Quantitative Understanding in Biology Short Course
Session 3
Working with Count Data;
Fisher Test and Contingency Tables

Jason Banfelder
Luce Skrabanek

March 25th, 2025

1 Introduction

In our last session, we emphasized that if you are analyzing data that are fractions or percentages, you should transform the collected data into logit space before analyzing. While this is correct, it turns out that you probably won't be doing this as often as you might think. This is because most of the time, when you measure fractions or percentages, like the fraction of mice with a particular knockout that die after one week, the raw data that you collect are not the fractions, but the raw counts; e.g., of alive and dead mice.

When you turn these two observations into a single fraction, you're throwing away some information. Your intuition should tell you that throwing away information is generally a bad idea, and that we can get more out of our data if we keep it all. Moreover, as we've started to see how important large n is to getting narrow confidence intervals and statistically significant results, the realization that you're throwing away information about n when you compute the fraction of mice that survived exposure should really give you pause.

Statistically speaking, we can say…

$$\frac{17}{23} \neq \frac{170}{230}$$

… because the second proportion is associated with a much higher n, and thus should give us more statistical power to draw inferences.

2 Bernoulli Trials and the Binomial Distribution

Many studies measure a proportion of subjects that produce a yes/no outcome. Statisticians would say that each individual outcome is the result of a Bernoulli trial.

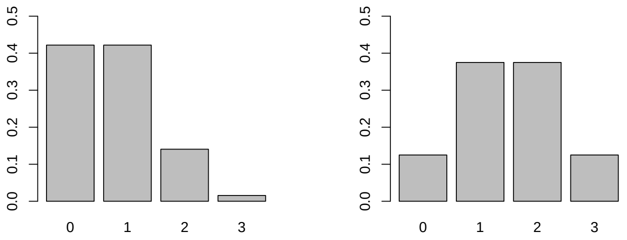
The canonical example of a Bernoulli is a coin flip. In this case, assuming that the coin is fair, the two possible outcomes (heads or tails) are considered equally likely, and we'd say that the true probability of getting heads is 0.50. Just like when we're measuring the means of heights of people, when we investigate a coin by repeatedly flipping it and observing how many heads and tails result, we will never exactly know the true probability of obtaining heads. Similarly, since coins are not precision machined and balanced, we should recognize that practically no coin will have a true probability of obtaining heads of exacly 0.5; with enough flips (perhaps millions or billions), we could detect and quantify an underlying bias. But if it is small enough, it probably wouldn't be of interest. This would be akin to detecting a statistically, but not biologically, significant difference in your work.

Statisticians tend to label one of the outcomes of a Bernoulli trial as 'success', and the other as 'failure.' We should recognize that this distinction is often arbitrary. But it is important that one is clear which outcome is which; in most cases 'death' would be equated with 'failure,' but if you're trying to drive cancer cells to undergo apoptosis, 'death' might be more naturally labeled as 'success.'

Although the canonical Bernoulli trial is a coin flip, don't lose sight of the fact that the probability of the 'successful' outcome of an arbitrary Bernoulli process is not necessarily 0.5. For example, if you're mating a pair of heterozygous mice for offspring that you hope will express a recessive trait that follows pure Mendelian inheritance patterns, the probability of expressing the trait (which we might label as 'success') is 0.25.

We normally are not interested in a single Bernoulli trial, but rather in how many successes (or failures) we observe from a fixed number of trials. The results of these meta-experiments follow what is known as a Binomial distribution. You can work out simple cases of binomial distributions in your head. For example, if your meta-experiment is to flip two coins (or the same coin twice), and count the number of heads you observe, you'd expect there to be a 25% chance that you get no heads (an outcome of 0), a 50% chance that you get 1 head, and a 25% chance that you get 2 heads.

Similarly, if you breed three offspring from your heterozygous mating pair, you can predict that there is a $(0.75)^3 = 42.2\%$ chance that none of your mice would express the recessive phenotype. With a little more work, you could work out and plot the probabilities of the four possible outcomes from this meta-experiment, as shown on the left:



These are two plots of different binomial distributions where $n = 3$. The left plot shows the probability of getting a certain number of mice with the recessive phenotype; the right shows the probability of getting a certain number of heads in three flips of a fair coin. On the left, $p = 0.25$, while on the right $p = 0.5$. We follow the (somewhat confusing) standard nomenclature here for the binomial distribution. Don't confound the p here (which is the probability of success of the underlying Bernoulli process), with a p-value.

