A Note On Sample Sizes Needed To Detect Differences In Proportions[1]

Vince Stanford

December 7, 2016 (revised March 20, 2019)

Sample size for detecting a given difference in proportions

We discuss well known techniques for the determining the sample sizes needed to allow us to detect differences between two specified proportions.[1].

The mathematics of Sample Size Determination

Suppose the proportions found in the two samples are p_1 and $p_2$with a common sample size n. Suppose further that n is large enough for the Central Limit Theorem to The statistic (temporarily ignoring the continuity correction) for testing the significance of their difference is:

$$z = \frac{p_1 - p_2}{\sqrt{\frac{2\bar{p}\bar{q}}{n}}}$$

where

$$\bar{p} = \frac{1}{2}(p_1 + p_2)$$

and

$$\bar{q} = 1 - \bar{p}$$

Now fix the Type I error as α. Thus z will be significant if

$$|z| > Z_{\alpha/2}$$

where $Z_{\alpha/2}$ is the denotes the threshold such that $\alpha/2$ probability mass of the Standard Normal probability density function.

Now, if the difference between the underlying true proportions is actually $\Delta P = P_2 - P_1$ so we wish to have a probability of rejecting $H_0 : P_2 - P_1 = 0$ in favor of $H_1 : P_2 - P_1 = \Delta P$ of $1 - \beta$. Thus we must find a value of n such that when $\Delta P = P_2 - P_1$ is the true difference in proportions

$$Prob \left\{ \frac{|p_2 - p_1|}{\sqrt{\frac{2\bar{p}\bar{q}}{n}}} > Z_{\alpha/2} \right\} = 1 - \beta$$

Which is the sum of the two probabilities:

$$Prob \left\{ \frac{p_2 - p_1}{\sqrt{\frac{2\bar{p}\bar{q}}{n}}} > Z_{\alpha/2} \right\} + Prob \left\{ \frac{p_2 - p_1}{\sqrt{\frac{2\bar{p}\bar{q}}{n}}} < Z_{\alpha/2} \right\} = 1 - \beta$$

[1] *Determining Sample Sizes Needed to Detect a Difference Between Two Proportions,* Chapter 2 of *Statistical Methods for Rates and Proportions.* Joseph L. Fleiss, John Wiley & Sons, New York, 1973.

If we hypothesize that $P_2 > P_1$we can ignore the second term above since it will be very small, so we can find

$$1 - \beta = Prob\left\{ \frac{p_2 - p_1}{\sqrt{\frac{2\bar{p}\bar{q}}{n}}} > Z_{\alpha/2} \right\}$$

Further assuming large samples, the law of large numbers allows us to equate $P_1 \approx p_1$ and $P_2 \approx p_2$. Thus

$$E\left(p_2 - p_1\right) = P_2 - P_1$$

and

$$s.e.\left(p_2 - p_1\right) \approx \sqrt{\frac{\left(P_1Q_1 + P_2Q_2\right)}{n}}$$

where $Q_1 = 1 - P_1$ and $Q_2 = 1 - P_2$.

$$1 - \beta = Prob\left\{ p_2 - p_1 > Z_{\alpha/2}\sqrt{\frac{2\bar{p}\bar{q}}{n}} \right\}$$

and

$$1 - \beta = Prob\left\{ \frac{\left(p_2 - p_1\right) - \left(P_2 - P_1\right)}{\sqrt{\frac{\left(P_1Q_1 + P_2Q_2\right)}{n}}} > \frac{C_{\alpha/2}\sqrt{\frac{2\bar{p}\bar{q}}{n}} - \left(P_2 - P_1\right)}{\sqrt{\frac{\left(P_1Q_1 + P_2Q_2\right)}{n}}} \right\}$$

and Z tends toward normality as n increases we have

$$Z = \frac{\left(p_2 - p_1\right) - \left(P_2 - P_1\right)}{\sqrt{\frac{\left(P_1Q_1 + P_2Q_2\right)}{n}}}$$

Let $Z_{1-\beta}$ be the value such that

$$1 - \beta = Prob\left\{Z > Z_{1-\beta}\right\}$$

Combining the above equations, we have two corresponding elements

$$Z_{1-\beta} = \frac{Z_{\alpha/2}\sqrt{\frac{2\bar{p}\bar{q}}{n}} - \left(P_2 - P_1\right)}{\sqrt{\frac{\left(P_1Q_1 + P_2Q_2\right)}{n}}}$$

$$= \frac{Z_{\alpha/2}\sqrt{2\bar{p}\bar{q}} - \left(P_2 - P_1\right)\sqrt{n}}{\sqrt{P_1Q_1 + P_2Q_2}}$$

Note that for large samples we can substitute for $\sqrt{2\bar{p}\bar{q}}$

$$\bar{P} = \frac{P_1 + P_2}{2}$$

Algorithm 0.1 Estimated Sample Size Function from Fleiss. *R* implementation.

```
fleiss_function <- function(alpha, beta, p1, p2) {
  pbar <- (p1 + p2) / 2
  qbar <- (1 - pbar)
  q1 <- (1 - p1)
  q2 <- (1 - p2)
  c_alpha_over_2 <- qnorm(alpha / 2)
  c_1_minus_beta <- qnorm(1 - beta)

  n <- (c_alpha_over_2 * sqrt(2 * pbar * qbar) - c_1_minus_beta *
        sqrt(p1 * q1 + p2 * q2)) ^ 2 / (p2 - p1) ^ 2

  # Continuity correction
  (n / 4) * (1 + sqrt(1 + 8/(n * abs(p1 - p2)))) ^ 2
}

# Checking our work against table of results provided by Fleiss
fleiss_function(0.05, 0.05, 0.05, 0.1) # Expect 796
```

$$\bar{Q} = 1 - \bar{P}$$

and

$$n = \frac{\left(Z_{\alpha/2}\sqrt{2\bar{P}\bar{Q}} - Z_{1-\beta}\sqrt{P_1Q_1 + P_2Q_2}\right)^2}{\left(P_2 - P_1\right)^2}$$

We get our final sample size estimator by applying the continuity correction of Kramer and Greenhouse[2] as follows:

$$\hat{n} = \frac{n}{4}\left(1 + \sqrt{1 + \frac{8}{(n\,|P_2 - P_1|)}}\right)^2$$

See algorithm 0.1 for implementation in *R*.

[2]Determination of sample size and selection of cases. M. Kramer and S. Greenhouse. NAS/NRC publication 583, p. 356-371, *Psychopharmacology: Problems in Evaluation.* Washington D.C.