Statistical Tests - 2

The Power of a Test

In the first part we relied mostly on controlling Type I errors, but what about Type II? As defined, it measures the risk of accepting a hypothesis that is false. A moment's reflection shows that this probability, often denoted with the letter β, depends on what the true value is. Since we don't know the true value (why would we be doing statistics, if we did?), it is more appropriate to talk about a *function* $\beta(\mu)$, which depends on what the true value of the expectation is. So, to be as clear as possible:

> $\beta(\mu)$ is the probability of not rejecting the Null Hypothesis, when the true value of the mean is μ. In particular, $\beta(\mu_0)$ is precisely the probability of accepting that $\mu = \mu_0$ if that's indeed the case. It follows that $\beta(\mu_0) = 1 - \alpha$, 1 minus the probability of a Type I error. That's not the interesting case, but it sets the starting point.

Actually, it is more common to discuss in terms of $1 - \beta(\mu)$, called the *power* of the test—but this is only a question of taste.

> The power function, we'll denote it here with a capital Greek pi letter, $\Pi(\mu)$, is *the probability of rejecting the Null Hypothesis, if the true value of the mean is* μ. Just as we mentioned above, we have that $\Pi(\mu_0) = \alpha$, the probability of a Type I error.

To understand how we can use the notion of power (or of Type II error) to clarify what our testing really says, let's take one of the examples from the previous module, and look closer at the alternate hypothesis.

Consider the case where our sample, with $n = 9$, returned a sample mean $\bar{X} = 10.5$, in a test in which the Null Hypothesis was $\mu \leqslant 10$, with the variance assumed to be known and equal to 1. Like we noted already, the p-value of this test was 0.144, which is pretty high, so that it would not lead us to reject the hypothesis.

But what if we were wrong? Well, suppose μ was actually larger than 10. Let's assume that we have set our significance level to a fixed value, say $95\% = 1 - \alpha$. Then, we have a threshold value, precisely

$$x_\alpha = \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}} = 10 + 1.645 \cdot \frac{1}{3} = 10.548$$

(z_α is a common symbol to indicate the value such that, for a standard normal random variable Z, $P[Z > z_\alpha] = \alpha$). If $\bar{X}$ is greater than this threshold value we would reject the hypothesis, but if it is less we would not. Now, if the true value was $\mu \geqslant \mu_0$, with our protocol, the probability of rejecting the Null hypothesis would be the probability of a normal random value, mean μ, standard deviation $\sigma/\sqrt{n} = 1/3$ (because that's what we already know), to be greater than 10.548. In formulas, denoting by Y_μ such a variable,

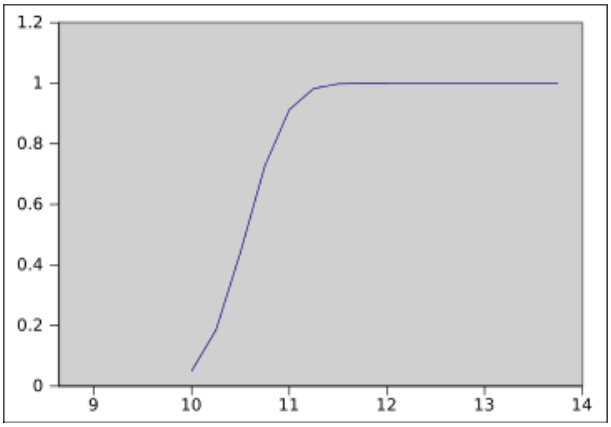
$$\Pi(\mu) = P[Y_\mu > x_\alpha] = P\left[\frac{Y_\mu - \mu}{\sigma/\sqrt{n}} > \frac{x_\alpha - \mu}{\sigma/\sqrt{n}}\right] = P\left[Z > \frac{10.548 - \mu}{1/3}\right]$$

This is function of μ, and we can use tables or, more conveniently, a spreadsheet, to compute some of its values. Programing the function `1-normsdist(31.645-3*[μ])` (here, $[\mu]$ is the spreadsheet cell where we wrote in a value for μ), we can come up with the following table (obviously, all values are rounded)

μ	Π
10	0.05
10.25	0.185
10.5	0.442
10.75	0.727
11	0.912
11.25	0.982
11.5	.998
11.75	.9998
12	1
12.25	1

Table 1.

It may be more striking to graph this table:



How should we read this table? If we set ourselves the goal to be right 90% of the time, we learn that we will be right in rejecting the Null Hypothesis whenever the true value of the mean is 11 or higher. If the true value is closer to 10 than that, things are not as clear cut. Even if $\mu = 10.75$, we will correctly reject the Null Hypothesis 73% of the time, but that means we will have been going overboard 27% of the time, which is a bit uncomfortable.

