avance

Converting life science technology into value

A statistical view on success rates

Null-hypothesis, type I/II errors

Statistics is a bit of the younger brother of mathematics. But it is still capable to confuse us quite a bit. This newsletter is about statistics and Bayes' theorem and will leave you still confused, but hopefully on a higher level.

The goal of a statistical test is to check whether a hypothesis is true or false. In a clinical trial we typically test whether a drug has a certain curing, moderating, inhibiting, or other desired effect. This hypothesis is compared to the null-hypothesis, which represents a default position, i.e. nothing happens, or the drug has no effect.

In order to do so many patients are observed. Unfortunately, the observed results do not always reflect the reality. We all know that patients sometimes show a positive reaction even though the drug they have been given has no effect at all – the so-called placebo effect. And the opposite can also be true, that some patients do not react (or fail to say that they improve) even though the drug actually has the desired effect.

In an unlucky setting the trial encompasses just such atypical patients and yields a misleading result. In general 4 different types of results are possible: Either the drug works or does not work and the trial correctly shows the effect. These two scenarios are what we expect from a trial, that it correctly tells us what is going on. But it is also possible that the trial indicates that the drug works even though it does not in reality. This is a type I error, where the null-hypothesis is wrongly rejected. It is also called a false positive (the trial reveals a positive result that is wrong). Imagine a pregnancy test that indicates that a man is pregnant. This would be a type I error. The last possible scenario is a false negative, i.e. the trial makes us believe that the drug does not work even though it does work. This is a type II error. Coming back to the pregnancy test a pregnant woman would receive the result that she is not pregnant.

Table 1: Possible scenarios. The greyed scenarios correspond to a successful trial.

	Drug works	**Drug doesn't work**
Positive trial result	Correct result	Type I error False positive
Negative trial result	Type II error False negative	Correct result

What does it mean in business?

Obviously, everybody hopes that the drug works and that the trial correctly shows so. The development can then be continued as planned.

For a biotech company one might think that a type I error (or a false positive) is the best of all catastrophes. The drug doesn't work but evidence supports the continuation of the trials. Even though the drug does not work – but nobody knows – it is regarded as a success and even increases in value. While this is very nice in the short-term disaster strikes just at a later point in time, when even more money has been spent.

As a contrast a correct negative result is very disappointing but allows

you to focus your money on projects that actually might work. If you fail you better fail early.

A type II error is tricky. First, we need to be aware that nobody knows whether it is a type II error or whether the drug simply does not work. But the sheer possibility of a type II error lets many scientists and executives believe that maybe the trial result is wrong and it might be worthwhile running another trial. But in any case a type II error unfairly destroys value in the short-term. The drug actually works but fate plays an evil turn against the company and the hopeful patients.

What do the numbers say?

But how likely is each of the scenarios? The health authorities are predominantly worried about spending money on drugs that actually do not work. So if they approve a drug, they want to be fairly certain that it actually works. Therefore the limit of a false positive is set at 5%. This significance level is denoted by α and corresponds to the p-value of the trials.

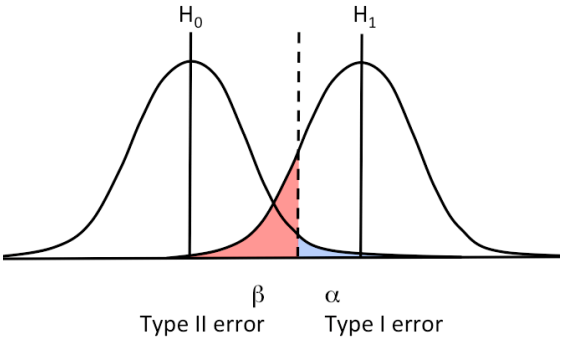

Figure 1: Type I and II errors in a weakly powered trial.

A trial sponsor on the other hand would like to keep the probability of a type II error as low as possible. The probability of a type II error occurring is denoted by β and can be governed by the power of the trial, which is $1-\beta$. A trial is more powerful if the probability of wrongly rejecting the new hypothesis, i.e. that the drug actually works, is less. A more powerful trial can typically be achieved by increasing the number of patients enrolled (but without increasing the observed parameters). So, the probability of bad luck can be reduced, but obviously at a cost.

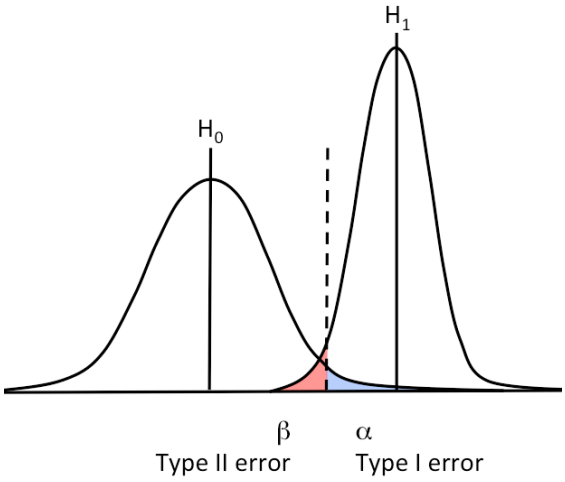

Figure 2: Type I and II errors in a highly powered trial.