3 Confidence Interval of a Proportion

While understanding binomial distributions is helpful in predicting how many successes we might expect to observe in a meta-experiment that can be thought of a battery of Bernoulli trials, we often need to work this backwards. We have observed a certain number, x, of successes out of n trials, and we wish to estimate the true probability of success, p, of the underlying Bernoulli process. For example, we may observe that 17 out of 23 KO mice animals die within one week. We would state that $17/23 = 74\%$ of the animals die, and this would be our best guess of the true value of p. Of course, we wish to compute a CI for this result.

In Prism, you can do this by creating a simple, two-line "Parts of Whole" table, and then analyzing it with a "Fraction of Total" analysis. Although Prism now recommends that the "Wilson/Brown" methods be used, the results presented throughout these notes are based on the more commonly used "Clopper/Pearson" option. According to this, the 95% CI for the true probability of success is 52% to 90%.

Just as with t-tests and the normal distribution, we expect that a 99% CI would be wider, and in fact for this example it is 45% to 93%. Similarly, the 80% CI is narrower: 59% to 86%.

At the same confidence level, CIs become narrower as more data is collected. For example, if we observed the same proportion of mice dying, but with ten times larger n (i.e., if 170 out of 230 mice died), our 95% CI becomes 68% to 79%.

CIs from proportional data are not symmetric, and are wider when near $p = 0.5$.

Unlike with logit transformed fractions, computing CIs of proportions of count data works just fine even when the number of success is zero. Try computing the 95% CI for the true probability of success for 23 attempts with no success.

In addition to computing CIs for outcomes of binomial processes, you can also compare the data to a hypothesized underlying probability of success. In Prism, this is done with the "Compare observed distribution with expected" analysis. When using this, you probably want to enter the expected values as percentages. The percentages you enter constitute the null hypothesis, and Prism will compute a p-value that tells you the probability of getting the data that you observed (or something more extreme) given your hypothesized underlying probability of success. As usual, if the p-value is low, you reject the null hypothesis. If it is higher than your pre-selected cutoff, you admit that the hypothesized value of p is plausible given the data you collected. This does not, though, "prove" that the hypothesized value of p is correct! Make sure you internalize this point!

As an example, you should be able to show that, at a 95% confidence level, you'd conclude that a coin that yielded 17 heads out of 23 coin flips is not fair, and a coin that yielded 16 heads out of 23 flips is plausibly fair.

4 Contingency Tables and Fisher's Exact Test

The binomial test just described is nice and easy, but our hypothetical experiment is poorly designed. To say that 74% of our knockout animals died within a week is not informative unless we also have a control group (maybe there is something very wrong with the food we've given all of our animals). If we did the experiment with controls, we would be in a position to formulate a contingency table:

	Outcome X	Outcome Y	Total
Group I: Experimental	17 (A)	6 (B)	23
Group II: Control	3 (C)	22 (D)	25
Total:	20	28	48

The relative probability of outcome X with respect to Y is:

$$\frac{P_I}{P_{II}} = \frac{\frac{A}{A+B}}{\frac{C}{C+D}} = \frac{\frac{17}{23}}{\frac{3}{25}} = 6.16 \qquad (1)$$

In the epidemiological literature, this ratio of proportions is known as the relative risk; this language implies that outcome X is worse than outcome Y.

In this case, just by looking at the data it is pretty clear that there is a significant difference in one-week survival due to the knockout. We would like to quantify what that difference is.

Unfortunately, although relative probability is easy to understand, results such as these are often expressed in terms of odds, not probabilities. You may recall that odds are defined as:

$$\text{odds} = \frac{p}{1-p} \tag{2}$$

That is, the 'odds' is defined as the ratio of the probability of an event happening to the probability of it not happening. If $p = 0.75$, the odds are 3:1, or just 3. Whereas $0 \leq p \leq 1$, the range of odds is much larger: $0 \leq \text{odds} < \infty$. For rare events, the odds is approximately equal to the probability.

Just as we computed a relative probability, we can compute the relative odds, or, as the literature calls it, the odds ratio:

$$\begin{pmatrix} \text{Odds} \\ \text{Ratio} \end{pmatrix} = \frac{A/B}{C/D} \tag{3}$$

You can enter contingency data like this into Prism using the "Contingency" option. Prism can then compute the odds ratio of success, and its CI, by choosing the "Chi-square (and Fisher's exact) test" from the "Contingency table analyses" section. For the data in this example, the odds ratio is 20.78, and the 95% CI is 4.78 - 77.5. As this CI does not include 1, which is the odds ratio that would correspond to no difference between the two groups, we can conclude that there is a statistically significant difference between the control and experimental group. Also note that the confidence interval is not symmetric.

