Hypothesis Testing

Sticking with our fish example 📖, the biologists wanted to see whether or not the fish population had increased due to some management intervention. Thus, on the assumption that the numbers of fish caught in their netting efforts reflect the population size, the hypothesis they tested is: **H**: average number of fish per net is higher than 12 (the historical value). Does the data support this hypothesis?

We need a philosophical pause here. There are some circumstances wherein it will be very easy to answer the question, without dressing up the data in a cloak of statistics.

Suppose they caught fish in each of 10 nets and and the average in their sample was 12.24. Is this the evidence they seek? The short answer is no (my opinion). Their new average is indeed a larger number than 12, but not in a biologically meaningful way. Indeed, in this case, I would argue against applying any sort of formal test, since simple common sense would argue here that that there is no reasonable evidence of *biologically significant* improvement.

Suppose they caught fish in each of 10 nets and and the average in their sample was 42.54 (and suppose that the smallest number in a single net was 24). Is this the evidence they seek? The answer is easily yes, and dramatically so. This also requires no formal testing. The evidence is overwhelmingly clear. Formal statistical tools are not required.

Now, suppose they caught fish in each of 10 nets and and the average in their sample was 15.4 (sound familiar?). Is this the evidence they seek? Read on.

Null and Alternate Hypotheses, Alpha

A good analogy to scientific hypothesis testing is a court trial. An individual is brought to trial if there is some reason to believe they may have committed a particular crime. The starting point for the trial is a presumption of innocence, and ask whether or not the evidence would be plausible if in fact he or she were indeed innocent. In science one does a hypothesis test to determine whether a postulated (or hoped for) phenomenon has occurred. A null hypothesis is formed that denies that occurrence, and the data is examined to see if it is consonant with or if it contradicts the null hypothesis. In a court trial, if the hypothesis of innocence is rejected the defendant is declared guilty. In a statistical hypothesis test, we say that the null hypothesis has been rejected. Another commonly used phrase is to say that we have a statistically significant result, meaning only that the result does not appear plausible if the null were in fact true. In particular, statistically significant does *not* mean “significant” in our English use of the term. It does not mean “important”.

In a court trial, there is some chance that an innocent person is found guilty; similarly, there is a chance that we could reject the null hypothesis when in fact it is (at least approximately) true, a so-called false rejection or false significance. A difference is that in a hypothesis test, we can *choose* the amount of risk we are willing to take of a false significance. For most conventional scientific use, this chance, the so-called alpha-level, is set at 0.05. The particular choice of 0.05 is a historical artifact, but it does reflect the sense among scientists that in most situations, 1 chance in 20 of a false significance is an acceptable risk Larger values (0.10, 0.15) are often used in pilot studies, where the consequence of false significance is that some ultimately non-significant variables continue to be measured in the main study (at some expense). Lower alpha levels (0.01, say) might be used when a false significance has large and unpalatable consequences.

So we have null and alternate hypotheses, and have chosen an alpha level. The question at hand is how do we measure the evidence? There are several possibilities that occur to most people when they ponder this question; I’ll address them in turn. The first idea is to calculate the probability of our result (the mean of 15.4) if the null were true. Low values indicate evidence against the null. The second is to measure how far the observed value is from the null (counting in SDs of the statistic). The further away, the stronger the evidence. Both of these are flawed; the first fatally, the second only somewhat. I'll discuss both of them in turn, then discuss our current approach, based on so-called *p*-values.

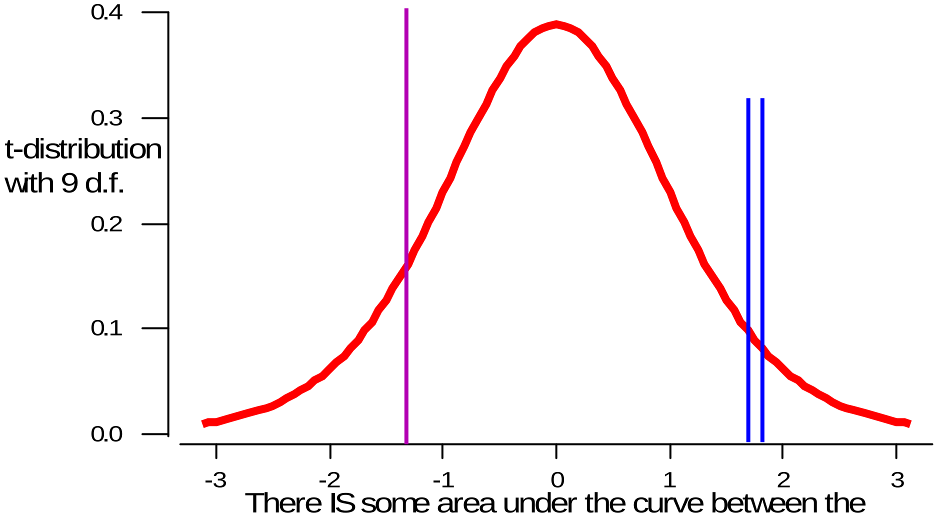
The Probability of the Observed Result

Using the fish example, one thought is to ask whether our result (a mean of 15.4 fish per net) has high or low probability of having occurred under the condition that the population mean is still 12.5. If it has a suitably low probability, we could reject the null hypothesis. This sensible sounding idea suffers from a fatal flaw: the probability of *any* single value is technically zero. The chance of getting 12.5 is zero; the chance of getting 100 is zero. How can this be? We *did* get 15.4! How could the probability be zero? This is surely paradoxical.

