Hypothesis Testing

Point estimation *qua* point estimation -- the simplest form of inferential statistics -- is actually quite rare in psychology. We usually don't just want to learn the value of, for example, the population mean, μ, or even set up a confidence interval for μ; more often, we want to know if μ is some specific value (or, conversely, not the same as some specific value). In other words, we want to test the working hypothesis that $\mu = V$, where V is some theoretically important value.

OK, now I know that you're probably saying to yourself: "self, I rarely want to test whether $\mu = V$, either. I run experiments, so I usually want to know if two population means are the same or different." Well, as I'll show a bit later, almost all cases of comparing two means actually boils down to being the same thing as testing something against a fixed value, V, so, actually, you do want to know how to do this. But, more than this, it is easier to introduce the logic of hypothesis testing using single sets of values against a target, so that's how I'm going to do it. We'll expand it to more complicated situations later, I promise.

The Logic of Hypothesis Testing

The basic logic of hypothesis testing is the exact opposite of what you might think. This is why some people -- especially Bayesians -- refer to the method as being "bass-ackwards." When you want to know if $\mu = V$, you do not calculate a probability that $\mu = V$ given your sample and its mean, $\overline{X}$. Instead, you actually calculate the probability of having observed $\overline{X}$ on the assumption that μ really is equal to V. (Please read the previous sentence more than once.) If this probability turns out to be very, very small -- in most empirical sciences, including psychology and neuroscience, less than 5% is the rule -- then you reject the idea that $\mu = V$. Contrariwise, if the probability of getting $\overline{X}$ assuming that $\mu = V$ is not very small, then you retain the previous idea that $\mu = V$.

This is very different from point estimation. In point estimation, we make a best guess about the population value (or a range of values, as in confidence intervals) and assign a probability to this best guess. In contrast, when we have a specific hypothesis to test, we do the testing by calculating the probability of observing the data that we have, assuming that the hypothesis is true. We do not calculate the probability of the hypothesis being true given the data; we calculate the probability of getting the data given the hypothesis. Note that this isn't just a difference between point estimation and hypothesis testing; it is also one of the critical differences between traditional hypothesis testing and the Bayesian approach. Bayesians do point estimation in the same way as hypothesis testers. But they don't switch over to the "bass-ackwards" approach for hypothesis testing; they continue to assign probabilities to hypotheses, instead.

I hope that you all appreciate how hard it was for me to write that last part. I am not a Bayesian. I worry that Bayesians will cause great damage to empirical science. And, yet, I managed to write some (snotty) sentences about hypothesis testing from their point of view. (Of course, I'm also a fan of Sun Tzu who said: "to know your enemy, you must become your enemy" and "if you know your enemies and know yourself, you will not be imperiled in a hundred battles," so don't take those few sentences as evidence that I'm changing my mind.)

Similar to point estimation, for hypothesis testing we use the sample plus some assumptions to estimate the spread and the shape of the sampling distribution of $\overline{X}$. (In fact, we use exactly the

same methods and assumptions; spread = $s/\sqrt{N}$ and shape = normal.) But we don't center this hypothetical distribution on $\bar{X}$, as we did for creating a confidence interval. Instead, we center this distribution on the to-be-tested value, *V*. Then we look up (using a *t* table, not a *z* table, because we had to estimate spread) the probability of observing a value of $\bar{X}$ that is as far from the center of this distribution as the one that we got. Again, if this probability is very small (under 5%), then we say "this $\bar{X}$ clearly did not come from this distribution, which is centered on *V*, so *V* can't be the correct value for μ." In statistical parlance, we "reject the null hypothesis" (since it was the null hypothesis that told us that μ was equal to *V*).

☞ Note: logically, when the probability of $\bar{X}$ is very low, we could try to keep the idea that $\mu = V$ and reject one or more of the assumptions that we have also made along the way, such as the assumption that the sampling distribution is normal or that $s/\sqrt{N}$ is a good estimate of error. But we don't do this. We only ever reject the idea that $\mu = V$. If we didn't do this -- if we opened the door to the idea that other things could be "wrong," instead -- then we would never get anywhere. Note, however, that this special, un-questionable status that we give to the assumptions implies that making sure that these assumptions are warranted is very important. That's why we covered these before now and why careful statisticians always test their assumptions.