When the values in a contingency table are very large, Fisher's exact test can be computationally intensive to compute. The Chi-square test is an alternative that uses some approximations that break down when your table has small entries. On a modern computer, you can usually just use the Fisher test. If you are performing many, many tests, you may want to look into alternatives (there are other issues in multiple hypothesis testing that we will touch on in another session).

Consider another contingency table:

	Outcome X	Outcome Y	Total
Group I: Experimental	4 (A)	246 (B)	250
Group II: Control	1 (C)	249 (D)	250
Total:	5	495	500

In this case, the experimental group seems to be roughly four times more likely to have outcome X. However, a Fisher test shows that there may be no difference at all between the groups; it is not unreasonable that the variation we observed is due to random sampling, as the 95% CI of the odds ratio is 0.67 - 50.

At this point, you may be wondering why we have elected to work with odds ratios instead of the more natural relative proportions. Thus far, all of our hypothetical examples have been of what are termed 'experimental studies'. In these studies, we define two groups, and then perform two different actions on the members of those groups. The outcomes are results of a Bernoulli trial. For experimental studies, there really is no good reason to introduce and work with odds instead of probabilities. The reason why this is done will become apparent in a little while; be patient.

Another kind of study, called a prospective study, is similar. In this kind of study, we define two groups, as before. However, the two groups are defined by some pre-existing difference. In an epidemiological study, this may be some prior exposure to a hypothesized risk factor for a disease. For example, if you hypothesize that people working in the meat-packing industry are at higher risk for contracting vCJD, one group would consist of those that work in the meat-packing industry, and the second would consist of subjects who do not. In this kind of study, once the subjects are selected and assigned to their groups, you let nature run its course, and, at the end of the study, observe how many subjects in each group present 'successful' or 'unsuccessful' outcomes.

The mathematics of the analysis of a prospective study is similar to that of an experimental study. Again, there is no particular motivation to use odds in lieu of probabilities in a prospective study. One of the advantages of a prospective study over an experimental study is that you don't need to manipulate, poke, prod, etc. your subjects; you are simply observing what would normally happen anyway. When engaging in research on human subjects, this is a big deal.

One of the problems with prospective studies is that, for rare outcomes, they need to be quite large in order to generate statistically significant results. Look again at the contingency table and the results of the Fisher test in the last example, now interpreting it as data from a prospective study. Our hypothetical study involved 500 patients, yet produced a very wide confidence interval: the 95% CI of the odds ratio is between 0.67 and 50. An informative exercise is to see how large our study would have to be to produce a statistically significant result.

We can somewhat crudely and artificially vary the size of the study by multiplying all elements of the contingency table by a constant factor and re-running the Fisher test.

If you try this, you can show that a study that can demonstrate that there is any significance at all between the two groups would require about 1,500 subjects, and to narrow the CI to something reasonable, we would need 16,000 subjects.

A corollary of the above example is that the values you use when computing Fisher's exact test (or any test that uses counts, for that matter, such as the binomial test), must be the absolute number of counts that were observed. You cannot use counts/min or incidents per 100,000 in a population, etc. Some laboratory equipment, such as scintillation counters, often report observations/minute; be sure to determine the absolute numbers of scintillations detected if you use such count data in statistical tests that depend on absolute counts.

The bottom line here is that prospective studies that investigate rare outcomes usually need to be large, and can be expensive, and time consuming. Consider that not only do we have to track a large number of patients, but we have do it for quite a while since we have to wait for the disease to manifest itself in the population.

The alternative is to do a retrospective study. In this case, we form two groups based on the outcome, and then look back in time to see if a hypothesized risk factor can be implicated. A contingency table might look like the following:

	Outcome X	Outcome Y	Total
Group I:	40 (A)	25 (B)	65
Group II:	10 (C)	25 (D)	35
Total:	50	50	100

In this design, we select the column totals, whereas in the prospective case we selected the row totals. While in our examples, the totals are the same, this does not have to be the case.

