4.

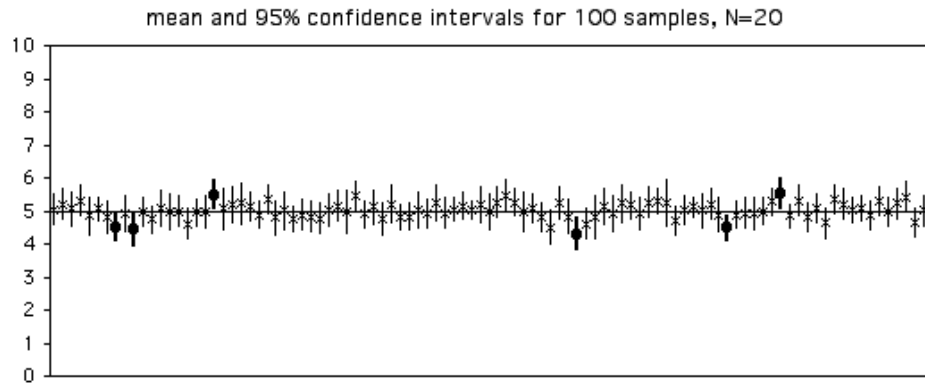
Most people use 95% confidence limits, although you could use other values. Setting 95% confidence limits means that if you took repeated random samples from a population and calculated the mean and confidence limits for each sample, the confidence interval for 95% of your samples would include the parametric 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.



CONFIDENCE LIMITS

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



When you calculate the confidence interval 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 this confusing, pedantic distinction, but others are very picky about it, 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

$$=(\text{STDEV}(\text{Ys})/\text{SQRT}(\text{COUNT}(\text{Ys}))) * \text{TINV}(0.05, \text{COUNT}(\text{Ys})-1)$$

where Ys is the range of cells containing your data. You add this value to and subtract it 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 and 97.3. You could also report this as “87 ±10.3 (95% confidence limits).” People report both confidence limits and standard errors 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 (statpages.org/confint.html). 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 \pm something," instead you'd have to say "0.10 with 95% confidence limits of 0.0124 and 0.3170."

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, if you calculate the confidence limits using the normal approximation on 0.10 with a sample size of 20, you get -0.03 and 0.23, which is ridiculous (you couldn't have less than 0% 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, more people use the confidence limits based on the normal approximation than use 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.

Statistical testing with confidence intervals

This handbook mostly presents "classical" or "frequentist" statistics, in which hypotheses are tested by estimating the probability of getting the observed results by chance, if the null is true (the P value). An alternative way of doing statistics is to put a confidence interval on a measure of the deviation from the null hypothesis. For example, rather than comparing two means with a two-sample t -test, some statisticians would calculate the confidence interval of the difference in the means.

This approach is valuable if a small deviation from the null hypothesis would be uninteresting, when you're more interested in the size of the effect rather than whether it exists. For example, if you're doing final testing of a new drug that you're confident will have some effect, you'd be mainly interested in estimating how well it worked, and how confident you were in the size of that effect. You'd want your result to be "This drug reduced systolic blood pressure by 10.7 mm Hg, with a confidence interval of 7.8 to 13.6," not "This drug significantly reduced systolic blood pressure ($P=0.0007$)."

Using confidence limits this way, as an alternative to frequentist statistics, has many advocates, and it can be a useful approach. However, I often see people saying things like “The difference in mean blood pressure was 10.7 mm Hg, with a confidence interval of 7.8 to 13.6; because the confidence interval on the difference does not include 0, the means are significantly different.” This is just a clumsy, roundabout way of doing hypothesis testing, and they should just admit it and do a frequentist statistical test.

There is a myth that when two means have confidence intervals that overlap, the means are not significantly different (at the $P < 0.05$ level). Another version of this myth is that if each mean is outside the confidence interval of the other mean, the means are significantly different. Neither of these is true (Schenker and Gentleman 2001, Payton et al. 2003); it is easy for two sets of numbers to have overlapping confidence intervals, yet still be significantly different by a two-sample t -test; conversely, each mean can be outside the confidence interval of the other, yet they’re still not significantly different. Don’t try compare two means by visually comparing their confidence intervals, just use the correct statistical test.

Similar statistics

Confidence limits and standard error of the mean serve the same purpose, to express the reliability of an estimate of the mean. When you look at scientific papers, sometimes the “error bars” on graphs or the \pm number after means in tables represent the standard error of the mean, while in other papers they represent 95% confidence intervals. I prefer 95% confidence intervals. When I see a graph with a bunch of points and error bars representing means and confidence intervals, I know that most (95%) of the error bars include the parametric means. When the error 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 $t \times 0.05$ 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. A surprising number of papers don’t say what their error bars represent, which means that the only information the error bars convey to the reader is that the authors are careless and sloppy.

Examples

Measurement data: The blacknose dace data from the central tendency web page has an arithmetic mean of 70.0. The lower confidence limit is 45.3 (70.0–24.7), and the upper confidence limit is 94.7 (70+24.7).

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 $\pm 30\%$, 10% , 3% , and 1% for $n=10$, 100 , 1000 , and $10,000$, respectively. This is why the “margin of error” in political polls, which typically have a sample size of around 1,000, is usually about 3%. Of course, this rough idea is no substitute for an actual power analysis.