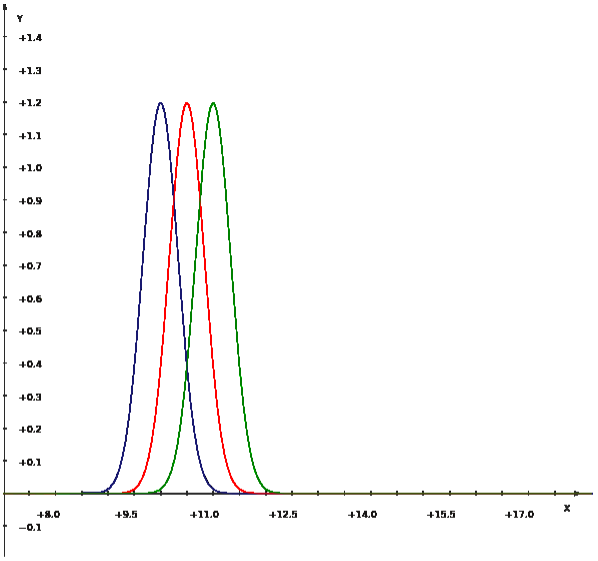
"Resolution Power"

As should be clear by now, statistical statements are not as clear cut as we might like them to be. However, we can now look at our hypothetical experiment and be a bit more specific:

> By setting a 95% confidence level, in not rejecting the hypothesis that $\mu \leqslant 10$, we can safely say that we will be able to exclude that the true value is 11 or greater (we will be right 90% or more of the time). We cannot be as confident that the true value is not somewhere between 10 and 11.

Although this is not standard terminology, we can think of this calculation as zeroing in on the *resolution power* of the test. Imagine you are taking a photo. There are two sources of light out there. Each sends a ray of light (a "photon", more precisely), and this ray generates a blip on your film or your solid state light receiver. If the two blips are too close, you won't be able to tell them apart. There is a minimum distance between the blips that will allow you to say that there were two sources. This is the "resolution power" of your camera/lens/film combo. Similarly, tests have a "resolution power": they can tell when two hypotheses are far enough apart to be sorted out, but they cannot distinguish between two situations when the two alternatives are too close together.

To drive this further home, let's look at the graphs for the distributions of normal random variables with standard deviation $\frac{1}{3}$ (that's the standard deviation of our means), and, respectively, $\mu = 10$ (blue), $\mu = 10.5$ (red), and $\mu = 11$ (green). Thinking of these as the blobs produced by photons hitting on a film plate (these would be, approximately, Gaussian-shaped too), you may appreciate how it is much harder to tell the blue and red apart, than it is to tell the blue and green:



Now, pushing our hypothetical experiment further, what if we really needed to be able to tell 10 and 10.5 apart? Clearly, our experiment doesn't cut it. We need another experiment, but how can we make it do the job? Well, as you can imagine, if we have no control on σ (that is, we have no access to a more precise instrument), we can still improve our resolution power, because we are testing means: their standard deviation—which determines the spread of the "bell curve"—is $\sigma/\sqrt{n}$, hence to have tighter curves, and hence less overlap for given values of μ, we can increase n. If your test does not have a good enough resolution power, one way to improve it is to take a larger sample. It may take a much larger sample: notice how your standard deviation will decrease as the *square root* of n: to decrease it by a factor of 2, you need 4 times as many observations, but that's the way it is.

Other Tests

The discussion here has concentrated on the simplest possible test: testing for the mean of a normal random variable, with known variance. Hopefully, the basic idea was highlighted, as we were not encumbered by subtleties like handling strange distribution families. With proper care, the same approach works in any other situation.

For example working on a two-tailed test (as in H_0: $\mu = \mu_0$, H_1: $\mu \neq \mu_0$), things are very similar, except you have to worry about both values smaller as well as larger than μ_0.

Also, it is pretty clear how we should proceed if the variance was unknown (the more common case). In this case, we would be working with $\sqrt{n}\frac{\bar{X}-\mu}{s}$, a random variable with t_{n-1} distribution. Again, we would repeat the same strategy, and working with a spreadsheet makes life much easier, since you have then access to the function tdist, the *survival function* ($1 -$ the cumulative distribution function) of the Student distribution (you will have to specify the x, the number of degrees of freedom, and whether you are considering a one-tailed or a two-tailed test). Otherwise, things follow the same steps.