Sample size and z test for two independent Bernoulli samples with the same size

1 Introduction: things to keep in mind about ...

1.1 Bernoulli random variables and normal approximation of the sample mean

Given $X_1, \dots, X_{n_1}$ a random sample of size n_1 drawn from a Bernoulli population X of parameter p_1 and $Y_1, \dots, Y_{n_2}$ a random sample of size n_2 drawn from a Bernoulli population Y of parameter p_2 independent of X, we are interested in studying the relationship between sample size and effect siz e $|p_1 - p_2|$ in comparing the sample mean $\overline{X}_{(n_1)} = \frac{1}{n_1}\sum_{i=1}^{n_1} X_i$ and $\overline{Y}_{(n_2)} = \frac{1}{n_2}\sum_{i=1}^{n_2} Y_i$ at a given significance level α and, in a second step, with a given power $1-\beta$. For our purposes we use samples of the same size, namely $n_1 = n_2 = n$ so we can refer to the sample means simply as: $\overline{X}$ and $\overline{Y}$.
Remark: in general the effect size for comparing two means is given by $\frac{|\mu_0 - \mu_1|}{\sigma}$; as it will be clear soon, here, since we use an upper bound for the standard deviation of a Bernoulli distribution, we can neglect σ.

The expected value and the variance of a Bernoulli random variable of parameter p are respectively p and $p \cdot (1-p)$. It is also easy to show that the variance is bounded above by 0.25, ad the maximum is reached at $p = 1/2$. Therefore $E(X) = p_1$, $var(X) = p_1 \cdot (1-p_1) \leq 1/4$ and $E(Y) = p_2$, $var(Y) = p_2 \cdot (1-p_2) \leq 1/4$. Being the sample mean an unbiased estimator for the expected value, it holds that $E(\overline{X} = p_1)$ and $E(\overline{Y} = p_2)$.

Let $D = \overline{X} - \overline{Y}$.
It holds that: $E(D) = p_1 - p_2$ and, since $X \perp Y$, it also holds that:
$var(D) = var\left(\overline{X} - \overline{Y}\right) = \frac{var\left(\sum_{i=1}^{n_1} X_i\right) + var\left(\sum_{i=1}^{n_1} Y_i\right)}{n^2} = \frac{p_1\cdot(1-p_1)+p_2\cdot(1-p_2)}{n}$, that implies

$$\sigma_D = \sqrt{\frac{p_1 \cdot (1-p_1) + p_2 \cdot (1-p_2)}{n}} \leq \sqrt{\frac{1}{2 \cdot n}} \tag{1}$$

Let D^* the standardization of D, namely

$$D^* = \frac{D - E(D)}{\sqrt{var(D)}} = \frac{(\overline{X} - \overline{Y}) - (p_1 - p_2)}{\sqrt{\frac{p_1\cdot(1-p_1)+p_2\cdot(1-p_2)}{n}}} \tag{2}$$

The sum of n independent and identically distributed Bernoulli random variables with parameter p follows a binomial distribution whose parameters are n and p. If we suppose $n \gg 1$, the binomial distribution can be approximated by a Normal distribution, and the approximation is already good starting from values of n on the order of magnitude of 10^2. In this approximation it holds that D follows a Normal distribution and $D^* \sim N(0,1)$, so it is possible to compute a sufficient sample size starting from the Normal random variable.
Remark: This observation is important because, in order to apply the Central Limit Theorem, n must tend to infinity; if the theorem is used to find the value of n, it is necessary to verify whether for the original distribution (in our case a Bernoulli distribution) that value is large enough to allow a good approximation of the sample mean with a normal distribution.

1.2 standard Normal variables

The cumulative distribution function of a standard Normal variable D^* is called Φ, that is $\Phi(x) = P(D^* \leq x)$.

As the standard Normal distribution is an even function, called z_α the α-quantile of the distribution, it holds that

$$z_\alpha = -z_{1-\alpha}, \text{ and, equivalently: } |z_\alpha| = |z_{1-\alpha}| \tag{3}$$

Remark: sometimes in the literature z_α is referred to as the upper quantile α and is such that $P(D^* \geq z_\alpha)$. It is easy to show that

$$P\left(|D^*| \leq k\right) = 2 \cdot \Phi(k) - 1 \tag{4}$$

Since a linear function of a Normal random variable is still a Normal random variable, it holds that

$$D \sim N\left(E(D), \sqrt{var(D)}\right) \tag{5}$$

1.3 hypothesis testing

Consider the **null hypothesis** $H_0 : p_1 = p_2 = p$ vs. $H_1 : p_1 \neq p_2$.

