7 SIGNIFICANCE TESTS

7.06 Type I and type II errors

Think about a courtroom trial. The point of departure is that the accused is innocent. The prosecutor tries to convince the jury or the judge that the accused is guilty. The burden of proof to convince jury or judge is on the prosecutor. The defendant is found guilty only if the prosecutor presents strong evidence against the defendant's presumed innocence. In a trial, there are four possible outcomes. First, the defendant could be convicted while he or she is indeed guilty. That's a correct decision. Second, the accused could be judged not guilty while he or she is indeed innocent. That's a correct decision too. Third, however, it is also possible that a defendant is convicted who is actually innocent. That's a wrong decision. And fourth, an accused who is in fact guilty could also be set free. That's a wrong decision too.

This is exactly what might also happen if we conduct a significance test. The point of departure that the accused is innocent is analogous to the null hypothesis being true. And the situation where the defendant is guilty is the equivalent of the null hypothesis being false. Convicting the defendant is analogous to rejecting the null hypothesis and setting him or her free is the equivalent of not rejecting the null hypothesis. This leads to four possible scenarios. In two of them you make a correct decision. This is the case if the null hypothesis is true and you don't reject it or if the null hypothesis is false and you do reject it. But it is also possible that you make a wrong decision. This is the case if the null hypothesis is true but you decide to reject it, or if the null hypothesis is false but you decide not to reject it. The first wrong decision is what we call a **type I error** or a **false positive**, and the second one is what we call a **type II error** or a **false negative**.

Let me give you an example. Imagine your null hypothesis is that in the American population of certified scuba-divers 50% have more than 35 hours of diving experience. In other words, Pi equals 0.5. The alternative hypothesis is that it is another percentage. In other words, Pi is not 0.5. You asked a simple random sample of 500 American divers how much diving experience they have, and you find that a proportion of 0.56 has more than 35 hours of diving experience.

Now suppose that your null hypothesis is actually true. A type I error occurs if you decide, on the basis of your sample data, to reject the true null hypothesis. If the null hypothesis is true, the sampling distribution looks like this. If your significance level alpha is equal to 0.05, this is your rejection region. In the z-table you can find that z-scores of -1.96 and 1.96 form your critical values. Now, if your test statistic falls within the rejection region, and is, in other words, a value more extreme than the critical values, you reject the null hypothesis. The probability that this happens is 0.025 plus 0.025 equals 0.05. This means that the probability of a Type I error is equal to the significance level. Or, in other words, P(type I error$|H_0$ is true) equals alpha.

It might therefore seem to be tempting to just decrease the significance level. This is, however, not necessarily a good idea. If you decrease the probability of rejecting the null hypothesis while it is actually true, you increase the probability of *not* rejecting a null hypothesis that is actually false. In other words, by decreasing alpha, you decrease the probability of making a type I error, but you *increase* the probability of making a type II error. The probability of making a type II error is what we call beta. It's complicated to compute beta; it depends on various factors, such as the value of alpha, the sample size and the true value of the parameter. For that reason we won't compute the value of beta here. But it is important that you realize that when we try to decrease the probability of one type of error, the probability of the other type increases.

When the null hypothesis is false and you're conducting a test, you want the **power** of that test to be high. The power of a test is the probability of rejecting the null hypothesis given that it is false. Or, in other words, the power of a test equals 1 minus the probability of a type II error. That is the same as 1 minus beta. Why is power so important? Well, before you've conducted a study, it can help you to determine how many participants you need. After you've conducted a study it can help you, for instance, to make sense of results that are not statistically significant.

One final note: in practice you will never know if a decision was correct or not. The only thing we can do is control the *probability* of making an incorrect decision.