1 Hypothesis testing

To date, we've been using statistics as a means for answering the question, "What should we expect to see?" This is explicit in notions of, for example, mean and median, and is still the case with concepts of variance (which addresses the question of how much dispersion we should anticipate in our data). Hypothesis testing is only a slight deviation from this concept; rather than addressing what we should expect to happen numerically, hypothesis testing provides a statistical mechanism for determining whether a particular statement is true or false.

As a concrete example, suppose I – having learned nothing in basic economics or statistics classes – buy a lottery ticket and miraculously win big tomorrow. I decide to use my winnings to buy an island in the South Pacific, and there just happen to be two for sale. Since I also love piña coladas, I'd like to live on the island which has a greater coconut yield. So I borrow a friend's seaplane and pay a visit to the islands, which happen to be unimaginatively named A and B. On my short visits, I see that A produces 5 coconuts while B produces 8, so I decide to purchase B and set up house. We can see why this may not be the wisest decision: having taken just one sample, it's possible that A underproduced and B overproduced by fluke, and I could have enjoyed many more piña coladas if I'd bought B instead. On the other hand, it also doesn't make much sense to visit each island until the day I die: even though if I waited this long I'd know for sure which one would yield more coconuts, I'd be out of time to enjoy my frosty treats! What is missing is a method for determining *with a reasonable amount of evidence, and a reasonable level of confidence* which decision is better (in this case, which island has a higher mean coconut yield). Hypothesis testing gives us a formal structure for determining with statistical precision whether A or B will produce more coconuts in the long run.

More formally, when we are interested in testing the veracity of a statement, we term this statement the *null hypothesis* and denote it H_0; simply enough, the opposing hypothesis – something else which can be the case (and not necessarily everything else which can be the case) – is the *alternate hypothesis*, denoted H_1. When we test H_0, we find data sufficient either to reject H_0 (and know with some degree of confidence that it is not the case) or we do not, and we fail to reject H_0. This may seem like a fine point, but failing to reject the null hypothesis is not the same as accepting it as true! Consider this: suppose I'd like to show that every horse is brown; I set $H_0 =$"Every horse is brown" and $H_1 =$"Some horse is not brown." I sample 10 horses, all of which are brown; through hypothesis testing, I cannot demonstrate that every horse is brown in this manner, I can only demonstrate that I have insufficient data to show H_1, that some horse is not brown. This is a good thing in general; in this example in particular, it's a good outcome because we know that not all horses are brown. In the wording of the lecture slides, if the data observed are "consistent" with the alternative hypothesis but not the null, we reject the null hypothesis in favor of the alternative; on the other hand, if the data observed are "consistent" with the null, we merely fail to reject it. Science is a process of falsifiability (and not so much direct verification) and in this sense failing to falsify a hypothesis is as strong a statement as can often be made.

Hypothesis testing assumes we do not know the true state of the world (otherwise, there would be no reason to test!). So with any test yielding a decision – whether to reject or fail to reject the null – there will be some level of error associated. Such errors can be viewed as flukes in the data or infrequent sequences of observations from a distribution; in testing which bus line gets me to campus the quickest in the morning, I had some early success with a route which later proved to be quite slow. But just because I'm pretty certain it's worse now doesn't mean it can't be better sometimes! And if we only view these "sometimes" we'll make the wrong decision about what's going on in the world.

Consider what is a successful result of a hypothesis: either the null hypothesis is true and we fail to reject it, or the alternative hypothesis is true and we successfully reject the null. Then it's pretty clear what is an erroneous decision: the null hypothesis is true and we reject it, or the alternative hypothesis is true and we fail to reject the null. When we reject the null even though it is true, we have *type I error*, while if we fail to reject the null even though the alternative is true we have *type II error.* If it helps, the type "number" of the error is the subscript of the hypothesis which is true, plus one. That is, in type I error the null H_0 is true ($0+1=1$), while in type II error the alternative H_1 is true ($1+1=2$).

It is worth considering that in different cases, type I error may be much worse than type II error (and vice-versa). As a relevant example, consider medical testing for some form of cancer. Here, a null hypothesis might be that a person has cancer and the test is looking for evidence to refute this. A false rejection of the null means that a patient with cancer has been told they are healthy, while a false failure to reject the null means that a patient without cancer is held for further testing. Hopefully it's obvious that the former kind of error is worse.