How to calculate confidence limits

Spreadsheets

The descriptive statistics spreadsheet (www.biostathandbook.com/descriptive.xls) calculates 95% confidence limits of the mean for up to 1000 measurements. The confidence intervals for a binomial proportion spreadsheet (www.biostathandbook.com/confidence.xls) calculates 95% confidence limits for nominal variables, using both the exact binomial and the normal approximation.

Web pages

This web page (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 (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;
  datalines;
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 and 94.7.

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

CONFIDENCE LIMITS

```
data gus;
  input paw $;
  datalines;
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.

References

- Payton, M. E., M. H. Greenstone, and N. Schenker. 2003. Overlapping confidence intervals or standard error intervals: what do they mean in terms of statistical significance? *Journal of Insect Science* 3: 34.
- Schenker, N., and J. F. Gentleman. 2001. On judging the significance of differences by examining overlap between confidence intervals. *American Statistician* 55: 182-186.

Student's t -test for one sample

Use Student's t -test for one sample when you have one measurement variable and a theoretical expectation of what the mean should be under the null hypothesis. It tests whether the mean of the measurement variable is different from the null expectation.

Introduction

There are several statistical tests that use the t -distribution and can be called a t -test. One is Student's t -test for one sample, named after "Student," the pseudonym that William Gosset used to hide his employment by the Guinness brewery in the early 1900s (they had a rule that their employees weren't allowed to publish, and Guinness didn't want other employees to know that they were making an exception for Gosset). Student's t -test for one sample compares a sample to a theoretical mean. It has so few uses in biology that I didn't cover it in previous editions of this Handbook, but then I recently found myself using it (McDonald and Dunn 2013), so here it is.

When to use it

Use Student's t -test when you have one measurement variable, and you want to compare the mean value of the measurement variable to some theoretical expectation. It is commonly used in fields such as physics (you've made several observations of the mass of a new subatomic particle—does the mean fit the mass predicted by the Standard Model of particle physics?) and product testing (you've measured the amount of drug in several aliquots from a new batch—is the mean of the new batch significantly less than the standard you've established for that drug?). It's rare to have this kind of theoretical expectation in biology, so you'll probably never use the one-sample t -test.

I've had a hard time finding a real biological example of a one-sample t -test, so imagine that you're studying joint position sense, our ability to know what position our joints are in without looking or touching. You want to know whether people over- or underestimate their knee angle. You blindfold 10 volunteers, bend their knee to a 120° angle for a few seconds, then return the knee to a 90° angle. Then you ask each person to bend their knee to the 120° angle. The measurement variable is the angle of the knee, and the theoretical expectation from the null hypothesis is 120°. You get the following imaginary data:

Individual	Angle
A	120.6
B	116.4
C	117.2
D	118.1
E	114.1
F	116.9
G	113.3
H	121.1
I	116.9
J	117.0

If the null hypothesis were true that people don't over- or underestimate their knee angle, the mean of these 10 numbers would be 120. The mean of these ten numbers is 117.2; the one-sample t -test will tell you whether that is significantly different from 120.

Null hypothesis

The statistical null hypothesis is that the mean of the measurement variable is equal to a number that you decided on before doing the experiment. For the knee example, the biological null hypothesis is that people don't under- or overestimate their knee angle. You decided to move people's knees to 120°, so the statistical null hypothesis is that the mean angle of the subjects' knees will be 120°.

How the test works

Calculate the test statistic, t , using this formula:

$$t_s = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

where \bar{x} is the sample mean, μ is the mean expected under the null hypothesis, s is the sample standard deviation and n is the sample size. The test statistic, t , gets bigger as the difference between the observed and expected means gets bigger, as the standard deviation gets smaller, or as the sample size gets bigger.

Applying this formula to the imaginary knee position data gives a t -value of -3.69 .

You calculate the probability of getting the observed t value under the null hypothesis using the t -distribution. The shape of the t -distribution, and thus the probability of getting a particular t value, depends on the number of degrees of freedom. The degrees of freedom for a one-sample t -test is the total number of observations in the group minus 1. For our example data, the P value for a t value of -3.69 with 9 degrees of freedom is 0.005, so you would reject the null hypothesis and conclude that people return their knee to a significantly smaller angle than the original position.

Assumptions

The t -test assumes that the observations within each group are normally distributed. If the distribution is symmetrical, such as a flat or bimodal distribution, the one-sample t -test is not at all sensitive to the non-normality; you will get accurate estimates of the P value, even with small sample sizes. A severely skewed distribution can give you too

many false positives unless the sample size is large (above 50 or so). If your data are severely skewed and you have a small sample size, you should try a data transformation to make them less skewed. With large sample sizes (simulations I've done suggest 50 is large enough), the one-sample t -test will give accurate results even with severely skewed data.

Example

McDonald and Dunn (2013) measured the correlation of transferrin (labeled red) and Rab-10 (labeled green) in five cells. The biological null hypothesis is that transferrin and Rab-10 are not colocalized (found in the same subcellular structures), so the statistical null hypothesis is that the correlation coefficient between red and green signals in each cell image has a mean of zero. The correlation coefficients were 0.52, 0.20, 0.59, 0.62 and 0.60 in the five cells. The mean is 0.51, which is highly significantly different from 0 ($t=6.46$, 4 d.f., $P=0.003$), indicating that transferrin and Rab-10 are colocalized in these cells.

Graphing the results

Because you're just comparing one observed mean to one expected value, you probably won't put the results of a one-sample t -test in a graph. If you've done a bunch of them, I guess you could draw a bar graph with one bar for each mean, and a dotted horizontal line for the null expectation.

Similar tests

The paired t -test is a special case of the one-sample t -test; it tests the null hypothesis that the mean *difference* between two measurements (such as the strength of the right arm minus the strength of the left arm) is equal to zero. Experiments that use a paired t -test are much more common in biology than experiments using the one-sample t -test, so I treat the paired t -test as a completely different test.

The two-sample t -test compares the means of two different samples. If one of your samples is very large, you may be tempted to treat the mean of the large sample as a theoretical expectation, but this is incorrect. For example, let's say you want to know whether college softball pitchers have greater shoulder flexion angles than normal people. You might be tempted to look up the "normal" shoulder flexion angle (150°) and compare your data on pitchers to the normal angle using a one-sample t -test. However, the "normal" value doesn't come from some theory, it is based on data that has a mean, a standard deviation, and a sample size, and at the very least you should dig out the original study and compare your sample to the sample the 150° "normal" was based on, using a two-sample t -test that takes the variation and sample size of both samples into account.

How to do the test

Spreadsheets

I have set up a spreadsheet to perform the one-sample t -test (www.biostat handbook.com/onesamplettest.xls). It will handle up to 1000 observations.

Web pages

There are web pages to do the one-sample t -test (http://vassarstats.net/t_single.html and www.graphpad.com/quickcalcs/oneSampleT1/?Format=C).

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 joint position sense data above. Note that "H0" parameter for the theoretical value is "H" followed by the numeral zero, not a capital letter O.

