*From:* HANDBOOK OF BIOLOGICAL STATISTICS SECOND EDITION by JOHN H. MCDONALD

University of Delaware

© 2009 by John H. McDonald

See http://udel.edu/~mcdonald/statpermissions.html for details.

**Basic concepts of hypothesis testing**

*The basic idea of a statistical test is to identify a null hypothesis, collect some data, then estimate the probability of getting the observed data if the null hypothesis were true. If the probability of getting a result like the observed one is low under the null hypothesis, you conclude that the null hypothesis is probably not true.*

**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.**

Example: 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 tested a treatment, 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 tried a different treatment and 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 your treatment really changed the sex ratio.

However, what if you had 31 females and 17 males?

That's definitely more females than males, but is it really so unlikely to occur due to chance that you can reject the null hypothesis? To answer that, you need more than common sense, you need to calculate the probability of getting a deviation that large due to chance.

**P-values**



*Probability of getting different numbers of males out of 48, if the parametric*

*proportion of males is 0.5.*

In the figure above, the BINOMDIST function of Excel was used 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 percent of chickens are male, is 0.030."

**Significance levels**

Does a probability of 0.030 mean that you should reject the null hypothesis, and conclude that your treatment 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 probability value (P) 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; people could have agreed upon

0.04, or 0.025, or 0.071 as the conventional significance level.

The significance level you use depends on the costs of different kinds of errors.

With a significance level of 0.05, you have a 5 percent chance of rejecting the null hypothesis, even if it is true. If you try 100 treatments on your chickens, and none of them really work, 5 percent of your experiments will give you data that are significantly different from a 1:1 sex ratio, just by chance. This is called *a "Type I error," or "false positive."*

If there really is a deviation from the null hypothesis, and you fail to reject it, that is called a *"Type II error," or "false negative."*

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.

You must choose your significance level before you collect the data, of course.