We commit a **type I error** if we reject H_0 when H_0 is true.

The **significance level** α of the test is the probability of committing a type I error, that is $\alpha = P_{H_0}$ (reject H_0).

Given a statistic T_n, a **critical region** C at significance level α is the region of rejection of H_0, that is C is such that $P_{H_0}(T_n \in C) = \alpha$.

We commit a **type II error** if we do not reject H_0 when H_0 is false.

If we call β the probability of committing a type II error, that is $\beta = P_{H_1}$ (not reject H_0), the **power** of the test is the probability $1 - \beta$ of rejecting H_0 when H_0 is false, that is $1 - \beta = P_{H_1}$ (reject H_0).

2 Sample size at a given significance level α for a two-sided test

Consider the null hypothesis $H_0 : p_1 = p_2 = p$ vs. $H_1 : p_1 \neq p_2$ at significance level α (assume $\alpha < 0.5$ without loss of generality). Requiring a **significance level** α is equivalent to requiring

$$P_{H_0}(|D > c|) \leq \alpha \tag{6}$$

Let us call $c > 0$ *critical value.*

If **If** H_0 **is true**, then $E(D) = 0$, and

$$\begin{aligned} \alpha \geq P_{H0}(|D| \geq c) = \\ P_{H0}(|D| \geq c) = P(|D/\sigma_D| \geq c/\sigma_D) = \\ P(|D^*| \geq c/\sigma_D) = 2 - 2 \cdot \Phi(c/\sigma_D) \end{aligned} \tag{7}$$

that implies

$$\Phi\left(\frac{c}{\sigma_D}\right) \geq 1 - \frac{\alpha}{2} \tag{8}$$

The minimum value c/σ_D satisfying the equation in (8) is therefore the $(1 - \alpha/2)$-quantile of the standard Normal distribution, and we **reject** H_0 if $|D^*| \geq z_{1-\alpha/2}$, that is if

$$\left| \frac{(\overline{X} - \overline{Y})}{\sqrt{\frac{2 \cdot p \cdot (1-p)}{n}}} \right| \geq z_{1-\alpha/2} \tag{9}$$

In order to turn $\left| \frac{(\overline{X} - \overline{Y})}{\sqrt{\frac{2 \cdot p \cdot (1-p)}{n}}} \right|$ into a statistic (getting rid of the unknown parameter p), note that

$$\sigma_D \leq \sqrt{\frac{1}{2 \cdot n}} \text{ implies } c/\sigma_D \geq \sqrt{2 \cdot n} \cdot c \tag{10}$$

and that, being Φ is a monotonic increasing function, it holds that

$$\Phi(c/\sigma_D) \geq 1 - \alpha/2 \text{ for } c/\sigma_D \geq z_{1-\alpha/2} \tag{11}$$

Therefore we can reason in the worst case

$$\left| \frac{(\overline{X} - \overline{Y})}{\sqrt{\frac{2 \cdot p \cdot (1-p)}{n}}} \right| \geq \left| \frac{(\overline{X} - \overline{Y})}{\sqrt{\frac{1}{2 \cdot n}}} \right| \geq z_{1-\alpha/2} \tag{12}$$

Solving with respect to n we obtain:

$$n \geq \frac{1}{2} \cdot \left(\frac{z_{1-\alpha/2}}{\overline{X} - \overline{Y}} \right)^2 \tag{13}$$

If we want to be able to detect a significant difference at least equal to d (effect size) at level α it is therefore sufficient that

$$n \geq \frac{1}{2} \cdot \left(\frac{z_{1-\alpha/2}}{\overline{X} - \overline{Y}} \right)^2 \geq \frac{1}{2} \cdot \left(\frac{z_{1-\alpha/2}}{d} \right)^2 \tag{14}$$

It is worth to note that n is inversely proportional to the square of the effect size.
Fig. 4 shows the sufficient sample size n as a function of effect size at different significance levels α.