It is important to recognize that a contingency table from a retrospective study gives us no information about the prevalence or rarity of the outcomes. From these data alone, we don't know if outcome X or Y is rare or common. However, as we shall show in a moment, the odds ratio (but not the relative probability) of the groups computed from a contingency table is correct. Before we demonstrate this, however, we will introduce one more experimental design…

A cross-sectional study is a design where subjects are chosen without regard to either risk factor or outcome. You simply randomly select from the population, and tabulate the results in a contingency table. The analysis of a cross-sectional study is the same as a prospective study. The ultimate cross-sectional study is to sample the entire population (often this is only possible as a thought experiment).

Now, we can show how odds ratios can be computed from retrospective study data. Begin by considering a complete cross-sectional study of the whole population:

	Outcome X	Outcome Y	Total
Group I:	(A)	(B)	$(A+B)$
Group II:	(C)	(D)	$(C+D)$
Total:	$(A+C)$	$(B+D)$	$(A+B+C+D)$

If you prefer to think in more concrete examples, consider the hypothetical case of an outbreak of a disease in a small town. The population is 10,000, and half of the population works in the local sausage plant. There have been 100 cases of the disease reported in the town; 80 of the affected people are workers in the plant.

The relative probability and the odds ratio are computed as follows:

$$\left(\begin{array}{c}\text{Relative}\\\text{Probability}\end{array}\right) = \frac{\left(\frac{A}{A+B}\right)}{\left(\frac{C}{C+D}\right)} \tag{4}$$

$$\left(\begin{array}{c}\text{Odds}\\\text{Ratio}\end{array}\right) = \frac{\left(\frac{A}{B}\right)}{\left(\frac{C}{D}\right)} \tag{5}$$

Now, in a prospective study, we sample some fraction of the population, f_I, in Group I, and some other fraction, f_{II}, of the population in Group II. The data in our contingency table are:

	Outcome X	Outcome Y	Total
Group I:	$f_I \cdot A$	$f_I \cdot B$	$f_I \cdot (A+B)$
Group II:	$f_{II} \cdot C$	$f_{II} \cdot D$	$f_{II} \cdot (C+D)$
Total:	$f_I \cdot A + f_{II} \cdot C$	$f_I \cdot B + f_{II} \cdot D$	$f_I \cdot (A+B) + f_{II} \cdot (C+D)$

The table has six variables, and we don't know any of them! But we do know four of the products.

We can compute the relative probability and the odds ratio:

$$\left(\begin{array}{c}\text{Relative}\\\text{Probability}\end{array}\right) = \frac{\left(\frac{f_I A}{f_I A + f_I B}\right)}{\left(\frac{f_{II} C}{f_{II} C + f_{II} D}\right)} = \frac{\left(\frac{A}{A+B}\right)}{\left(\frac{C}{C+D}\right)} \tag{6}$$

$$\left(\begin{array}{c}\text{Odds}\\\text{Ratio}\end{array}\right) = \frac{\left(\frac{f_I A}{f_I B}\right)}{\left(\frac{f_{II} C}{f_{II} D}\right)} = \frac{\left(\frac{A}{B}\right)}{\left(\frac{C}{D}\right)} \tag{7}$$

So far, so good…

Now consider a retrospective study. This time, instead of sampling the groups by row, we are sampling the groups by column. We are sampling some fraction, f_X, of those subjects with outcome X, and another fraction, f_Y, of those with outcome Y. Typically (but not necessarily), for rare diseases, f_X is quite large (we look at a sizable fraction of reported cases), while f_Y is very, very small (we consider a tiny sliver of the whole population to be used as a control group). The data we have are

	Outcome X	Outcome Y	Total
Group I:	$f_X \cdot A$	$f_Y \cdot B$	$f_X \cdot A + f_Y \cdot B$
Group II:	$f_X \cdot C$	$f_Y \cdot D$	$f_X \cdot C + f_Y \cdot D$
Total:	$f_X \cdot (A + C)$	$f_Y \cdot (B + D)$	$f_X \cdot (A + C) + f_Y \cdot (B + D)$

Incidentally, retrospective studies are often also called case-control studies. The cases are those with a disease, and the controls are those without it.