```
DATA jps;
  INPUT angle;
  DATALINES;
120.6
116.4
117.2
118.1
114.1
116.9
113.3
121.1
116.9
117.0
;
PROC TTEST DATA=jps H0=50;
VAR angle;
RUN;
```

The output includes some descriptive statistics, plus the *t*-value and *P* value. For these data, the *P* value is 0.005.

DF	t Value	Pr > t
9	-3.69	0.0050

Power analysis

To estimate the sample size you need to detect a significant difference between a mean and a theoretical value, you need the following:

- the effect size, or the difference between the observed mean and the theoretical value that you hope to detect;
- the standard deviation;
- 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 G*Power program will calculate the sample size needed for a one-sample *t*-test. Choose "t tests" from the "Test family" menu and "Means: Difference from constant (one sample case)" from the "Statistical test" menu. Click on the "Determine" button and enter the theoretical value ("Mean H0") and a mean with the smallest difference from the theoretical that you hope to detect ("Mean H1"). Enter an estimate of the standard deviation. Click on "Calculate and transfer to main window". Change "tails" to two, set your alpha (this will almost always be 0.05) and your power (0.5, 0.8, or 0.9 are commonly used).

As an example, let's say you want to follow up the knee joint position sense study that I made up above with a study of hip joint position sense. You're going to set the hip angle to 70° (Mean H0=70) and you want to detect an over- or underestimation of this angle of 1°, so you set Mean H1=71. You don't have any hip angle data, so you use the standard

deviation from your knee study and enter 2.4 for SD. You want to do a two-tailed test at the $P < 0.05$ level, with a probability of detecting a difference this large, if it exists, of 90% ($1 - \beta = 0.90$). Entering all these numbers in G*Power gives a sample size of 63 people.

Reference

McDonald, J.H., and K.W. Dunn. 2013. Statistical tests for measures of colocalization in biological microscopy. *Journal of Microscopy* 252: 295-302.

Student's t -test for two samples

Use Student's t -test for two samples when you have one measurement variable and one nominal variable, and the nominal variable has only two values. It tests whether the means of the measurement variable are different in the two groups.

Introduction

There are several statistical tests that use the t -distribution and can be called a t -test. One of the most common is Student's t -test for two samples. Other t -tests include the one-sample t -test, which compares a sample mean to a theoretical mean, and the paired t -test.

Student's t -test for two samples is mathematically identical to a one-way anova with two categories; because comparing the means of two samples is such a common experimental design, and because the t -test is familiar to many more people than anova, I treat the two-sample t -test separately.

When to use it

Use the two-sample 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 , is calculated using a formula that has the difference between the means in the numerator; this makes t 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 gets larger as the means get farther apart, the variances get smaller, or the sample sizes increase.

You calculate the probability of getting the observed t value under the null hypothesis using the t -distribution. The shape of the t -distribution, and thus the probability of getting

a particular t 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. Fortunately, it is not at all sensitive to deviations from this assumption, if the distributions of the two groups are the same (if both distributions are skewed to the right, for example). I've done simulations with a variety of non-normal distributions, including flat, bimodal, and highly skewed, and the two-sample t -test always gives about 5% false positives, even with very small sample sizes. If your data are severely non-normal, you should still try to find a data transformation that makes them more normal, but don't worry if you can't find a good transformation or don't have enough data to check the normality.

If your data are severely non-normal, *and* you have different distributions in the two groups (one data set is skewed to the right and the other is skewed to the left, for example), *and* you have small samples (less than 50 or so), then the two-sample t -test can give inaccurate results, with considerably more than 5% false positives. A data transformation won't help you here, and neither will a Mann-Whitney U-test. It would be pretty unusual in biology to have two groups with different distributions but equal means, but if you think that's a possibility, you should require a P value much less than 0.05 to reject the null hypothesis.

The two-sample t -test also assumes homoscedasticity (equal variances in the two groups). If you have a balanced design (equal sample sizes in the two groups), the test is not very sensitive to heteroscedasticity unless the sample size is very small (less than 10 or so); the standard deviations in one group can be several times as big as in the other group, and you'll get $P < 0.05$ about 5% of the time if the null hypothesis is true. With an unbalanced design, heteroscedasticity is a bigger problem; if the group with the smaller sample size has a bigger standard deviation, the two-sample t -test can give you false positives much too often. If your two groups have standard deviations that are substantially different (such as one standard deviation is twice as big as the other), and your sample sizes are small (less than 10) or unequal, you should use Welch's t -test instead.

