**Hypothesis Tests**

So it seems that this analysis presumes the following generic setup. So we have a population, and are concerned about some population statistic, η. We make a measurement, m, and would like to know whether this casts doubt or not on the accepted value, η0, of this statistic η.

**Null Hypothesis**

We have a Null Hypothesis about the population statistic, which is the ‘commonly accepted value’, η = η0. The value of this statistic may or may not be sufficient to determine the probability distribution of the measurements Pη0(m). If the Null Hypothesis is ‘the probability of having a boy or girl is 50-50’, then our statistic is η = probability of having, say, a boy = 0.50. And knowledge of p is enough to determine the probability distribution of having a certain number of boys or girls in n tries. In other cases, it isn’t enough. Say we have two drugs to treat a disease. The Null hypothesis is that ‘they are both equally effective’, that both random variables come from the same distribution. But our population statistic, η = the effectiveness = probability of recovery, p, would be undetermined. So I guess we wouldn’t know what η0 is. And so we wouldn’t know the supposed probability distribution of having x recoveries in n tires. Sometimes we may be able to cleverly carry out the necessary calculations without knowing this value. Other times, we would need to make some educated guess. Cause either way, we need to know enough about the probability distribution of the population to calculate the probability distribution of measurement outcomes, Pη0(m).

**Alternative Hypothesis**

Then we have the Alternative Hypothesis about η. It’s basically, that ‘η ≠ η0’. More concretely, it’ll be something like, ‘the odds of having a boy is different than 50-50’. Or ‘this drug has different effectiveness than the other’. Importantly, the Alternative Hypothesis and Null Hypothesis are exclusive (and when framed this way, I guess mutually exhaustive). Technically, this alternative hypothesis is a two-tailed alternative hypothesis, as opposed to a single-tailed alternative hypothesis, since we’re looking at the case η ≠ η0, rather than η > η0 or η < η0. These sorts of alternative hypotheses are discouraged?

**The test**

So what we do is presume the null hypothesis is true, and look to see if our sample measurement conforms to or discredits that notion. So we take some measurement of the sample statistic and get m = m\*. And then from the presumption of the null hypothesis being true, we calculate the probability distribution of that measurement, Pη0(m). And then we calculate the total probability of m taking on values as or more extreme than m\*. Say this equals some probability, *p*. This is called the *p-value*, or the *confidence level*. For instance, if by ‘as extreme as’, we means ‘as large as’, then we’d calculate:



This is a right-tailed test. Or if by ‘as extreme as’, we mean ‘as small as’, then we’d calculate:



This is a left-tailed test. Or if by ‘as extreme as’, we mean, ‘as far from m = 0 (presumably) as’, then we’d calculate,



This is a two-tailed test. If p were 0.90, then the probability of measurements as extreme as ours would be so likely, assuming the Null hypothesis is true, that we would not reject the Null hypothesis. If p were 0.10, then we might have second thoughts. The commonly accepted threshold is p = 0.05. So if p < 0.05, then we say we reject the Null hypothesis and consequently accept the Alternative hypothesis. So we can *kind of* say that p reflects our confidence in the Null hypothesis. Note hypothesis testing is best used to detect a difference in behavior of η from the null, not to determine the direction of that difference (greater, lesser). That’s probably best left to the confidence interval stuff? Might note that a p-value of say 0.03 means that there is a 3% chance we get our results even though the Null hypothesis is true. So there is a 3% chance that we’d incorrectly accept the Alternative hypothesis. So there is a 3% false-positive rate.

In this file we’ll do examples pertaining to characterizing a population based on a sample. We’ll look at estimating means and variances. And we’ll also look at estimating proportions in both binomial and multinomial situations. To that end, we’ll make use of the results in the Single Variable and Multiple Variable PDF files, especially the Normal Distribution, Student’s T distribution, χ2 distribution, etc.

**Attention!**

A pitfall. If you do a test and get a p-value of 0.06, you could add a little more data to it and see if p falls below 0.05 or not (presuming you’d like to accept the Alternative hypothesis). But having already potentially lucked out with a test with p-value close to 0.05, there’s like a 50% chance, say, of p increasing or decreasing at this point. And so there is a disproportionately high chance of p dropping below 0.05, when adding data to the present test. Instead, to be objective, you really have to do the whole test over again, with different (and more if desired) data. A so-called *Power Analysis* will tell you what a good sample size is.

A related problem is that you could keep doing tests until you *finally* find one with a p-value of 0.02, say. But that’d be cherry picking. If you have a bunch of tests, then you should fill their p-values into the False Discovery Rate (formula)? This will give you a bunch of adjusted p-values, a little bit higher, that can give you insight into how to interpret the collection of p-values.

