Quantitative Understanding in Biology Short Course
Session 2
Data Transformations; p-values and Formal Statistical Tests

Jason Banfelder
Luce Skrabanek

March 18th, 2025

1 Review

In our last class, we learned:

- We can put bounds on the plausible size of an effect of some treatment or difference between two groups by computing a confidence interval.
- We typically use 95% CIs to define "plausibility", but this threshold is arbitrary, and different cutoffs can be used.
- If the plausible range of CIs includes "no effect," then we must admit the possibility that there is no difference between our groups.
- If the measurements from each of two groups are normally distributed, we can use a t-test to compute the CI of the difference of the group means.
- Paired studies have much more statistical power than unpaired studies; mechanically, we use an unpaired or paired t-test, as appropriate to the study design.

2 Introduction

An important theoretical requirement of using a t-test is that the measurements are samples from a normal distribution. We mentioned that this is almost always strictly not the case, because it would imply that negative measurements, as well as arbitrarily large values, are possible; for many measurements (e.g., heights of people, counts of cells), this makes no sense.

In general, if your data are not normally distributed, you have a few choices:

- Use statistical tests like the t-test anyway; many are quite robust and work well, even for moderate deviations from normality. However, some tests are not, and you may want to research the robustness of a particular method before using it.
- Transform your data so they are normally distributed, and analyze the transformed data.
- Use a non-parametric test. These are usually based on rank-order of data, rather than the actual numerical values themselves. Non-parametric tests are usually substantially less powerful than their parametric counterparts. This means that you have a higher chance of your analysis being inconclusive, and may need to increase n in your experimental design to get good results.
- Use tests specifically designed for a particular non-normal distribution. These can be difficult to find and apply correctly. We'll see one common example in the next session, but, in general, it may be wise to work with a professional statistician if you need to go down this road.

3 Transforming Data

When your data are clearly not normally distributed, one option is to transform the data so that they are. The ultimate purpose of a data transformation is usually not to slightly adjust the shape of the distribution so that a histogram looks a bit more like the bell curve, but rather to take account of some fundamental property of the data being collected.

One such fundamental property of data is how differences in measured values are interpreted. For example, when we are measuring heights of people, we consider the one inch difference between 5'8" and 5'9" to be of about the same importance as the difference between 6'2" and 6'3". This idea is baked into a t-test, and if it doesn't hold, a t-test is probably not appropriate for your data.

We shall consider three common cases where this idea is not applicable.

3.1 Measurements of Fractions or Percentages

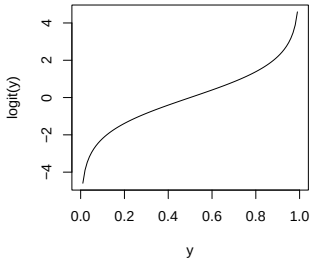
When observing fractions of things, a difference in the observed fraction usually has a different interpretation, depending on where in the range you are. For example, if 48.3% of the cells in your culture undergo apoptosis in one replicate, but 48.8% undergo apoptosis in another, you'd probably not think the observed 0.5% difference important. However, if the data were 0.01% of cells in one culture, and 0.51% of cells in the other, you'd probably consider that a huge difference even though the absolute difference is also 0.5%. After all, the second culture shows about 50× the rate of apoptosis! Similarly, if the data were

99.99% in the first culture, and 99.49% in the second, you'd probably consider the difference huge as well since there is 50× increase in cells not undergoing apoptosis.

Realize that, when measuring fractions, all measurements are bound to the range of zero to one. A common way to deal with the different interpretations of changes at the edges vs. in the middle of this range is to rescale (non-linearly) the range so that deltas have a consistent import regardless of where on the scale you are. This can be achieved with the so-called logit transformation:

$$\text{logit}(y) = \ln\left(\frac{y}{1-y}\right) \tag{1}$$

A graph of this transformation looks like:



Here we can see that in the middle of the range (from about 0.2 to 0.8) the curve is linear, so equal changes in that range are assigned equal importances. Near the ends, however, the curve is steeper, indicating that small changes are more important.