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'll 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; you will get the exact same P value from a two-sample t -test and from a one-way anova, even though you calculate the test statistics differently. The t -test is easier to do and is familiar to more people, but it is limited to just two categories of data. You can do a one-way anova on two or more categories. I recommend that if your research always involves comparing just two means, you should call your test a two-sample 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 the same thing.

The Mann-Whitney U-test is a non-parametric alternative to the two-sample t -test that some people recommend for non-normal data. However, if the two samples have the same distribution, the two-sample t -test is not sensitive to deviations from normality, so you can use the more powerful and more familiar t -test instead of the Mann-Whitney U-test. If the two samples have different distributions, the Mann-Whitney U-test is no better than the t -test. So there's really no reason to use the Mann-Whitney U-test unless you have a true ranked variable instead of a measurement variable.

If the variances are far from equal (one standard deviation is two or more times as big as the other) and your sample sizes are either small (less than 10) or unequal, you should use Welch's t -test (also known as Aspin-Welch, Welch-Satterthwaite, Aspin-Welch-Satterthwaite, or Satterthwaite t -test). It is similar to Student's t -test except that it does not assume that the standard deviations are equal. It is slightly less powerful than Student's t -test when the standard deviations are equal, but it can be much more accurate when the standard deviations are very unequal. My two-sample t -test spreadsheet (www.biostat handbook.com/twosamplettest.xls) will calculate Welch's t -test. You can also do Welch's t -test using this web page (graphpad.com/quickcalcs/ttest1.cfm), by clicking the button labeled "Welch's unpaired t -test".

Use the paired t -test 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.

Use the one-sample t -test when you have just one group, not two, and you are comparing the mean of the measurement variable for that group to a theoretical expectation.

How to do the test

Spreadsheets

I've set up a spreadsheet for two-sample t -tests (www.biostat handbook.com/twosamplettest.xls). It will perform either Student's t -test or Welch's t -test for up to 2000 observations in each group.

Web pages

There are web pages to do the t -test (graphpad.com/quickcalcs/ttest1.cfm and vassarstats.net/tu.html). Both will do both the Student's t -test and Welch's t -test.

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 @@;
  DATALINES;
2pm 69  2pm 70  2pm 66  2pm 63  2pm 68  2pm 70  2pm 69
2pm 67  2pm 62  2pm 63  2pm 76  2pm 59  2pm 62  2pm 62
2pm 75  2pm 62  2pm 72  2pm 63
5pm 68  5pm 62  5pm 67  5pm 68  5pm 69  5pm 67  5pm 61
5pm 59  5pm 62  5pm 61  5pm 69  5pm 66  5pm 62  5pm 62
5pm 61  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 labeled "Pooled", and the P value for Welch's t -test is on the line labeled "Satterthwaite." For these data, the P value is 0.2067 for Student's t -test and 0.1995 for Welch's.

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 expensive space mouse ($N_1/N_2=3.0$).

The G*Power program will calculate the sample size needed for a two-sample *t*-test. Choose "t tests" from the "Test family" menu and "Means: Difference between two independent means (two groups)" from the "Statistical test" menu. Click on the "Determine" button and enter the means and standard deviations you expect for each group. Only the difference between the group means is important; it is your effect size. Click on "Calculate and transfer to main window". Change "tails" to two, set your alpha (this will almost always be 0.05) and your power (0.5, 0.8, or 0.9 are commonly used). 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 want to know whether people who run regularly have wider feet than people who don't run. You look for previously published data on foot width and find the ANSUR data set, which shows a mean foot width for American men of 100.6 mm and a standard deviation of 5.26 mm. You decide that you'd like to be able to detect a difference of 3 mm in mean foot width between runners and non-runners. Using G*Power, you enter 100 mm for the mean of group 1, 103 for the mean of group 2, and 5.26 for the standard deviation of each group. You decide you want to detect a difference of 3 mm, at the $P<0.05$ level, with a probability of detecting a difference this large, if it exists, of 90% ($1-\beta=0.90$). Entering all these numbers in G*Power gives a sample size for each group of 66 people.

Independence

Most statistical tests assume that you have a sample of independent observations, meaning that the value of one observation does not affect the value of other observations. Non-independent observations can make your statistical test give too many false positives.

Measurement variables