**Types of Errors**

There are two types of errors. Type I is better known as a False Positive. This would occur when we falsely reject the Null Hypothesis, i.e., falsely accept the Alternative Hypothesis. And Type II is a False Negative, which would occur when we falsely accept the Null Hypothesis, i.e., falsely reject the Alternative Hypothesis.

**Example**

Say we have a population of 10,000 wolves. We sample 100 of them and find for our sample that the mean is = 82cm (that’s the measurement m\*). Let’s say the sample standard deviation is 8. Does our measurement rule out a previous estimate that the average height is 80cm, at the 95% confidence level?

So we need Pμ(). This is a Gaussian distribution with mean μ and variance σ2/n. There’s a problem though. We need to know the population variance, σ2, too. An approximate way to deal with this is to take the variance of our sample and naively equate it to the population variance (should use the 1/√(n-1) formula here, though, rather than 1/√n formula right?). Okay, well if we do, then Pμ() would be, in the large n limit, normally distributed with mean μ , and std = s/√n. So we want to calculate,



Can do this sort of calculation with scipy,



So we *would* reject the Null Hypothesis at the 95% confidence level (because p > 0.05), and conclude that our measurement invalidates the prior result. Maybe that’s because all the wolves grew in the interim. What if we didn’t make the approximation that s = σ? Then we’d use the Student’s T-distribution. We have to formulate things in terms of the z-score: T = ( – μ)/√(s2/n). And our T-distribution would have ν = n-1 = 99 d.o.f. Our T\* is (82 – 80)/√(82/100) = 2.5. So we want to calculate:



And again, with scipy,



So we’d still reject the Null Hypothesis, though we see there is a slightly greater probability of it being True, thanks to the fact that the Student’s T distribution has fatter tails.

**Example**

Suppose we’re manufactoring a drug, which is supposed to have an active ingredient with amount 100μg ± 1μg. We take a random sample of 50 tablets, and find a sample standard deviation of 1.5μg. What are the odds that we’d get this standard deviation, if the actual standard deviation were in fact 1μg?

We can do a χ2 test. So recall from the Multiple Variable PDF file that the sample variance, S2, is χ2 distributed with n-1 d.o.f., according to,



And n = 50. Let our measurement M be W. We want the probability that we’d W > W\* = (50-1)(1.5)2/(1)2 = 110.25. So this is:



Can do with scipy,



**Example**

Say we take a sample of 1000 people in a hospital and find 530 male births, but 470 female births. We want to know if this invalidates the hypothesis that male and female births occur in equal proportions. So now we’re doing a proportion hypothesis test.

H0: Our null hypothesis would be: ‘the odds of a male vs female birth are 50-50 (the statistic) and the plurality of male births (our measurement) is just a product of random chance’. So we assume η0 = p = 0.50.

HA: Our Alternative Hypothesis would be that there is a difference in the probabilities. So p ≠ 0.50.

Test: So to test we assume the null hypothesis that p = 0.50, and then calculate the probability distribution of measuring x male births out of 1000. So let the measurement be:



where X is the random variable for the number of male births. Pη0(x) is presumed binomial with probability of male birth being p = 0.50, and not male being q = 0.50 as well. So the probability distribution for x male births out of n is:



Can approximate with normal distribution, as easier to work with:



And want to know probability that something as apparently unlikely as getting x > 530 will happen. This means we want to know the probability of getting something like x > 530, or x < 470. Easiest is to reduce to standard normal distribution and figure out how many std’s from the mean this is. So let our measurement random variable now be:



and then its distribution will be:



And our measurement is:



So we’d want to know



This is about 5.8%. So maaaybe random chance. According to the generally accepted threshold of p = 0.05, we’d accept the Null hypothesis (or, fail to reject it). With scipy,



Can also get this directly from *statsmodels*.



**Example**

Let’s revisit the previous example. Say we take a sample of 1000 people in a hospital and find 530 male births, but 470 female births. We said our Null Hypothesis was that the births were 50-50, and the plurality of male births (our measurement) was just a product of random chance’. Let’s revisit testing this Null Hypothesis with the χ2 test. So we know that:



is χ2 distributed with ν = 1 d.o.f. So let’s form our Z2 guy.



And our χ2 distribution would have 1 d.o.f. So we want to calculate,



and this matches our previous result. So we’d fail to reject our Null Hypothesis.

**Example**

Say we conduct a presidential poll of 1000 people. And 45% are voting for the Democrat candidate, 45% the Republican candidate, and 10% the Independent candidate. Does this invalidate other polling which suggests there is 48% support for the Democrat candiate, 47% support for the Republican candidate, and 5% for the Independent candidate?

We’ll use the χ2 test. So recall in the Multivariable PDF’s file, we said that for a multinomial distribution, the ‘Z2’ variable,



was χ2 distributed (x = Z2),



and ν would be ν = 3-1 = 2 in this case. Let’s work out Z2.



And so we calculate,



So it would seem our results do invalidate the prior polling. Or are at least inconsistent with it.