Once you've transformed your fraction data with logit, you can usually use t-tests and the like to compare groups. The logit transformation maps the range $(0, 1)$ to the range $(-\infty, \infty)$, which is the natural range for the normal distribution. You get means and CIs in "logit space", which you then have to reverse transform back into normal space.

Note that in normal space, your CIs will **not** be symmetric; that is expected, but it implies that when you report CIs, you'll write something like: "the fraction of affected cells is 0.040 (95% CI: 0.010 - 0.099)."

Note that the logit transformation blows up when the fraction observed is exactly zero or one. Different programs and practitioners deal with this in disparate ways; some add a small number (e.g., 0.025 or 0.05) to each value; other remap the range $(-\epsilon, 1+\epsilon)$ to $(-\infty, \infty)$ so that zero and one map to finite logit values.

If you deal with your data in percentages instead of fractions (i.e., if you work with numbers ranging from zero to 100), then of course you remap that range instead.

For completeness, we should mention there is a more advanced way to deal with data like this; you can use a "generalized linear model (glm) of a binomial distribution with a logit link function." This is much more complex, and Prism can't do it; a simple logit transformation is much easier to get your head around, and is good enough for most work that you'll encounter. We won't go into this any further, except to note that if you're reading a paper and the authors used a glm to analyze their fraction or percentage data, they did the right thing.

3.2 Logarithmic transformations

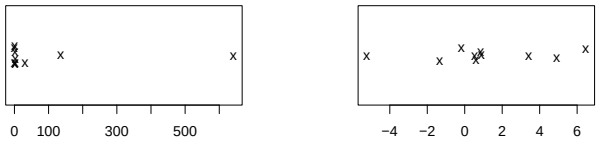
Another common case where data transformations are needed involve those where the data are measuring the sizes (e.g., mass or volume) of things, or counts of things, and the values span a few or more orders of magnitude. The sizes of colonies of cells or animals, and bank balances (which are counts of dollars), are canonical examples. In these cases, we again find that absolute differences have disparate interpretations, depending on the context of the absolute value.

For example, if you are a financially destitute student who attends seminars solely for the free food, someone giving you $1,000 could really change your quality of life, at least for a month or two. Conversely, to an internet billionaire, an extra $1,000 probably doesn't mean much at all. A good approximation of this psychological interpretation of the value of incrementally more money is that it is the change in the order of magnitude that matters, not the change in absolute value[1].

Since a logarithm of a value is essentially that value's order of magnitude, transforming such data into log space can be useful. If the log of your data looks more or less normally distributed, you might conclude that your data follows a lognormal distribution. If this makes sense, you'd do your t-tests in log space, compute CIs, and then transform back to report the CIs in "normal" space.

Lognormal data tends to look like it contains outliers. As an example, the plot on the left below shows 10 datapoints sampled from a lognormal distribution. You might be tempted to conclude that the rightmost point is an outlier. However, transforming the data into log space makes the rightmost point seem much more plausible as a *bona fide* measurement, as shown on the right.

[1] While we use a psychological example here, there is often a physical basis for such a relationship in physical and biological systems, e.g., the Weber-Fechner law.



Be careful to take due account of units of measure when you take logs of measured values, as it is mathematically incorrect to take a logarithm of a quantity that is not dimensionless. You'll often need to divide all measurements by a standard value to formally justify this operation.

Also, be careful that the value you're taking the log of is an absolute value. It doesn't make sense to take the log of a temperature measured in C or °F, since those are not absolute values. You'd want to work in K for that case (0 C or 0°F does not mean no heat, but zero K or R does) – this point probably applies only to astrophysics researchers, but the principle always applies!

When working in Prism, you often don't have to do the transformation into log space and back yourself. Instead, work with geometric mean, SD, and related quantities like CIs. You can appreciate why this works by considering the definition of the geometric mean (another measure of central tendency):

$$GM_x = \sqrt[n]{x_1 \cdot x_2 \cdot x_3 \cdots x_n} = \sqrt[n]{\prod x_i} = e^{\frac{1}{n}\sum \ln x} \tag{2}$$

The last terms shows that the geometric mean is the antilog of the average of the logs, which is exactly what we've been saying: transform data into log space, do your analysis (compute the average, in this case), and then transform back.