One of the assumptions of most tests is that the observations are independent of each other. This assumption is violated when the value of one observation tends to be too similar to the values of other observations. For example, let's say you wanted to know whether calico cats had a different mean weight than black cats. You get five calico cats, five black cats, weigh them, and compare the mean weights with a two-sample t -test. If the five calico cats are all from one litter, and the five black cats are all from a second litter, then the measurements are not independent. Some cat parents have small offspring, while some have large; so if Josie the calico cat is small, her sisters Valerie and Melody are not independent samples of all calico cats, they are instead also likely to be small. Even if the null hypothesis (that calico and black cats have the same mean weight) is true, your chance of getting a P value less than 0.05 could be much greater than 5%.

A common source of non-independence is that observations are close together in space or time. For example, let's say you wanted to know whether tigers in a zoo were more active in the morning or the evening. As a measure of activity, you put a pedometer on Sally the tiger and count the number of steps she takes in a one-minute period. If you treat the number of steps Sally takes between 10:00 and 10:01 a.m. as one observation, and the number of steps between 10:01 and 10:02 a.m. as a separate observation, these observations are not independent. If Sally is sleeping from 10:00 to 10:01, she's probably still sleeping from 10:01 to 10:02; if she's pacing back and forth between 10:00 and 10:01, she's probably still pacing between 10:01 and 10:02. If you take five observations between 10:00 and 10:05 and compare them with five observations you take between 3:00 and 3:05 with a two-sample t -test, there's a good chance you'll get five low-activity measurements in the morning and five high-activity measurements in the afternoon, or vice-versa. This increases your chance of a false positive; if the null hypothesis is true, lack of independence can give you a significant P value much more than 5% of the time.

There are other ways you could get lack of independence in your tiger study. For example, you might put pedometers on four other tigers—Bob, Janet, Ralph, and Loretta—in the same enclosure as Sally, measure the activity of all five of them between 10:00 and 10:01, and treat that as five separate observations. However, it may be that when one tiger gets up and starts walking around, the other tigers are likely to follow it around and see what it's doing, while at other times all five tigers are likely to be resting. That would mean that Bob's amount of activity is not independent of Sally's; when Sally is more active, Bob is likely to be more active.

Regression and correlation assume that observations are independent. If one of the measurement variables is time, or if the two variables are measured at different times, the

data are often non-independent. For example, if I wanted to know whether I was losing weight, I could weigh myself every day and then do a regression of weight vs. day. However, my weight on one day is very similar to my weight on the next day. Even if the null hypothesis is true that I'm not gaining or losing weight, the non-independence will make the probability of getting a P value less than 0.05 much greater than 5%.

I've put a more extensive discussion of independence on the regression/correlation page.

Nominal variables

Tests of nominal variables (independence or goodness-of-fit) also assume that individual observations are independent of each other. To illustrate this, let's say I want to know whether my statistics class is more boring than my evolution class. I set up a video camera observing the students in one lecture of each class, then count the number of students who yawn at least once. In statistics, 28 students yawn and 15 don't yawn; in evolution, 6 yawn and 50 don't yawn. It seems like there's a significantly ($P=2.4 \times 10^{-8}$) higher proportion of yawners in the statistics class, but that could be due to chance, because the observations within each class are not independent of each other. Yawning is contagious (so contagious that you're probably yawning right now, aren't you?), which means that if one person near the front of the room in statistics happens to yawn, other people who can see the yawner are likely to yawn as well. So the probability that Ashley in statistics yawns is not independent of whether Sid yawns; once Sid yawns, Ashley will probably yawn as well, and then Megan will yawn, and then Dave will yawn.

Solutions for lack of independence

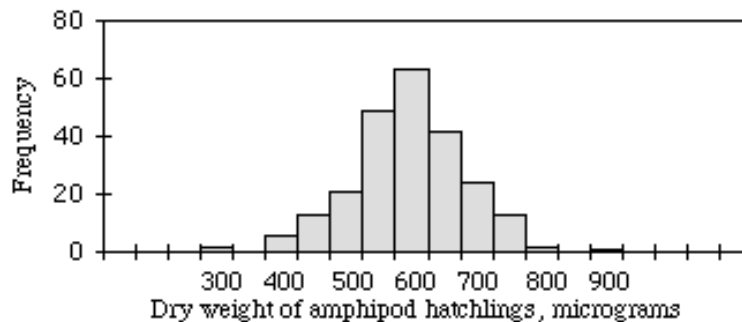
Unlike non-normality and heteroscedasticity, it is not easy to look at your data and see whether the data are non-independent. You need to understand the biology of your organisms and carefully design your experiment so that the observations will be independent. For your comparison of the weights of calico cats vs. black cats, you should know that cats from the same litter are likely to be similar in weight; you could therefore make sure to sample only one cat from each of many litters. You could also sample multiple cats from each litter, but treat "litter" as a second nominal variable and analyze the data using nested anova. For Sally the tiger, you might know from previous research that bouts of activity or inactivity in tigers last for 5 to 10 minutes, so that you could treat one-minute observations made an hour apart as independent. Or you might know from previous research that the activity of one tiger has no effect on other tigers, so measuring activity of five tigers at the same time would actually be okay. To really see whether students yawn more in my statistics class, I should set up partitions so that students can't see or hear each other yawning while I lecture.

