Contents

1 Sample Size and Power **2**

1.1 Short review 2

1.2 Sample Size for Comparing Two Proportions 2

1 Sample Size and Power

The goal for most studies employing logistic regression is to test whether different predictors have an effect on a binary response variable.

We want to discuss the required sample size for detecting an effect of a given size.

1.1 Short review

When planning a study for estimating a success probability π with a $(1-\alpha)\times 100\%$ confidence interval within a margin of error of M, then the required sample size is

$$n \geq \pi(1-\pi)\ \left(\frac{z^*}{M}\right)^2$$

This is based on the large sample Z-confidence interval

$$\hat{\pi} \pm z^* \sqrt{\frac{\pi(1-\pi)}{n}}$$

with Margin of Error of $z^*\sqrt{\frac{\pi(1-\pi)}{n}}$.
Then we choose the sample size so that we can be certain the Margin of Error will not exceed M, resulting in the formula for the sample size given above.

1.2 Sample Size for Comparing Two Proportions

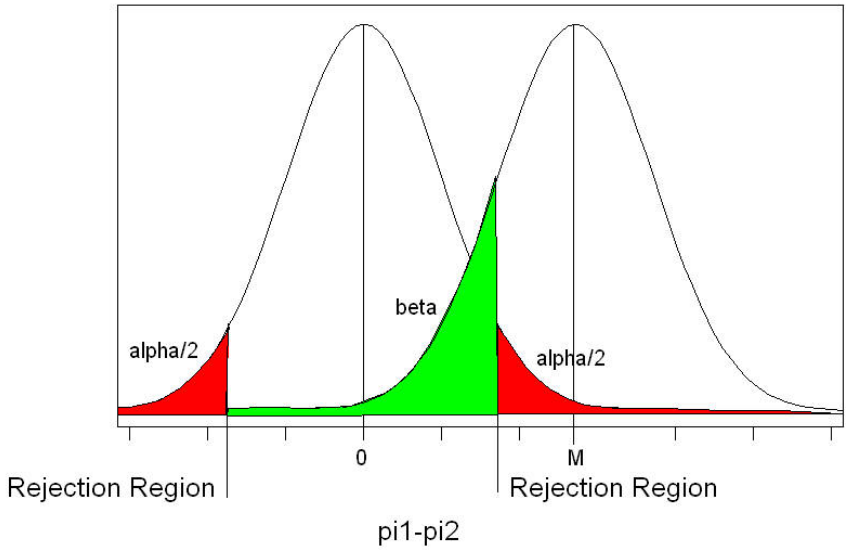
Most approaches for determining the sample size are based on a different approach.

We ask: What is the required sample size to detect an effect of a certain size using a statistical test at significance level of α, and and error probability for the error of type 2 of at most β (or power of $1-\beta$)?

For two proportions we would specify,

- How large is the effect we want to be able to detect: $M = \pi_1 - \pi_2$
- The acceptable error probability for an error of type 1: α
- The required power of the test if $\pi_1 - \pi_2 = M$: $1-\beta$

These choices mean: If $|\pi_1 - \pi_2| > M$, then the probability for not rejecting $H_0 : \pi_1 - \pi_2 = 0$ at significance level of α should not exceed β.
That is: when using a test at significance level of α, an error of type II shall be unlikely (β) if H_0 is violated by at least M.



Using equal sample sizes for the two samples requires approximately:

$$n_1 = n_2 = \frac{(z_{\alpha/2} + z_\beta)^2(\pi_1(1-\pi_1) + \pi_2(1-\pi_2))}{(\pi_1 - \pi_2)^2}$$

Note that we do not just need to know M but good estimates for π_1 and π_2.
The test for $H_0 : \pi_1 = \pi_2$ is equivalent to the test if the slope in a logistic regression model is zero, when the predictor is binary. Thus this is relevant for finding sample sizes in logistic regression.

Example 1
Assume we want to test if the probability for remission is the same for treatment A and treatment B. Former studies have shown that for treatment A 75% of patients achieve remission and we hope for treatment B to improve this rate by at least 5%.
We will use a test at significance level of 5% and require a power of at least 80% if this rate of improvement is correct.
The required sample size is at least

$$n_1 = n_2 = \frac{(1.96 + 0.84)^2(0.75(0.25) + 0.80(0.2))}{(0.75 - 0.80)^2} = 1089.76$$

Thus we would need at least 1090 observations for each treatment. (Reason: 5% improvement is not that large).

Based on the same idea, when considering the logistic regression model

$$\text{logit}(\pi) = \alpha + \beta_1 x$$

and intends to test $H_0 : \beta_1 = 0$, the sample size depends on the distribution of the values for the predictor, x. One needs to be able to *guess* the probability of success, $\bar{\pi}$, when $x = \bar{x}$.

The effect size is the odds ratio ϑ comparing $\bar{\pi}$ to the probability of success one standard deviation above the mean, $\bar{\bar{\pi}}$.
Let$\lambda = \log(\vartheta)$. an approximate sample size formula for a one sided test is then

$$n = [z_\alpha + z_\beta \exp(-\lambda^2/4)]^2(1 + 2\bar{\pi}\delta)/(\bar{\pi}\lambda^2)$$

with

$$\delta = [1 + (1 + \lambda^2)\exp(5\lambda^2/4)]/[1 + \exp(-\lambda^2/4)].$$

The usefulness of this approach and formula is limited, since one usually would not know the mean of the predictor x and $\bar{\pi}$.
Using a similar approach to multiple logistic regression would require even more information before one even collects data, and the outcome becomes even more guesswork.