Again, we have six variables, of which we know none. But we do know four products. When we blindly compute a relative probability…

$$\left(\begin{array}{c}\text{Incorrect}\\ \text{Relative}\\ \text{Probability}\end{array}\right) = \frac{\left(\frac{f_X A}{f_X A + f_Y B}\right)}{\left(\frac{f_X C}{f_X C + f_Y D}\right)} \neq \frac{\left(\frac{A}{A+B}\right)}{\left(\frac{C}{C+D}\right)} \tag{8}$$

…we see the result is incorrect. However, the odds ratio ‘magically’ works:

$$\left(\begin{array}{c}\text{Odds}\\ \text{Ratio}\end{array}\right) = \frac{\left(\frac{f_X A}{f_Y B}\right)}{\left(\frac{f_X C}{f_Y D}\right)} = \frac{\left(\frac{A}{B}\right)}{\left(\frac{C}{D}\right)} \tag{9}$$

Note that in the middle expression above, the numerator is not the correct odds of outcome X to outcome Y. However, due to the cancellation of the fractions, the computed ratio is still correct. It is because we are unable to cancel the fractions when computing the relative probability that we don’t obtain the correct result there.

Now we are in a position to understand why statisticians like to use odds ratios. It is a consistent quantity that works for all of the experimental designs considered: experimental, prospective, retrospective, and cross-sectional. That said, it is possible to compute CIs for relative probabilities in experimental and prospective and cross-sectional studies. While Prism can do this, some argue that methods based on odd-ratios are preferred because they allow comparison of results across the variety of experiment types.

Recall that for rare diseases, the odds are approximately the same as the probability. So, for rare diseases, as a bonus, you can use the odds ratio from a retrospective study as a good approximation for a relative probability (aka relative risk).

Now let us look at our retrospective study's contingency table again, and run our Fisher test. With only one hundred subjects, we have a statistically significant result. We also see that the odds ratio is close to the relative risk (BTW: in this example, the disease is not all that rare in our hypothetical population; diseases are often measured in incidents per 100,000 or million). Finally, note that the CI is about as wide as a prospective study with 4,000 subjects.

One of the principle advantages of a retrospective study is that they can be performed relatively quickly, since you don't need to select subjects and then wait for nature to run its course. For diseases with a long incubation period, this is a critical concern. They can often be performed by inspection of medical records (although there are assumptions that come into play).

As you might imagine, you can also design and perform matched pairs case-control studies. In these studies, the controls are selected to be similar to the cases in variables that are unrelated to the groupings. In our sausage plant example, for each patient that has the disease (cases), we would select a control from our population that has a similar age, weight, household income, kind of pet, etc. Except to state that these studies have additional statistical power over grouped case-control studies, we won't go into the details of experimental design or analysis of results here (you don't use contingency tables to analyze the results, as it masks the extra information inherent in the matched pairs).

Again, always remain aware that relative risk alone tells you nothing of the prevalence of outcomes. If someone tells you that you are sixteen times more likely to contract vCJD from eating beef if you vacation in the UK instead of France (vCJD outbreaks in the UK were a big deal in the '90s), you might consider altering your travel plans. Now consider that the odds of contracting vCJD were estimated at 5 in 10,000,000 for dining in the UK for a month, vs. 3 in 100,000,000 for dining in France for a month. Finally, consider that the odds of dying in a motor vehicle accident are roughly 1.4 deaths per 100,000,000 miles travelled. This implies that your round trip taxi ride to Newark Airport from campus is a bit more risky than your exposure to vCJD would have been in the UK. This is not to say that we shouldn't protect our food supply (left unchecked, the odds may have gotten a lot worse) or avoid risky behaviors, but it is important to keep things in perspective.

5 Homework

1. You wish to compare the relative expression of Sox2 in cfos +/+ vs. cfos -/- cerebral cortex cells.

   For five fields of view from a cfos +/+ tissue sample, the total number of cells (as determined by DAPI staining) and the number of Sox2 expressing cells were counted:

total cells	143	122	156	135	117
Sox2+ cells	47	52	47	57	54

   (a) Quantify the level of Sox2 expression in this sample.

   For five fields of view from a cfos -/- tissue sample, the total number of cells and the number of Sox2 expressing cells is:

total cells	118	123	148	137	156
Sox2+ cells	78	85	74	79	78

   (b) Quantify the level of Sox2 expression in this sample.

   (c) Compare the level of Sox2 expression in these two samples. Is there a statistically significant difference?

   (d) What can you conclude about the relative expression of Sox2 in cfos +/+ vs. cfos -/- cerebral cortex cells from these data?

2. In preparation for the next session's discussion of multiple hypothesis testing, please read:

   - `https://arstechnica.com/science/2017/04/the-peer-reviewed-saga-of-mindless-eating-mindless-research-is-bad-too/`
   - `https://web.archive.org/web/20170312041524/http://www.brianwansink.com/phd-advice/the-grad-student-who-never-said-no`

   Start by reading the original post; then read the two addenda at the top of the page, and (at least some of) the comments.