For regression and correlation analyses of data collected over a length of time, there are statistical tests developed for time series. I don't cover them in this handbook; if you need to analyze time series data, find out how other people in your field analyze similar data.

Normality

Most tests for measurement variables assume that data are normally distributed (fit a bell-shaped curve). Here I explain how to check this and what to do if the data aren't normal.

Introduction



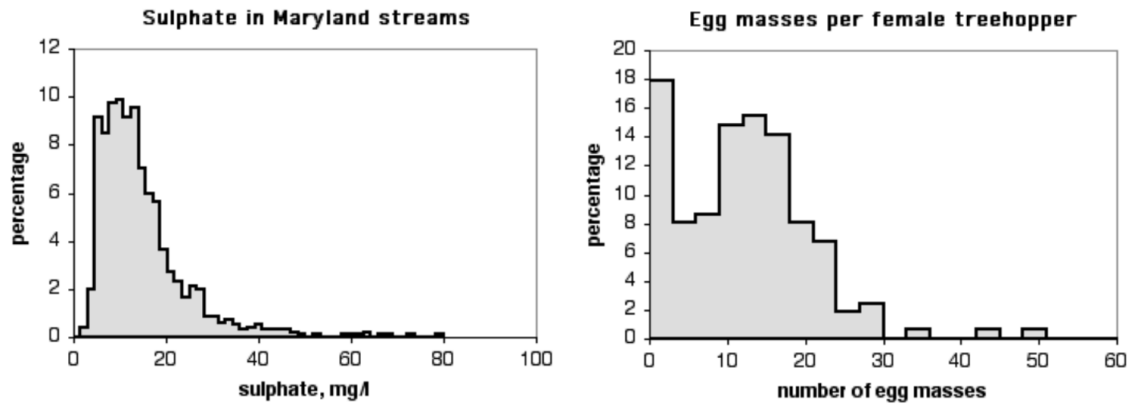
Histogram of dry weights of the amphipod crustacean *Platorchestia platensis*.

A probability distribution specifies the probability of getting an observation in a particular range of values; the normal distribution is the familiar bell-shaped curve, with a high probability of getting an observation near the middle and lower probabilities as you get further from the middle. A normal distribution can be completely described by just two numbers, or parameters, the mean and the standard deviation; all normal distributions with the same mean and same standard deviation will be exactly the same shape. One of the assumptions of an anova and other tests for measurement variables is that the data fit the normal probability distribution. Because these tests assume that the data can be described by two parameters, the mean and standard deviation, they are called parametric tests.

When you plot a frequency histogram of measurement data, the frequencies should approximate the bell-shaped normal distribution. For example, the figure shown at the right is a histogram of dry weights of newly hatched amphipods (*Platorchestia platensis*), data I tediously collected for my Ph.D. research. It fits the normal distribution pretty well.

Many biological variables fit the normal distribution quite well. This is a result of the central limit theorem, which says that when you take a large number of random numbers, the means of those numbers are approximately normally distributed. If you think of a variable like weight as resulting from the effects of a bunch of other variables averaged together—age, nutrition, disease exposure, the genotype of several genes, etc.—it's not surprising that it would be normally distributed.

NORMALITY



Two non-normal histograms.

Other data sets don't fit the normal distribution very well. The histogram on the left is the level of sulphate in Maryland streams (data from the Maryland Biological Stream Survey, www.dnr.state.md.us/streams/MBSS.asp). 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 right is the number of egg masses laid by individuals 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.

Parametric tests assume that your 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 a test that assumes normality.

What to do about non-normality

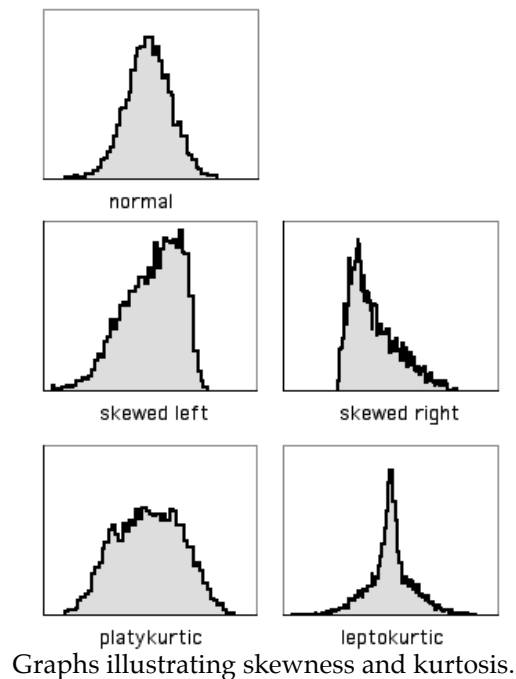
Once you have collected a set of measurement data, you should look at the frequency histogram to see if it looks non-normal. There are statistical tests of the goodness-of-fit of a data set to the normal distribution, but I don't recommend them, because many data sets that are significantly non-normal would be perfectly appropriate for an anova or other parametric test. 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 another result of the central limit theorem, which says that 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.

Because parametric tests are not very sensitive to deviations from normality, I recommend that you don't worry about it unless your data appear very, very non-normal to you. This is a subjective judgement on your part, but there don't seem to be any objective rules on how much non-normality is too much for a parametric test. You should look at what other people in your field do; if everyone transforms the kind of data you're collecting, or uses a non-parametric test, you should consider doing what everyone else does even if the non-normality doesn't seem that bad to you.

If your histogram looks like a normal distribution that has been pushed to one side, like the sulphate data above, you should try different data transformations to see if any of them make the histogram look more normal. It's best if you collect some data, check the normality, and decide on a transformation before you run your actual experiment; you don't want cynical people to think that you tried different transformations until you found one that gave you a significant result for your experiment.

If your data still look severely non-normal no matter what transformation you apply, it's probably still okay to analyze the data using a parametric test; they're just not that sensitive to non-normality. However, you may want to analyze your data 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/correlation. 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 will not be any better than a parametric one.

Skewness and kurtosis



