Hypothesis Testing, Type I & II Errors, and Power

The logic of hypothesis testing is not always clear to people, even after taking the statistics course. Herein is a brief summary of notions of hypothesis testing.

Let's talk in terms of z-scores, because then we can use the normal distribution tables. Further, let's assume that we're talking in terms of sample means, which means that the distribution we're using is the sampling distribution of the mean. Thus we're also assuming that \( \sigma \) is known, so the sampling distribution of the mean has a standard error of \( \frac{\sigma}{\sqrt{n}} \). As you may recall, the sort of context in which we might make use of such statistics is as follows.

Suppose that you are interested in seeing if a group of 25 students has been sampled randomly from the normal distribution of IQs (i.e., a population with \( \mu = 100 \) and \( \sigma = 15 \)). What you are saying, in essence, is that you want to test the null hypothesis (H\(_0\)) that your sample of n=25 came from the population with \( \mu = 100 \). The alternative (H\(_1\)) is that the mean of the population from which you drew your sample was different from 100 (i.e., H\(_1\): \( \mu \neq 100 \)).

If the null hypothesis is true, your sample mean would have been drawn from a sampling distribution of means with a mean (\( \mu \)) of 100 and a standard error of 3. Essentially the null distribution would be illustrated below:

As you can see, virtually all the distribution (~100%) falls between 91 and 109. Although scores below 94 are unusual and although scores above 106 are unusual, they do occur occasionally.

Using our ordinary approach to hypothesis testing, and an \( \alpha \)-level of .05, we would retain the null hypothesis for any sample means that fell between \( \pm 1.96 \) standard deviation units (z-score units) of the mean. This is because 47.5% of the scores fall between the mean (100) and 1.96 standard deviations above the mean. Similarly, 47.5% of the scores fall between the mean and -1.96 standard deviations. Thus, 95% of the scores from this distribution would fall between \( \pm 1.96 \) standard errors and 5% of the scores would fall beyond \( \pm 1.96 \) standard errors.

We would reject the null if our sample mean was greater than 1.96 or less than -1.96 standard deviation units of the mean. Thus, in this example, sample means greater than 105.88 \textit{would} occur about 2.5% of the time. Likewise, sample means less than 94.12 \textit{would} occur about 2.5% of the time. So, taken together, in this distribution about 5% of the time we would get sample means less than 94.12 or greater than 105.88. However, because these sample means occur so infrequently, we feel that it is safer to reject H\(_0\) and conclude that the sample means were actually drawn from a different population. (Note that we could be making a Type I error and the scores \textit{really did} come from a population with \( \mu = 100 \).

So a sample mean of 108 would lead us to reject H\(_0\). In so doing, we would be saying that we had selected our sample from a population whose mean was \textit{greater} than 100. Notice that we wouldn't simply say \textit{different from} 100, because if the score fell in the upper rejection region it would probably have been produced by sampling from a population with a mean greater than 100.

Okay, so we've rejected the H\(_0\). Are we correct? We'll never know. However, we can rest assured about some things. For instance, we know that if we're correct in rejecting the null, the sample
was drawn from a population with $\mu \neq 100$, and likely $> 100$. That is to say, we've sampled some other population, not the null population. That's why I typically show the "Real" distribution as well as the "Null" distribution (see below), to indicate that we might be sampling from some distribution other than the null distribution. In this instance, the Real distribution would have a larger $\mu$ than the Null distribution.

We also know that if we're incorrect, and have made a Type I error, that should only happen 5% of the time. It happened because the random number god was acting up and gave us a sample comprised largely of people from the upper tail of the null population (with $\mu = 100$). So if our research hypothesis is totally off the wall and the sample is drawn from the null distribution, we will get scores extreme enough to lead us to falsely reject the null hypothesis only 5% of the time. There should be some comfort in this fact as well. That is, should the null hypothesis be correct (and our sample was drawn from the null distribution) we would most often be led to retain the null hypothesis (95% of the time).

Suppose, however, that we could know for sure that we've been sampling from the Real distribution (with $\mu > 100$). (Of course, in real life we'd never know that this was the case.) We're still testing the likelihood of the null hypothesis, so scores greater than 105.88 will convince us that we've sampled from a distribution with $\mu > 100$. The question then becomes, "How likely is it that we'll get scores greater than 105.88 if we're sampling from the Real distribution?" In other words, what proportion of the Real distribution would give us scores greater than 105.88? As you can see in the figure below, that proportion is fairly large (maybe .70 or so). So the area of the Real distribution shaded with the oblique stripes represents the proportion of time that you'd correctly reject $H_0$, that is, it represents the power of your statistic.

Note that even though the Real distribution has $\mu > 100$, it would still occasionally yield sample means less than 105.88 (the portion of the curve with the horizontal shading). If we sampled from the Real distribution and obtained one of these sample means, we would be led (incorrectly) to retain $H_0$. Thus, we would be making a Type II error, which happens with probability of $\beta$. Although we can fix $\alpha = .05$, we cannot fix the probability of $\beta$. In fact $\beta$ is determined by the parameters of the Real distribution. If $\mu$ for the Real distribution is much larger than the $\mu$ of the hypothetical Null distribution, then $\beta$ would be small. If $\mu$ for the Real distribution is only a bit larger than the $\mu$ of the hypothetical Null distribution, then $\beta$ would be large.