Risk and Power

The output from a null hypothesis statistical test is dichotomous (i.e., it has only two values): you either "reject" the null hypothesis (i.e., you conclude that $\mu = V$ is not true after all) or you "retain" the null hypothesis. This can be thought of as the *decision* that is made by the statistical analyst.

The status of the null hypothesis (H_0) -- e.g., $\mu = V$ -- as a statement about the world, is also dichotomous: it is either true or false. This can be thought of as *reality*.

These two dichotomies set up or create a two-by-two table that describes the relationship between the decision that was made and reality. In two of the four cells, the analyst made the correct decision; in the other two, the analyst made an error.

	Reality: H_0 is true	Reality: H_0 is false
Decision: H_0 retained	*correct retention* $1 - \alpha$	*miss* or Type II error β
Decision: H_0 rejected	*false alarm* or Type I error α or "risk"	*correct rejection* or *hit* $1 - \beta$ or "power"

The two kinds of error are very different and occur under different realities: one can mistakenly reject a null hypothesis that really is true, and one can mistakenly retain a null hypothesis that really is false. These are called *false-alarm* (Type I) and *miss* (Type II) errors, respectively. The probability of making a false alarm -- which assumes that H_0 is true -- is denoted by α and is a value that is set by consensus. In most sciences, including psychology, it is .05 or 5% (which is the complement to the 95% that we use for confidence intervals, as mentioned above). The

probability of making a miss -- which assumes that H_0 is false -- is denoted by β. This depends on a variety of factors, only some of which are under the direct control of the experimenter. In psychology, there is a movement to try to set the value of β at .20 or 20%.

The other name for the probability of making a false-alarm error is *risk*. There is no special name for the probability of making a miss error, but the probability of correctly rejecting a false null hypothesis is also called *power*. Thus, in psychology we set risk to 5%, while also aiming for 80% power.

Why are the values of risk and power lopsided? Why do we allow ourselves to make more miss errors than false-alarm errors? Why we allow any errors at all?

Let's start with the last question first. We allow for errors because we have to. The normal distribution goes to infinity in both directions. Therefore, in point estimation (as already mentioned), the only way to set up a 100% confidence interval is to have lower and upper bounds of $-\infty$ and $+\infty$, respectively, which would make the CI useless. Now note that the 5% rule for hypothesis testing is merely the complement to the 95% rule for confidence intervals. Thus, never making false alarms in hypothesis testing is the same as requiring 100% confidence interval. And for the same reason that a 100% CI goes from $-\infty$ to $+\infty$, making anything possible, you can never reject the null hypothesis if risk is reduce to 0%.

As to the issue of lopsidedness, this comes from the conservative bias of most empirical science (cf. "reality has a well-known liberal bias," Colbert, 2006). In brief, because a false-alarm error is usually much more damaging to the field than a miss error (because, *inter alia*, the former will be published and make folks think that the hull hypothesis is false, while the latter probably won't ever be published), we need to avoid false alarms much more than misses. Furthermore, sciences have an easier time self-correcting for misses than for false alarms. If I told you that someone has already tested your (pet) theory's main prediction and found no evidence for it (which you view as a miss), you'd probably come up with a reason why their experiment was no good and go run your own test (and hopefully, for your sake, correct the miss). But if I told you that someone else's (pet) theory has already been tested and evidence in its favor was found (which you want to view as a false alarm, because you don't like their theory), you'd be much less likely to rerun the experiment (hoping to find no evidence this time around); plus, you'd probably worry that your experiment -- assuming that it failed to get any evidence in favor of the other person's theory -- just produced a miss error and, therefore, won't be published, so your replication won't help the field to self-correct the mistake.

The other main (and better) reason for the lopsided nature of risk and power needs to wait until we've had a chance to discuss the factors that influence power. We'll do that after *t*-tests.