Graphs illustrating 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 a parametric test 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.

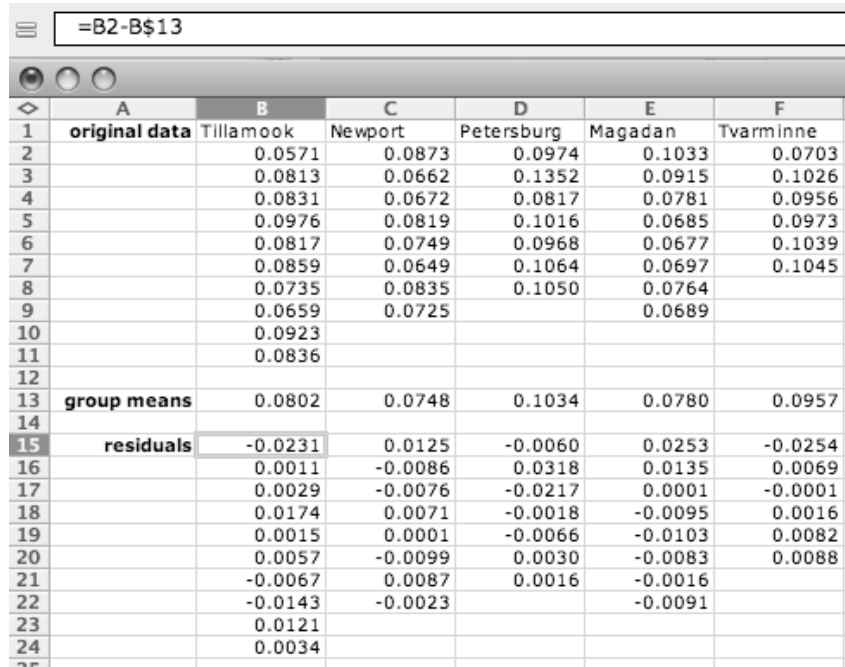
How to look at normality

Spreadsheet

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

NORMALITY

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.



	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				
25						

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 histogram calculator at www.shodor.com/interactivate/activities/Histogram/ is the best I've found.

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.

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. *Review of Educational Research* 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. *Journal of Educational Statistics* 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. *Review of Educational Research* 66: 579-619.

Homoscedasticity and heteroscedasticity

Parametric tests assume that data are homoscedastic (have the same standard deviation in different groups). Here I explain how to check this and what to do if the data are heteroscedastic (have different standard deviations in different groups).

Introduction

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

To illustrate this problem, I did 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 were homoscedastic (had the same standard deviation), the one-way anova on the simulated data sets were significant ($P < 0.05$) about 5% of the time, as they should be. However, when I made the standard deviations different (1.0 for population A, 2.0 for population B, and 3.0 for population C), I got a P value less than 0.05 in about 18% of the simulations. In other words, even though the population means were really all the same, my chance of getting a false positive result was 18%, not the desired 5%.

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. Heteroscedasticity is much less of a problem when you have a balanced design (equal sample sizes in each group). Early results suggested that heteroscedasticity was not a problem at all with a balanced design (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 of heteroscedasticity is much worse when the sample sizes are unequal (an unbalanced design) and the smaller samples are from populations with larger standard deviations; but when the smaller samples are from populations with smaller standard deviations, the false positive rate can actually be much less than 0.05, meaning the power of the test is reduced (Glass et al. 1972).

What to do about heteroscedasticity

You should always compare the standard deviations of different groups of measurements, to see if they are very different from each other. However, 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 you should not use a test that assumes homoscedasticity.

If you see a big difference in standard deviations between groups, the first things you should try are data transformations. A common pattern is that groups with larger means also have larger standard deviations, and a log or square-root transformation will often fix this problem. It's best if you can choose a transformation based on a pilot study, before you do your main experiment; you don't want cynical people to think that you chose a transformation because it gave you a significant result.

If the standard deviations 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 alternative to one-way anova is Welch's anova, 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. This means that non-parametric tests 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, or that a transformation will fix it.

Bartlett's test

There are several statistical tests for homoscedasticity, and the most popular is Bartlett's test. Use this test when you have one measurement variable, one nominal variable, and you want to test the null hypothesis that the standard deviations of the measurement variable are the same for the different groups.

Bartlett's test is not a particularly good one, because it is sensitive to departures from normality as well as heteroscedasticity; you shouldn't panic just because you have a significant Bartlett's test. It may be more helpful to use Bartlett's test to see what effect different transformations have on the heteroscedasticity; you can choose the transformation with the highest (least significant) *P* value for Bartlett's test. An alternative to Bartlett's test that I won't cover here is Levene's test. It is less sensitive to departures from normality, but if the data are approximately normal, it is less powerful than Bartlett's test.

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 equality of standard deviations 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 standard deviations.

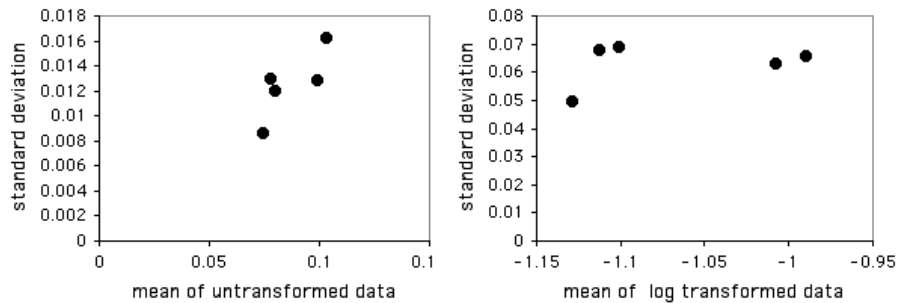
How to do Bartlett's test

Spreadsheet

I have put together a spreadsheet that performs Bartlett's test for homogeneity of standard deviations for up to 1000 observations in each of up to 50 groups (www.biostathandbook.com/bartletts.xls). 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 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 that will handle up to 14 groups (home.ubalt.edu/ntsbarsh/Business-stat/otherapplets/BartletTest.htm). You have to enter the variances (not standard deviations) 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;
```