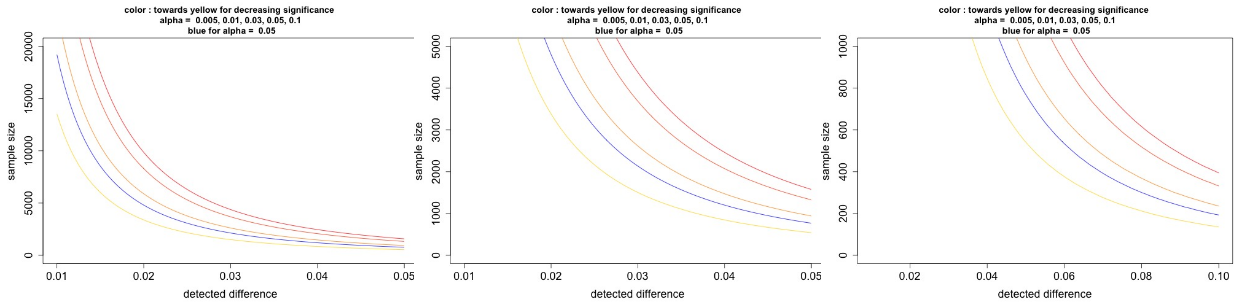

Figura 1: n as a function of effect size

3 Sample size at a given significance level α and power $1-\beta$ for a two-sided test

Now we take int account also type II error: not reject H_0 when it is false, whose probability is β.
If H_1 is true, we have to explicitly take into account the two means p_1 and p_2, since $E(D) = p_1 - p_2$. So let's restart from the very beginning. First of all recall that (type I error):

$$\begin{gathered} \alpha \geq P_{H0}(|D| \geq c) = 1 - \left(2 \cdot \Phi\left(\frac{c}{\sigma_D}\right) - 1 \right) \text{ that implies} \\ \Phi\left(\frac{c}{\sigma_D}\right) = 1 - \frac{\alpha}{2} \text{ and the critical values } c_1 = z_{1-\frac{\alpha}{2}} \cdot \sigma_D \text{ and } c_2 = -z_{1-\frac{\alpha}{2}} \cdot \sigma_D = z_{\frac{\alpha}{2}} \cdot \sigma_D. \end{gathered} \tag{15}$$

Then we require that (type II error): $P_{H1}(|D| \leq c) \leq \beta$, or, equivalently, that $P_{H1}(|D| \geq c) \geq 1 - \beta$, that is

$$\begin{aligned} 1 - \beta \leq P_{H1}(|D| \geq c) = P_{H1}(|D| \geq z_{1-\frac{\alpha}{2}} \cdot \sigma_D) = \\ P_{H1}\left(D > z_{1-\frac{\alpha}{2}} \cdot \sigma_D \cup D \leq z_{\frac{\alpha}{2}} \cdot \sigma_D\right) = P_{H1}\left(D > z_{1-\frac{\alpha}{2}} \cdot \sigma_D\right) + P_{H1}\left(D \leq z_{\frac{\alpha}{2}} \cdot \sigma_D\right) = \\ P\left(\frac{D - (p_1 - p_2)}{\sigma_D} > z_{1-\frac{\alpha}{2}} - \frac{(p_1 - p_2)}{\sigma_D}\right) + P\left(\frac{D - (p_1 - p_2)}{\sigma_D} \leq z_{\frac{\alpha}{2}} - \frac{(p_1 - p_2)}{\sigma_D}\right) = \\ P\left(D^* > z_{1-\frac{\alpha}{2}} - \frac{(p_1 - p_2)}{\sigma_D}\right) + P\left(D^* \leq z_{\frac{\alpha}{2}} - \frac{(p_1 - p_2)}{\sigma_D}\right) = \\ 1 - \Phi\left(z_{1-\frac{\alpha}{2}} - \frac{(p_1 - p_2)}{\sigma_D}\right) + \Phi\left(z_{\frac{\alpha}{2}} - \frac{(p_1 - p_2)}{\sigma_D}\right) \end{aligned} \tag{16}$$

Taking the upper bound $\sqrt{\frac{1}{2 \cdot n}}$ for σ_D, the plot of $1 - \beta$ against $(p_1 - p_2)$ for different significance levels and in the particular case of $n = 100$ is shown in Figure 2, while in Figure 3 the power function is plot for different sample sizes in the particular case of $\alpha = 0.05$.