3.3 Reciprocal transformations

You may have also heard of the harmonic mean. The harmonic mean is just the reciprocal of the average of the reciprocals of the measurements you have.

$$\frac{1}{H_x} = \frac{1}{n}\sum \frac{1}{x_i} \tag{3}$$

In this case the transformation is a reciprocal. In the U.S., the fuel economy of a car is reported in miles per gallon; many argue that it is more appropriate to measure fuel economy in gallons per mile (or liters per kilometer), as is done in Europe. When Congress legislates average fuel economy of the fleet of vehicles that a carmaker produces, what do

you think should be averaged? Think about what deltas imply about fuel consumption at different points in the scale; does improving the fuel economy of a fleet of taxi cabs from 45 mpg to 50 mpg have the same effect as improving the fleet from 10 mpg to 15 mpg?

4 Introduction to p-values

In our last session, we learned that when the 95% CI of a test like the t-test includes "no effect" as a plausible value, we have to admit the possibility that there is no difference in the true means of the groups from which measurements are being sampled. There is another line of reasoning that leads us to the same conclusion; this involves the logic of formal statistical testing, and the interpretation of p-values.

In this line of reasoning, we begin by stating a null hypothesis, H_0. The odd thing about the null hypothesis is that you will try to show that it is not plausible. By doing so, you will show that its complement, the alternative hypothesis, H_1, is likely to be true.

Once you have stated the null hypothesis, you compute a probability, called the p-value. Informally, the p-value is the probability of "getting the data that you actually got" under the assumption that the null hypothesis is true. The tricky part here is defining the colloquial "getting what you actually got" appropriately. More formally, it is the probability of getting a result as, or more, inconsistent with the null hypothesis than the data that you actually observed.

If the p-value from this computation is small (below some pre-determined cutoff value usually written as α), then you can conclude that the null hypothesis is unlikely to be true, and you reject it. You have a statistically significant result.

If the p-value is not below α, your test is inconclusive. You cannot conclude anything about the null or alternative hypotheses. Be sure to understand this point, and do not misinterpret the p-value. The p-value is **not** the probability of the null hypothesis being true.

For an unpaired t-test, the null hypothesis is that the means of the two sampled populations are the same. Using our notation from our previous session, we have:

$$H_0\text{: } \overline{y} - \overline{x} = 0$$

Now, the means of your two samples have some observed difference, Δ, which is presumably not zero. We can compute the probability of taking two groups of samples from the single hypothesized distribution, and obtaining sample group means as far apart, or further, from each other as the observed Δ. This probability is the p-value reported by a t-test.

As an exercise, use simulated data to compute p-values for three cases.

- One hundred samples in each of two groups. The true means of both groups is zero, and the SD for both groups is one. In this case, the null hypothesis is true. How often will it be rejected?
- Same as above, except that the true mean of group B is one.
- Same as above, except that the true mean of group B is 0.2.

5 A Caution for Cumulatively Collected Data

Consider the follow scenario: a researcher performs an experiment with $n = 6$ and obtains a p-value greater than the pre-determined $\alpha = 0.05$. Knowing that a larger n has more power, he decides to collect six more datapoints, and redo the t-test with all 12 values.

Now imagine that there really was no difference between the two groups being compared; i.e., the drug or treatment being tested actually has no effect. It should be clear that the probability of erroneously concluding that there is a statistically significant difference between the two groups is greater than 5% (as that is the probability of doing so just in the first step). Thus reporting the result as statistically significant with a 95% confidence is wrong! The confidence level is **much lower** in this case. In fact, it can be shown that if you follow this practice forever, you will, at some point, get a statistically significant result.

The bottom line here is that you cannot cumulatively add measurements to a dataset, and keep recomputing p-values until you get an answer that you like. Don't be tempted (or bullied) into doing this.

6 One-Tailed and Two-Tailed p-values

The p-values we have been computing so far are two-tailed p-values, because they compute the probability of seeing a difference as or more inconsistent with the null hypothesis in either direction. You will occasionally see one-tailed p-values reported. As you might have guessed, a one tailed p-value is usually just half of a two tailed p-value. This is the case for the t-test (because the t-distribution is symmetric), but isn't the case, for example, with Fisher's Exact test.