Typically, 80%[1] is accepted as a fairly powered trial. But this means that for every 5 working drugs you hold in your hand only 4 pass and miss dearly needed revenues. Of course you can increase the power of the trial, but usually the power is determined by the budget available for the trial, especially in a biotech setting. It should be precisely the other way round.

[1] http://williampcoleman.wordpress.com/2007/11/14/clinical-trial-design-for-beginners/

Mammograms

Nate Silver[2], the American star statistician, described how type I errors successfully prevent doctors to run mammograms in younger women. A mammogram incorrectly claims that a woman has breast cancer even though she has not in 10% of the cases. These are false positives, or type I errors. On the other hand, if the woman does have cancer, the mammogram detects it in 75% of the cases. The type II probability is therefore 25%. These rates sound fairly good. But let's consider that actually only 1.4% of women have breats cancer in her forties.

Table 2: Mammograms in women in their forties.

	Cancer	No cancer
Probability	1.4%	98.6%
Positive mammogram	75% of 1.4% **=1.1%**	10%*98.6% **=9.9%**
Negative mammogram	25% of 1.4% **=0.4%**	90%*98.6% **=88.7%**

The doctors face the dilemma that even though in 1.1%+88.7%=89.8% of the cases the mammogram returns a correct result a positive mammogram doesn't mean at all that a woman has breast cancer, quite the contrary. The mammogram will be positive about 1.1%+9.9%=11% of the time, but only 10% of these positives are true positives. A mammogram is quite good to rule out breast cancer, but not to diagnose breast cancer – as long as the overall probability to develop breast cancer is so low. This is the reason why women do not get mammograms until they are older and have a higher probability of developing breast cancer.

Clinical Trials

Let us make a similar analysis for clinical trials. Success rates for phase 2 clinical trials can be as low as 25%-30%, for phase 1 but also phase 3 trials they can be significantly higher and go up to 80% (typically the extremes are not in the same disease area). How telling are positive and negative trial results based on the "true" probability whether the drug should pass the trial or not?

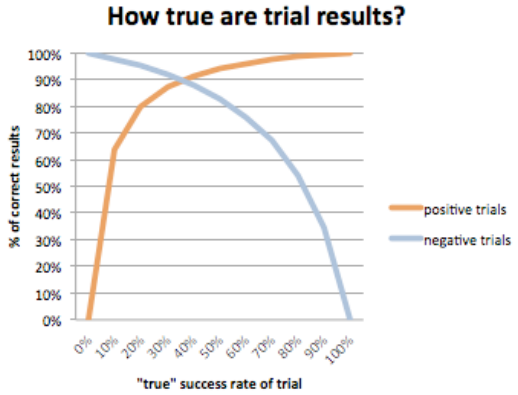

Figure 3: Trustworthiness of trial results depending on the true probability.

If we assume α=5% and β=20%, then the trials become already quite trustworthy (80%) even if we have a very low success rate of 20%

The above figures can also be obtained with Bayes' theorem, where the original probability is our initial guess of how true the hypothesis that the drug works really is. After observing the trial results we can modify that original assumption and become fairly certain that the drug works or not, depending on whether the trial was a success or a failure.

[2] Nate Silver, „The Signal and the Noise".

Again, the graph displays the tantalising possibility that the drug could actually work despite a negative trial result. At a success rate of 60% - fairly common for phase 3 trials – only 76% of the negatives are true negatives.

Figure 4 displays how better powered trials improve the picture. N increased power makes the failures trustworthier, because the probability of a false negative is reduced. If the trial exhibits a high p-value then it is more likely to be in deed true, as shown in the figure.

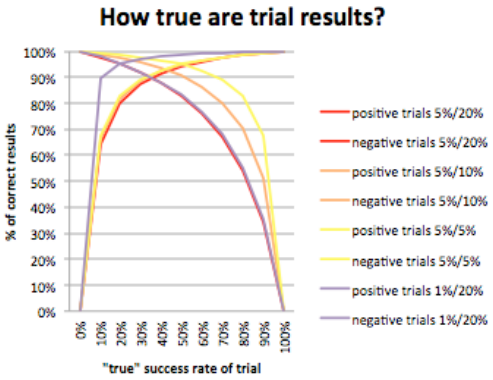

Figure 4: Trustworthiness of trials depending on significance and power.

That remaining uncertainty in a trial might lead to some attrition in the subsequent trial. While the phase 2 results are fine, it could have been a type I error and only the phase 3 trial shows that the drug actually does not work – or not sufficiently well. Figure 2 displays that especially trials with a low success rate still leave a high uncertainty even after positive results. This is also reflected by disease areas with low phase 2 success rates. These are usually followed by low phase 3 success rates again, like CNS. Of course, the attrition can also be attributed to a variety of other causes, but the false positives aggravate the situation.

An additional subtlety arises from the fact that the success rates are computed from trial results that are, as shown, a little flawed. In order to get the "true" success rate that is displayed on the x-axis, we need to reverse engineer the success rates as well.

In such a situation mathematicians and statisticians alike tend to say: This is left as an exercise.