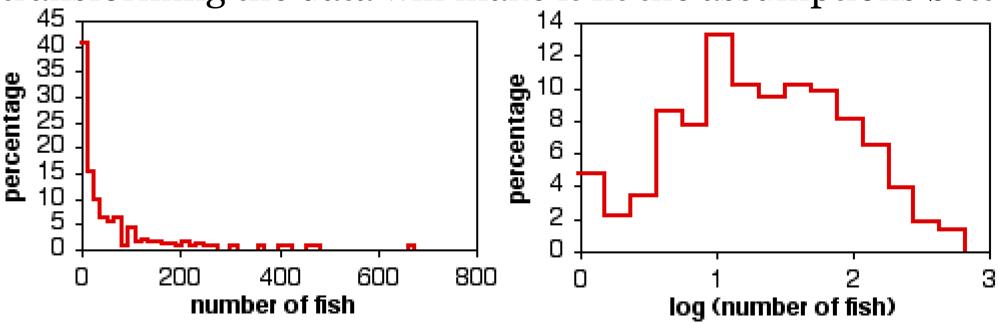
Data transformations

Summary

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.

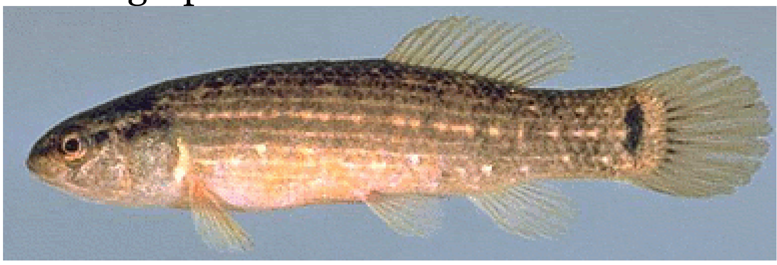
Introduction

Many biological variables do not meet the assumptions of parametric statistical tests: they are notnormally 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.



Eastern mudminnow (*Umbra pygmaea*).

Here are 12 numbers from the mudminnow data set; the first column is the untransformed data, the second column is the square root of the number in the first column, and the third column is the base-10 logarithm of the number in the first column.

Untransformed	Square-root transformed	Log transformed
38	6.164	1.580
1	1.000	0.000
13	3.606	1.114
2	1.414	0.301
13	3.606	1.114
20	4.472	1.301
50	7.071	1.699
9	3.000	0.954
28	5.292	1.447
6	2.449	0.778

4	2.000	0.602
43	6.557	1.633

You do the statistics on the transformed numbers. For example, the mean of the untransformed data is 18.9; the mean of the square-root transformed data is 3.89; the mean of the log transformed data is 1.044. 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 thenormality 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...× 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 nitrogen×water×sunlight×insects, and mathematically, this kind of function turns out to be log-normal.

Square-root transformation. This consists of taking the square root of each observation. The back transformation is to square the number. If you have negative numbers, you can't take the square root; you should add a constant to each number to make them all positive.

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.

```
DATA mudminnow;
   INPUT location $ banktype $ count;
   countlog=log10(count);
   countsqrt=sqrt(count);
   DATALINES;
Gwynn_1       forest 38
Gwynn_2       urban   1
Gwynn_3       urban  13
Jones_1       urban   2
Jones_2       forest 13
LGunpowder_1  forest 20
LGunpowder_2  field  50
LGunpowder_3  forest  9
BGunpowder_1  forest 28
BGunpowder_2  forest  6
BGunpowder_3  forest  4
BGunpowder_4  field  43
;
```

The dataset "mudminnow" contains all the original variables ("location", "banktype" and "count") plus the new variables ("countlog" and "countsqrt"). You then run whatever PROC you want and analyze these variables just like you would any others. Of course, this example does two different transformations only as an illustration; in reality, you should decide on one transformation before you analyze your data.

The SAS function for arcsine-transforming X is ARSIN(SQRT(X)).

You'll probably find it easiest to backtransform using a spreadsheet or calculator, but if you really want to do everything in SAS, the function for taking 10 to the X power is 10**X; the function for taking *e* to a power is EXP(X); the function for squaring X is X**2; and the function for backtransforming an arcsine transformed number is SIN(X)**2.

Reference

Picture of a mudminnow from The Virtual Aquarium.

⇐ Previous topic|Next topic ⇒

This page was last revised December 18, 2015. Its address is http://www.biostathandbook.com/transformation.html. It may be cited as:
McDonald, J.H. 2014. Handbook of Biological Statistics (3rd ed.). Sparky House Publishing, Baltimore, Maryland. This web page contains the content of pages 140-144 in the printed version.