1 Binomial distribution

The underlying process is constructed as a sequence of N Bernoulli trials with success rate $p^\star$.

2 Normal approximation

Let's assume that we prepared an actual realization of the process defined above. Because of the limited sample size ($N < \infty$ in any practical situations), the success rate p follows a binomial distribution with a finite standard deviation (decreasing as $1/\sqrt{N}$) around $p^\star$. In the limit of not too small N we can apply the normal approximation which implies that the probability density function (PDF) of the observed success rates $f(p)$ is given by:

$$f(p) \sim N\left(p^\star\,,\,\sqrt{\frac{p^\star(1-p^\star)}{N}}\right) \tag{1}$$

3 One proportion test

Our objective is to assign a p-value telling us the probability that the success rate p of the observed realization is significantly different than some baseline rate p_0. The first step is to calculate the z-score according to:

$$z = \frac{p - p_0}{\sqrt{p_0(1-p_0)}}\sqrt{N} \tag{2}$$

where p_0 is the baseline rate that we are testing against. Note that, at this point, we did not take into account the fact that the true underlying success rate is in fact $p^\star$ as would actually happen in a real experiment where the true rate would in fact be unknown. A small calculation show that the PDF of the observed z-scores should behave as:

$$f(z) \sim N\left(\mu_z,\ \sigma_z\right) \quad \text{with} \quad \begin{cases} \mu_z & = (p^\star - p_0)\sqrt{\frac{N}{p_0(1-p_0)}} \\ \sigma_z & = \sqrt{\frac{p^\star(1-p^\star)}{p_0(1-p_0)}} \end{cases} \tag{3}$$

4 One tailed p-value

One can then extract the one tailed p-value from the z-score as follows:

$$p = 1 - \Phi\left(z\right) \tag{4}$$

where Φ stands for the cumulative distribution function of the standard normal distribution $N(0,1)$. Another trivial calculation shows that the PDF of the observed p-values $\Psi(p)$ should follow:

$$\Psi(p) = \frac{e^{-\mu_z^2/2\sigma_z^2}}{\sigma_z} \exp\left[\left(1 - \frac{1}{\sigma_z^2}\right) \mathrm{erfinv}^2(1-2p)\right] \exp\left[-\frac{\sqrt{2}\mu_z}{\sigma_z^2}\mathrm{erfinv}(1-2p)\right] \tag{5}$$

where erfinv stands for the inverse error function.

5 Probability of significant p-values

We are now in a position to answer the original question: For a given tuple $\{p^\star, p_0, N\}$, what is the probability of observing a significant p-value? Well, this is simply given by the following equation:

$$\beta(p) = \int_{0.95}^{1.0} \Psi(p) \tag{6}$$

which can be estimated numerically.

Let's consider the special case when the true success rate $\tilde{p}$ is equal to the baseline rate p_0. In this case, the z-scores are distributed according to the normal distribution $N(0,1)$ and p-values are distributed uniformly between $[0,1]$. In this case, the probability of observing a significant p-value is equal to 5% regardless of the sample size. In other words, there is a 5% chance of concluding that the empirical data is better than the baseline rate even though we know that they are actually identical to each other. (Note that this is precisely what one would expect of a p-value...)

On the other hand, let's consider the case when $\{p_0 = 0.1; p^\star = 0.11; N = 2500\}$. This represents a true 10% increase compared to the baseline backed up by quite a lot of observations. Surely, one would expect that the p-value would be very significant. Surprisingly, it turns out that the probability of observing a significant p-value is only about 50%. This implies that in spite of these quite strong metrics we have a 50% chance of concluding that the experiment is worse than the baseline due to the slow convergence of β...

Note that the same analysis can also be carried out for situations in which the experiment is known to be worse than the baseline $p^\star < p_0$. For example, let's consider $\{p_0 = 0.1; p^\star = 0.096; N = 1000\}$. Despite the fact that the experiment is constructed to be 4% worse than the baseline and that we observed 1000 events, there is still an almost 2% chance that we may conclude the experiment to be (wrongly) better.