Whether you report a one- or two-tailed p-value is sometimes debatable (although one-tailed values are falling out of favor). If you are in a position to not care about or are willing to chalk up deviations in a direction other than that which you expect to random sampling only, it may be appropriate to use a one-tailed p-value. In a coin flipping example, if a casino were running a game where it wins when heads appears, and you were a gaming inspector looking for cheating casinos, you might use one-tailed p-values since you don't care about coins weighted towards tails. Doing this is tantamount to cutting off one side

of a CI on the basis that you know it is impossible for the difference between means to be, say, negative. If you are not comfortable with this, then you should not use a one-tailed p-value.

In general, it is safer to stick with two-tailed p-values. If you choose to use a one-tailed value, you should have good reasons for doing so, and should state them before you collect any data. Computing a two-tailed p-value of 0.07 and then using a one-tailed value so you report a significant result is certainly not proper. Additional arguments in favor of two tailed p-values are that they are more conservative, and that they are consistent with CIs.

7 p-Value Cutoffs

Throughout the bulk of our work so far, we have been computing 95% CIs and using a p-value cutoff of $\alpha = 0.05$. While this is very common, it is important to realize that there is nothing sacred about this arbitrary number. It is often appropriate to choose a different cutoff, and we will spend a little time exploring the motivations and implications of doing so. Most importantly, it is important to choose your cutoff before you collect any data.

Some like to joke that statistics is never having to say you are wrong. When we "reject" a null hypothesis, we are not saying it is wrong, just that it is unlikely to be true; we are always trying to hedge. One never rounds a p-value to zero. You will see low p-values reported as "$< 10^{-9}$", or something similar, indicating that the null hypothesis is 'very, very unlikely to be true', but not 'impossible'.

Unfortunately for the cautious, when we perform a statistical test, compute a p-value, and compare it to a pre-established α, we are usually planning to make some decision and take some action (or not) based on the result. The decision will have consequences, sometimes quite profound ones. It is therefore important to understand what errors we might be inclined to make, how likely they are, and what their consequences are.

When we perform a test and we find a statistically significant result when, in fact, the null hypothesis should not have been rejected, we have committed a Type I error. Whenever you test a hypothesis that is not, in fact, deserving of a statistically significant conclusion, the probability of making a Type I error is α. The overall rate of Type I errors across many tests is dependent on how many hypotheses you test.

When we fail to reject the null hypothesis when it is in fact false, we are committing a Type II error. As above, determining the Type II error rate requires additional information on the proportion of tests that should lead to a significant conclusion.

As α decreases, the probability of making a Type I error goes down, and the probability

of making a Type II error goes up. You can compensate for the latter by increasing the sample size. In order to quantify the Type II error rate, you need to know something about the rate of occurrence of the outcomes in the population(s) you are studying.

Consider the following statistical analyses:

Screening compounds for a drug that might have a desired biological effect:

Type I error: An ineffective compound is viewed as putatively effective. The consequence of making this error is the additional time and money invested in further testing an unhelpful compound.

Type II error: An effective compound is abandoned. The consequence of making this error is a missed opportunity to improve human health (or make big money for your employer).

In this case, you might want to increase α; even a value of 0.2 might be appropriate.

Phase III clinical trial of a drug to treat a disease that currently has no effective treatment:

Type I error: You determine that the drug is effective when it is not. The consequence is that patients are paying for a placebo, and unnecessarily enduring any side effects that might be present.

Type II error: You abandon an effective drug, and leave patients with no viable treatment. The consequence is a missed opportunity to improve human health.

In this case, you may consider increasing α, weighing the costs and side effects of the drug.

Phase III clinical trial of a drug to treat a disease that already has an effective treatment:

Type I error: You determine that your drug is more effective than the existing one when it is not. You have actively deprived patients of better care.

Type II error: You deprive patients of a better or less expensive treatment, but they are still adequately treated.

In this case, you may consider lowering α; a value as low as 0.01 might be appropriate. You will probably need a larger study to prove that your drug is better than the existing therapy; this seems appropriate when changing something that is recognized to work well.

8 Statistical Significance vs. Biological Significance

It is important to recognize that statistical significance and biological significance are two very different things, and that statistical tests cannot help you at all with assessment of biological significance.