There are two parts to the argument. The first is based on the assumption that the mean is a value from a continuous distribution. Soon, we'll need that distribution to be at least approximately Normal, but for now, simply continuous is enough. Probability is defined for intervals in such a distribution (think of area under a curve between two points). There is positive probability of a value between 15 and 16 (or any other biologically possible interval), but, technically, a probability of zero (no interval, so no area) is assigned to any given point. This is an odd idea, and doesn't sit comfortably with most folks the first time they encounter it (include me in that list!).

Here is the resolution to the paradox. When we measure anything, we always round our answers. So what we've recorded as 15.4 is really known only to be between 15.35 and 15.45. That *does* describe a narrow interval, so a measured answer of 15.4 (with rounding) *does* have non-zero probability (that is, whatever the probability is of being between 15.35 and 15.45. We could compute the probability of getting 15.4 plus and minus rounding, but then our answer would depend on the level of rounding, leading to somewhat arbitrary probabilities. The more precisely we measure (i.e. the narrower the rounding interval), the lower the probability, with no connection at all to degree of evidence for or against the null hypothesis.

Area under the curve (i.e. probability) exists only for intervals



There IS some area under the curve between the two blue lines, but none over the purple one.

Using the graph for illustration, we could compute the chance of being between 1.8 and 1.9 SDs above the mean, but there is no probability defined for, say, −1.3 SDs below the mean (or any other single point!).

Distance From Null

Another approach that occurs fairly readily is to calculate the distance between the observed mean and the null. Here it is 15.4 – 12.5 = 2.9. It becomes clear fairly quickly that we can't judge this value all by itself because we don't know if 2.9 is big or small; we need to assess it relative to the variation in the mean. If we scale it by the SD of the mean (which here is $4.72/\sqrt{10} = 1.49$), we get 2.9/1.49 = 1.94. Thus our observed mean is 1.94 SDs from the null. This in fact is the way we used to do hypothesis tests back in the old days (you remember: when we walked to school barefoot in the snow, and used pencils for computers…). Essentially, we would pick a cut-off point based on certain criteria (I’ll come back to these below, when I compare our current method (based on so-called *p*-values)). This would be called the “critical *t*-statistic”. If our observed *t*-statistic (referencing the fact that the distribution of the mean is assumed to follow a *t*-distribution) is larger than the critical value, we declare there to be sufficient evidence with which to reject the null hypothesis. This method has a weakness that I will address when I compare this method to the *p*-value approach.

The p-value of a Hypothesis Test

By definition, the *p*-value for a hypothesis test is the probability of obtaining a test statistic as extreme or more extreme than the observed test statistic, assuming the null hypothesis to be correct. The starting point for calculating the *p*-value is in fact the distance from the null (in SDs) of the observed statistic. (And remember we are talking about the SD of the distribution of the statistic, not the SD among individual data points.) If you wish, take a look at the linked Excel spreadsheet for a dynamic study of p-value calculations, exemplified for a two-tailed test.

To interpret a *p*-value, compare it to alpha. If the *p*-value is less than alpha, we reject the null hypothesis .

For our example, the observed mean of 15.4 fish per net is 1.94 SDs above the null mean of 12.4 fish per net. Assuming approximate Normality for the distribution of the mean, and then fudging the Normal to account for the estimated, as opposed to known, SD, we use the *t*-distribution with 9 degrees of freedom. In that distribution, the chance of being 1.94 SDs (or more) above the mean is 0.0415. This is marginally less than a typical choice of $\alpha = 0.05$, so we could say the result is marginally significant; there is some, but not overwhelmeing evidence (as determined by $\alpha = 0.05$). Note that our language would be different had we chosen $\alpha = 0.15$ or $\alpha = 0.01$. The degree of significance has to be judged in context of the amount of risk we are willing to take of false significance. Synonyms: alpha is sometimes called the significance level of the test, and the *p*-value the *observed* significance level.

P-value versus critical value approach

In the old days, one would do a hypothesis test by choosing an alpha, then determining (with respect to the appropriate distribution (a *t*-distribution with 9.d.f. for our example), a so-called critical value. This value was some number of SDs from that distribution such that the probability of being that far or further from the null was equal to alpha. Then, one computed the observed number of SDs from the null ones actual mean was. If this observed value was further than the so-called critical value, reject. If not, fail to reject. This is mathematically identical to, “if the *p*-value is less than alpha, reject. If not, fail to reject.” The utility of *p*-values is that, at a glance, a reader can compare a given *p*-value against *their own choice for alpha*, and come to their own conclusions. The use of *p*-values was made possible by modern computers; prior to the early 1980’s, people were restricted to such information as was published in tables (the kind commonly found in the back of statistics textbooks); space and computational limitations held us to the critical value approach.

One-tailed and Two-tailed Tests

In the fish example, the research question was aimed at detecting an increase. If there is interest in only one direction, or if there is theoretical justification for a given direction (“We’ve seen it go this way in the past, so we hypothesize it will do so again here”), then the test is a one-tailed test. Operationally, if the observed effect is *not* in the direction specified by the research hypothesis (and formally encoded in the alternate), stop. Done. There is clearly no evidence against the null in favor of your alternate. If the observed effect *is* in the hypothesized direction, then proceed to compute the *p*-value.

Computer packages routinely do two-tailed tests by default. You can over-ride that default with an option somewhere along the way. Alternately, you can get the *p*-value for your one-tailed test by dividing the two-tailed *p*-value by two. The *p*-value for one-tailed test (if the result is in the desired direction) is always precisely ½ that for a two-tailed test.

The test yielded a *p*-value of 0.0415, suggesting that there was pretty decent evidence (against an alpha of 0.10) in support of their management change. This *p*-value can be had from a statistics package easily: enter the data, and ask for a one-tailed test for an increase (or divide the default two-tailed *p*-value by two).