References

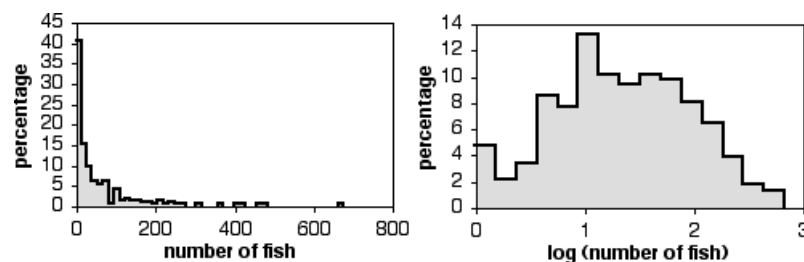
- Glass, G.V., P.D. Peckham, and J.R. Sanders. 1972. Consequences of failure to meet assumptions underlying fixed effects analyses of variance and covariance. *Review of Educational Research* 42: 237-288.
- Harwell, M.R., E.N. Rubinstein, W.S. Hayes, and C.C. Olds. 1992. Summarizing Monte Carlo results in methodological research: the one- and two-factor fixed effects ANOVA cases. *Journal of Educational Statistics* 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. *Review of Educational Research* 66: 579-619.

Data transformations

If a measurement variable does not fit a normal distribution or has greatly different standard deviations in different groups, you should try a data transformation.

Introduction

Many biological variables do not meet the assumptions of parametric statistical tests: they are not normally distributed, the standard deviations 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.

Here are 12 numbers from the 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. If you were comparing the fish abundance in different watersheds, and you decided that log transformation was the best, you would do a one-way anova on the logs of fish abundance, and you would test the null hypothesis that the means of the log-transformed abundances were equal.

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% 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 interval is not 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.

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 your use of data transformations.

There are an infinite number of transformations you could use, but it is better to use a transformation that other researchers commonly use in your field, such as the square-root transformation for count data or the log transformation for size data. Even if an obscure transformation that not many people have heard of gives you slightly more normal or

more homoscedastic data, it will probably be better to use a more common transformation so people don't get suspicious. Remember that your data don't have to be perfectly normal and homoscedastic; parametric tests aren't extremely sensitive to deviations from their assumptions.

It is also important that you decide which transformation to use before you do the statistical test. Trying different transformations until you find one that gives you a significant result is cheating. 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, you should use whatever transformation people in your field routinely use for your variable. For example, if you're studying pollen dispersal distance and other people routinely log-transform it, you should log-transform pollen distance too, even if you only have 10 observations and therefore can't really look at normality with a histogram.

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 $2.303... \times$ 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 the mean of your base-10 log-transformed data is 1.43, the back transformed mean is $10^{1.43}=26.9$ (in a spreadsheet, " $=10^{1.43}$ "). If the mean of your base- e log-transformed data is 3.65, the back transformed mean is $e^{3.65}=38.5$ (in a spreadsheet, " $=EXP(3.65)$ "). 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 $\text{nitrogen} \times \text{water} \times \text{sunlight} \times \text{insects}$, 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.

People often use the square-root transformation 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 0 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 transform 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 standard deviation 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 labeled "Paste All" and "Formulas," and check the box labeled "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.