When performing experiments, it is exceptionally difficult to produce two groups that truly have absolutely no difference between them. We usually have to admit that there is some small effect from nearly every manipulation we perform or choice that we make. When n is small, these tiny and uninteresting effects tend to go undetected, so traditionally this has not been a problem. However, when n is large, as is often true when high-throughput experimental platforms are used, we more often have the ability to resolve these subtle and uninteresting differences. When considering results of a statistical analysis with a p-value alone, we may not fully appreciate that the measured effect, which is statistically significant, might be too small to be biologically relevant.

For example, a large study of the efficacy of a diet pill might show that the drug "works" ($p = 0.034$). In this case, you can be moderately sure that there is some difference between the treated and control groups, but this result doen't tell you anything at all about how large a weight loss the pill resulted in. That, of course, can be answered by inspecting the 95% CI of the difference between the two groups. From the p-value, you can infer that the 95% CI doesn't include no effect, but that interval could be a loss of 1g - 3g of mass, or a loss of 5lb - 10lb. The first case would not be biologically significant, while the second clearly would.

Because CIs allow you to assess biological significance of results, they should be (although are often not) preferred over p-values.

9 Homework

The data in the following table show the fraction of cells successfully transfected with a particularly difficult virus. Data for thirty trials for each of two methods have been obtained, and you're asked to evaluate if one method is statistically significantly better than the other.

1. Analyze raw data
   (a) Plot the raw (untransformed) data. Show the 95% CI of the mean of each group.
   (b) To your eye, do the data from each method look normally distributed?
   (c) What is the mean transfection fraction from method A? How many of the sampled values are above the mean?
   (d) What is the mean transfection fraction from method B? How many of the sampled values are above the mean?
   (e) What is the 95% CI of the mean transfection rate for each of the methods?
   (f) Perform an unpaired t-test using the raw values. Based on this result, can you conclude that the two methods are statistically significantly different?

2. Analyze transformed data

   (a) Transform the data using the logit transformation, and prepare an appropriate plot.

   (b) What is the mean of the logit transformed data from method A? How many of the sampled transformed values are above the mean?

   (c) What is the mean of the logit transformed data from method B? How many of the sampled transformed values are above the mean?

   (d) Perform an unpaired t-test using the transformed values. Based on this result, can you conclude that the two methods are statistically significantly different?

3. Error bars

   (a) Compute the mean and 95% CI of the mean for the transformed data for method A and B (you can do this as part of the t-test above).

   (b) Transform the mean, and the upper and lower bounds of the 95% CI of the mean, back into real space. Compare these bounds to the 95% CIs you obtained in part 1.

   (c) **Challenge:** Add error bars for the back-transformed 95% CIs of the means to your plot.

   Hint: In Prism, you'll likely need to create a new datasheet choosing the option "Enter and plot error values already calculated elsewhere; Enter: Mean (or median), Upper/Lower limits."

   (d) **Challenge:** Build a Prism workbook where these error bars are updated automatically if the data changes. This is necessary for a fully reproducible workflow.

   Hint: You'll have to define a new transformation that is the inverse of logit, and you'll need to use Prism's "paste link" capability.

9.1 Tabular Data

These data are available on the course web site; there's no need to key this in by hand.

	methodA	methodB
1	0.0462	0.0062
2	0.0056	0.0021
3	0.0384	0.0097
4	0.0183	0.0011
5	0.0018	0.0090
6	0.0037	0.0124
7	0.0232	0.0753
8	0.0179	0.0038
9	0.0094	0.0006
10	0.0048	0.0333
11	0.0211	0.0005
12	0.0092	0.0066
13	0.0196	0.0322
14	0.0086	0.0203
15	0.0256	0.0030
16	0.0031	0.0077
17	0.0081	0.0022
18	0.0066	0.0022
19	0.0131	0.0167
20	0.0381	0.0028
21	0.0087	0.0077
22	0.0163	0.0023
23	0.0024	0.0016
24	0.0154	0.0077
25	0.0032	0.0023
26	0.1624	0.0007
27	0.0075	0.0014
28	0.0059	0.0020
29	0.0145	0.0008
30	0.0018	0.0033

