Sample Size Estimation

Precision and Power Analysis

Christopher Meaney

October 4, 2023

1 Background

In this research note, we review sample size estimation for precision and power analysis. We work through five examples discussed in Browner et al. [2022] :

1. Sample size for descriptive study estimating a CI for a continuous mean.
2. Sample size for descriptive study estimating a CI for a binomial proportion.
3. Sample size for a study testing whether a continuous mean is greater than some null value.
4. Sample size for a study testing whether a binomial proportion is greater than some null value.
5. Sample size for a study testing whether the bivariate Pearson correlation coefficient between two random variables is greater than some null value.

We introduce theoretical formula related to sample size estimation for precision analysis and power analysis [Chow et al., 2017]. We apply the theory to calculate sample size function addresssing each of the five examples discussed above.

2 Sample Size for Precision Analysis

Objective: Design a study such that (α-level) confidence statements about expected model parameters have desired precision (width).

We begin by stating the general formula needed to construct the lower and upper limits for the (Wald-style) confidence interval (LL, UL).

$$LL = \hat{\theta} - z_{\alpha/2} * \hat{se}(\hat{\theta})$$

$$UL = \hat{\theta} + z_{\alpha/2} * \hat{se}(\hat{\theta})$$

The total width (W) of the confidence interval is given as: W = UL - LL.

$$\begin{aligned} W &= UL - LL \\ &= (\hat{\theta} + z_{\alpha/2} * \hat{se}(\hat{\theta})) - (\hat{\theta} - z_{\alpha/2} * \hat{se}(\hat{\theta})) \\ &= 2 * z_{\alpha/2} * \hat{se}(\hat{\theta}) \end{aligned}$$

We assume an estimator for $\hat{se}(\hat{\theta}) = \frac{s}{\sqrt{n}}$.

$$\begin{aligned} W &= 2 * z_{\alpha/2} * \hat{se}(\hat{\theta}) \\ W &= 2 * z_{\alpha/2} * (\frac{s}{\sqrt{n}}) \\ \sqrt{n} * W &= 2 * z_{\alpha/2} * s \\ n &= \frac{4 * z_{\alpha/2}^2 * s^2}{W^2} \end{aligned}$$

From the final equation above, we note that sample size for precision analysis is governed by three input parameters: 1) $z_{\alpha/2}$, 2) s, and 3) W.

1. Confidence level ($z_{\alpha/2}$): As confidence level increases, the sample size increases.

2. Sample standard deviation (s): As sample standard deviation increases, the sample size increases.

3. Desired precision/width (W): As the confidence interval width decreases (precision increases), the sample size increases.

3 Sample Size for Power Analysis

Objective: To design a study such that it has a high probability of detecting a true effect should one exist. Or alternatively, designing a study to with a priori control of type-1 and type-2 error probabilities.

Example: Below we derive critical values, power, and sample size formula under the assumption of a 1-sample problem/design, and a 1-sided hypothesis test, assuming an "IID" (independent and identically distributed) data generating mechanism.

$$\begin{aligned} H_0 &: \mu = \mu_0 \\ H_1 &: \mu > \mu_0 \end{aligned}$$

If the null hypothesis is true, we want the probability of rejecting the null hypothesis to be α. We can write this in symbols as $\alpha = Pr(X \geq c | H_0)$.

Derive a general formula for the critical value of the hypothesis test:

$$\begin{aligned}
\alpha &= Pr(X \geq c | H_0) \\
&= 1 - Pr(X \leq c | H_0) \\
&= 1 - Pr\left(\frac{X - \mu_0}{\sigma_n} \leq \frac{c - \mu_0}{\sigma_n} \middle| H_0\right) \\
&= 1 - \Phi\left(\frac{c - \mu_0}{\sigma_n}\right) \\
1 - \alpha &= \Phi\left(\frac{c - \mu_0}{\sigma_n}\right) \\
z_{1-\alpha} &= \frac{c - \mu_0}{\sigma_n} \\
c &= \mu_0 + z_{1-\alpha}\,\sigma_n
\end{aligned}$$