Note that $1-\beta$ is an even function, see the plot it against $|p_1 - p_2|$. Therefore we can consider the case $(p_1 - p_2) > 0$. It holds that $\Phi\left(z_{\frac{\alpha}{2}} - \frac{(p_1-p_2)}{\sigma_D}\right) < \alpha/2$, and we can neglect it, getting

$$\beta \lessapprox \Phi\left(z_{1-\frac{\alpha}{2}} - \frac{(p_1 - p_2)}{\sigma_D}\right) \tag{17}$$

Therefore

$$z_\beta \simeq z_{1-\frac{\alpha}{2}} - (p_1 - p_2)/\sigma_D \text{ and, for (3) it holds that: } -z_{1-\beta} \simeq z_{1-\frac{\alpha}{2}} - (p_1 - p_2)/\sigma_D \tag{18}$$

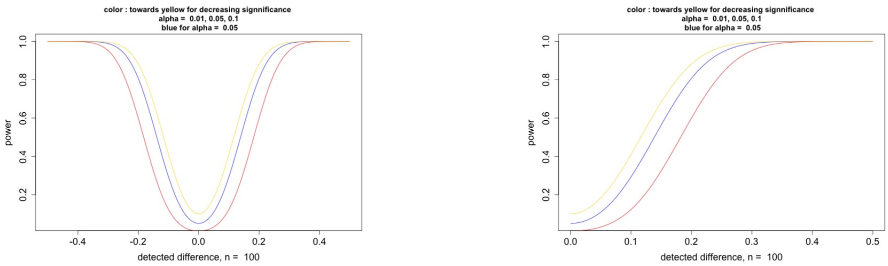

Figura 2: power as a function of effect size for different significance levels

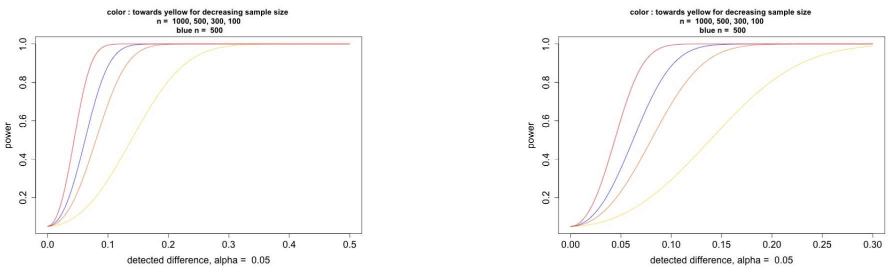

Figura 3: power as a function of effect size for different sample sizes

Considering the worst case, that is $\sigma_D \leq \sqrt{\frac{1}{2 \cdot n}}$, from:

$$z_{1-\beta} + z_{1-\frac{\alpha}{2}} \simeq d/\sigma_D \leq \sqrt{2 \cdot n} \cdot d \tag{19}$$

it is possible to compute the sufficient sample size n needed to detect a minimum effect size d at α significance level and power $1-\beta$, as follows:

$$n \geq \frac{1}{2} \cdot \left(\frac{z_{1-\beta} + z_{1-\frac{\alpha}{2}}}{d}\right)^2 \tag{20}$$

It is worth to note that n is inversely proportional to the square of the effect size. Fig. **??** shows the sufficient sample size n as a function of effect size at different powers at the given significance level $\alpha = 0.05$.

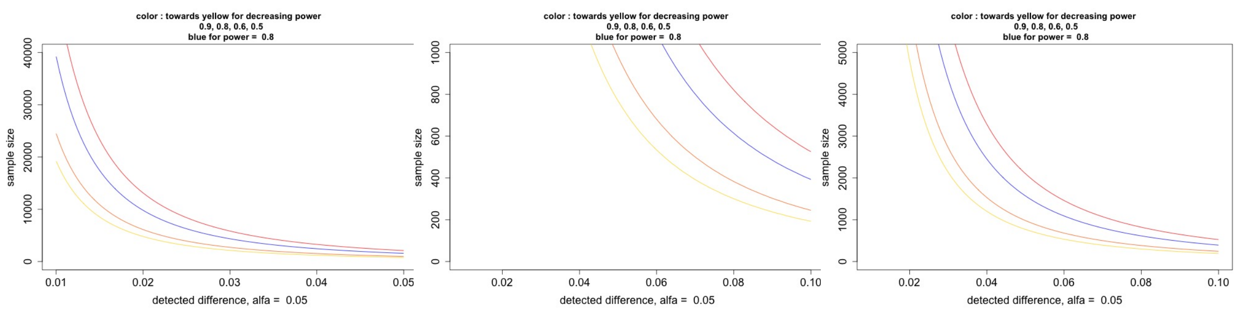

Figura 4: n as a function of effect size