How do we make these decisions? For this, we hark back to the central limit theorem and confidence intervals. For now, let's constrain ourselves to considering only hypotheses involving the mean of a distribution (since this makes appeal to the central limit theorem much more natural). Using confidence intervals, we were able to build – for example – a 95% confidence interval for the mean as $[\overline{X}-E, \overline{X}+E]$ for some E. Suppose we have a test about the mean of a particular distribution, $H_0 : \mu = 5$ (where 5 is a number picked out of thin air), and an alternative $H_1 : \mu \neq 5$. Following a set of observations from this distribution, we see $\overline{X} = 4$, and a 95% confidence interval which gives us $E = \frac{1}{2}$ (again, numbers picked out of thin air). Since a 95% confidence interval is $[3.5, 4.5]$ and $\mu = 5$ is not in this interval, we'd do alright to reject the null here. That is, if we have 95% confidence that $\mu \in [3.5, 4.5]$, the odds that $\mu = 5$ are not good, so we reject it outright.

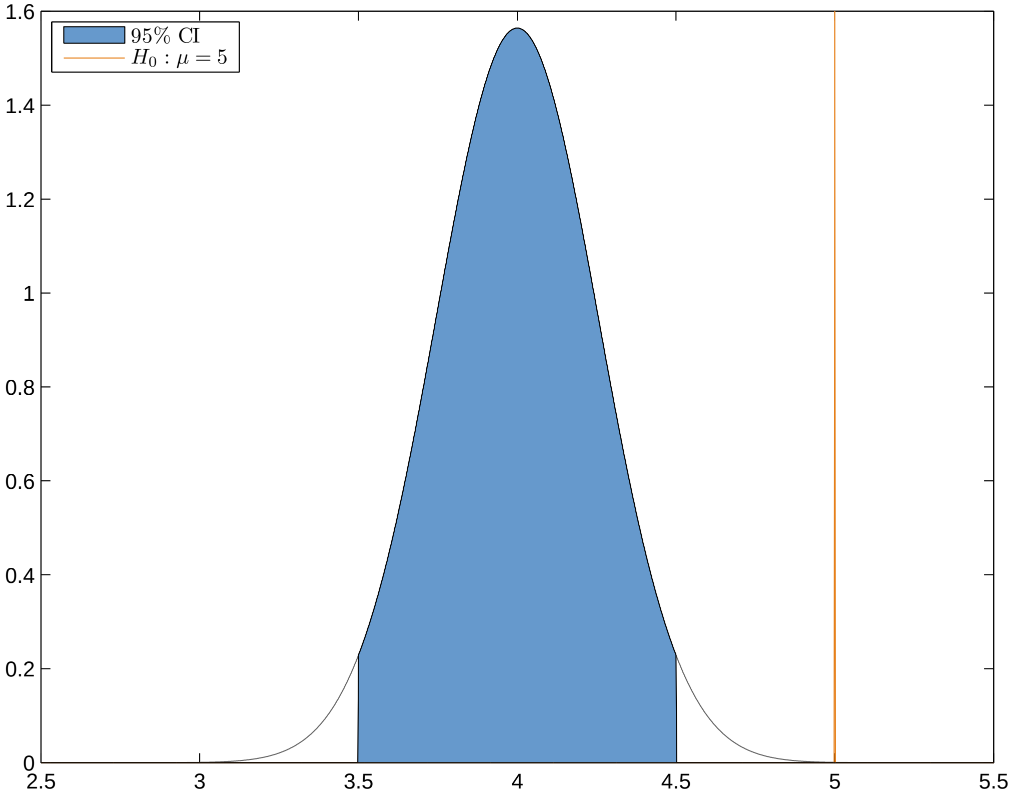

Figure 1: a graphical representation of the example with $\overline{X} = 4$ and $E = \frac{1}{2}$, testing $H_0 : \mu = 5$. We can see that the null hypothesis has μ lying well outside of the 95% confidence interval, so it is reasonable to reject the null hypothesis as false.

