1.4: Basic Concepts of Hypothesis Testing

Learning Objectives

- 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, 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.

Fig. 1.4.1 A giant concrete chicken in Vietnam.

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

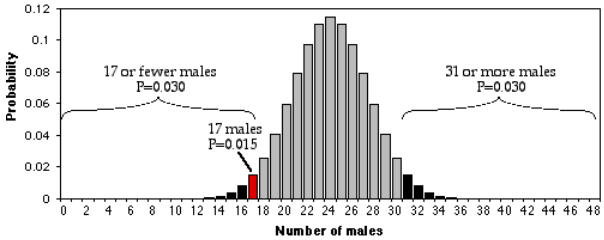

Fig. 1.4.2 Probability of getting different numbers of males out of 48, if the parametric proportion of males is 0.5.

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 17 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, in might even help you produce more female chickens someday. Any time a deviation in either direction would be interesting, you should use the two-tailed probability. In addition, people are skeptical of one-tailed probabilities, especially if a one-tailed probability is significant and a two-tailed probability would not be significant (as in 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 $2.45mm$, with a confidence interval on the difference of $\pm 1.98mm$."

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 conclude that 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 positives. 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 positives. 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 $1mm$ 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 $20mm$ with a confidence interval of $\pm 5mm$, 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.

Fig. 1.4.3 Two guinea pigs wearing hats.

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.

Reference

1. Picture of giant concrete chicken from Sue and Tony's Photo Site.
2. Picture of guinea pigs wearing hats from all over the internet; if you know the original photographer, please let me know.