From above we define a decision rule for whether we reject, versus fail to reject our null hypothesis as stated above:

$$\text{Fail to Reject } H_0 \text{ if: } X < \mu_0 + z_{1-\alpha}\,\sigma_n$$

$$\text{Reject } H_0 \text{ if: } X \geq \mu_0 + z_{1-\alpha}\,\sigma_n$$

Given a definition for a critical region (above), we next derive the power of the hypothesis test (under the assumption that the alternative hypothesis is true):

$$\begin{aligned}
\text{Power} &= Pr(X \geq \mu_0 + z_{1-\alpha}\,\sigma_n | H_1) \\
&= 1 - Pr(X \leq \mu_0 + z_{1-\alpha}\,\sigma_n | H_1) \\
&= 1 - Pr\left(\frac{X - \mu}{\sigma_n} \leq \frac{\mu_0 + z_{1-\alpha}\,\sigma_n - \mu}{\sigma_n} \middle| H_1\right) \\
&= 1 - Pr\left(\frac{X - \mu}{\sigma_n} \leq \frac{\mu_0 - \mu}{\sigma_n} + z_{1-\alpha} \middle| H_1\right) \\
&= 1 - \Phi\left(\frac{\mu_0 - \mu}{\sigma_n} + z_{1-\alpha}\right) \\
&= \Phi\left(\frac{\mu - \mu_0}{\sigma_n} - z_{1-\alpha}\right)
\end{aligned}$$

Given the power formula derived above, we can next derive a simple sample size formula, solving for n in the above expression.

$$
\begin{aligned}
1-\beta &= \Phi\left(\frac{\mu-\mu_0}{\sigma_n} - z_{1-\alpha}\right) \\
z_{1-\beta} &= \frac{\mu-\mu_0}{\sigma_n} - z_{1-\alpha} \\
\frac{1}{\sigma_n} &= \frac{z_{1-\beta}+z_{1-\alpha}}{\mu-\mu_0}
\end{aligned}
$$

The above expression is a general result. Assuming a particular parametric data generating model, and plugging in an estimator for σ_n allows for derivation of more concrete sample size estimators in particular application areas.

3.1 Normal Data

If the data are normally distributed, and we are testing a (1-sided) hypothesis, about a single mean, then we have the following results:

$$
\begin{aligned}
\bar{Y} &\sim N(\mu, \sigma^2/n). \\
\text{Power} &= \Phi\left(\frac{\mu-\mu_0}{\sigma/\sqrt{n}} - z_{1-\alpha}\right) \\
n &= \left(\sigma\frac{z_{1-\beta}+z_{1-\alpha}}{\mu-\mu_0}\right)^2
\end{aligned}
$$

3.2 Binomial Data

If the data are binomially distributed, and we are testing a (1-sided) hypothesis, about a single proportion, then we have the following results:

$$
\begin{aligned}
\hat{p} &\sim N\left(p, \frac{p(1-p)}{n}\right) \\
\text{Power} &= \Phi\left(\frac{p-p_0}{\sqrt{p(1-p)/n}} - z_{1-\alpha}\right) \\
n &= p(1-p)\left(\frac{z_{1-\beta}+z_{1-\alpha}}{p-p_0}\right)^2
\end{aligned}
$$

3.3 Bivariate Normal Data

If vector-valued data (X,Y) are distributed according to a bivariate normal distribution, and we are testing a (1-sided) hypothesis, about whether the population correlation coefficient is greater than some null value (often 0), we have the following results.

$$(X, Y) \sim \mathrm{BVN}(\mu, \Sigma)$$

$$\mu = (\mu_X, \mu_y)$$

$$\Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$$

$$\rho = \frac{\mathrm{COV}(X, Y)}{\sigma_X, \sigma_Y}$$

$$r = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{N}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{N}(Y_i - \bar{Y})^2}}$$

The Fisher transform: $\gamma = f(r)$, of the correlation coefficient (r) is given below. The Fisher transform is both a normalizing and variance stabilizing transformation. Following application of the Fisher transform, the distribution of the correlation coefficient is approximately normally distributed, with variance a function of the sample size (and independent of other model parameters) [Winterbottom, 1979].

$$\gamma = \frac{1}{2} * \log\left(\frac{1 - r}{1 + r}\right)$$

$$\hat{\gamma} \sim \text{Normal}\left(\gamma, \frac{1}{\sqrt(N - 3)}\right)$$

$$\text{Power} = \Phi\left(\frac{\rho - \rho_0}{1/\sqrt{N - 3}} - z_{1-\alpha}\right)$$

$$N = \left(\frac{z_{1-\beta} + z_{1-\alpha}}{\rho - \rho_0}\right)^2 + 3$$

References

Warren S Browner, Thomas B Newman, Steven R Cummings, and Deborah G Grady. *Designing clinical research.* Lippincott Williams & Wilkins, 2022.

Shein-Chung Chow, Jun Shao, Hansheng Wang, and Yuliya Lokhnygina. *Sample size calculations in clinical research.* CRC press, 2017.

Alan Winterbottom. A note on the derivation of fisher's transformation of the correlation coefficient. *The American Statistician*, 33(3):142–143, 1979.