Now, it's worth noting that instead of confidence levels, when running hypothesis tests we use the term *significance*. Significance is denoted α, and represents the probability of type I error – the probability of falsely rejecting the null hypothesis. This notation is intentially suggestive: recall that we constructed $100(1-\alpha)\%$ confidence intervals. In this sense, if the null hypothesis is true, we'd fail to reject it $100(1-\alpha)\%$

of the time, and we'll incorrectly reject it $100\alpha\%$ of the time. So the α term for significance is in some sense the same as that for confidence intervals. As an aside, we denote the probability of type II error by β; this has no real analogy to confidence interval construction.

2 Problems and examples

We should begin by pointing out that what we'll be doing is testing whether or not a particular sample is drawn from a known distribution. Hypothesis testing generalizes, so that we can test whether or not two samples are drawn from the same distribution (whether or not it is known), but the math gets a little uglier and we have not covered this generalization yet in lecture.

Question 1: *I ride the Metro 302 line to get to and from school; it's the same as the Metro 2 line, except it has fewer stops and ostensibly goes faster. Let's say I know that it takes a mean time of* 62.5 *minutes to get home on the 302, with a known variance of* 46 *minutes* ^{2}. *Sometimes I take the 2 anyway, since I'd rather be stuck on a bus than stuck at a bus stop. Over* 30 *samples, I discover that it takes a sample mean time of* 67 *minutes to get home when I take the Metro 2 line. Does this evidence support the claim (at a* 2.5% *significance level) that the Metro 2 line takes longer to get me home than the Metro 302 line?*

We begin by forming our hypotheses. Since we'd like to disprove the idea that the 2 and the 302 take the same time, our null hypothesis is $H_0 : \mu_2 = \mu_{302}$. Our alternative hypothesis, representing the fact that we think that the 2 should take longer than the 302, is $H_1 : \mu_2 > \mu_{302}$.

Let $\overline{X}$ represent the sample mean of commute times on line 2. We look to reject the null on its own terms[1]; if the null is true, then according to the central limit theorem the value

$$Z = \sqrt{N}\left(\frac{\overline{X} - \mu_{302}}{\sigma_{302}}\right)$$

should be distributed $N(0, 1)$. Since we know N, μ_{302}, and σ_{302}, this tells us that

$$Z = 0.8076\left(\overline{X} - 62.5\right) \sim N(0, 1)$$

Plugging in for our observed $\overline{X}$, we obtain a z-score of

$$z_{\overline{X}} = 0.8076\,(67 - 62.5) = 3.6341$$

To move forward, we need to interpret the meaning of a 2.5% significance level. Since the alternative hypothesis is $\mu_2 > \mu_{302}$, we are running a one-tailed hypothesis test corresponding to a one-sided confidence interval (that is, if we saw $\overline{X} = 30$ in this example, we wouldn't want to reject the null in favor of the alternative since the alternative performs even worse!). We define the *critical region* to be the range of z values which will allow us to reject the null; in this one-dimensional case, the cutoff value for the critical region is called the *critical value.* To obtain a 2.5% critical region, we appeal to the central limit theorem and find

$$\Pr(Z > z) = 2.5\% \quad\Longleftrightarrow\quad \Phi(z) = 97.5\% \quad\Longleftrightarrow\quad z = 1.96$$

That is, if we'd like to have a 2.5% chance of falsely rejecting the null in favor of $H_1 : \mu_2 > \mu_{302}$ we need that the sample mean $\overline{X}$ – and hence the standard-normalized version Z – lies above some value with a 2.5% probability (consider for a moment why we don't care in this case if the sample mean lies below some value with a 2.5% probability).

Then if our observed z-value is above 1.96, we should reject the null. Here, our observed z-value is $3.6341 > 1.96$, so we can safely reject the null. Therefore, we know with 2.5% significance that my commute along the Metro 2 line takes more time than along the Metro 302 line.

[1] Tests are performed assuming the null is true since this will give us a worst-case ability for rejection of the null. That

Question 2:

Question 3:

is, if we assumed somthing other than the null were true we'd be free to more or less assume anything; the assumptions we made would have an effect on the result of the test (rejection or failure of rejection). We could then reject the null arbitrarily, depending on the assumptions we made. If we assume the null is true and reject it anyway, this is the best possible evidence that it is not the case. This construction adds great strength to our notion of falsifiability.