Notes on Simulating Power

Prepared by Mauricio Cáceres

September 12, 2016

1 Parametric Power

We typically consider the model

$$Y_i = \alpha + \beta T_i + \gamma X_i + \varepsilon_i \tag{1}$$

for some treatment T_i at the individual level and the OLS estimator $\widehat{\beta}_{OLS}$ the difference in outcomes for the treatment and control groups. For $H_0 : \beta = 0$, consider the rejection probability function

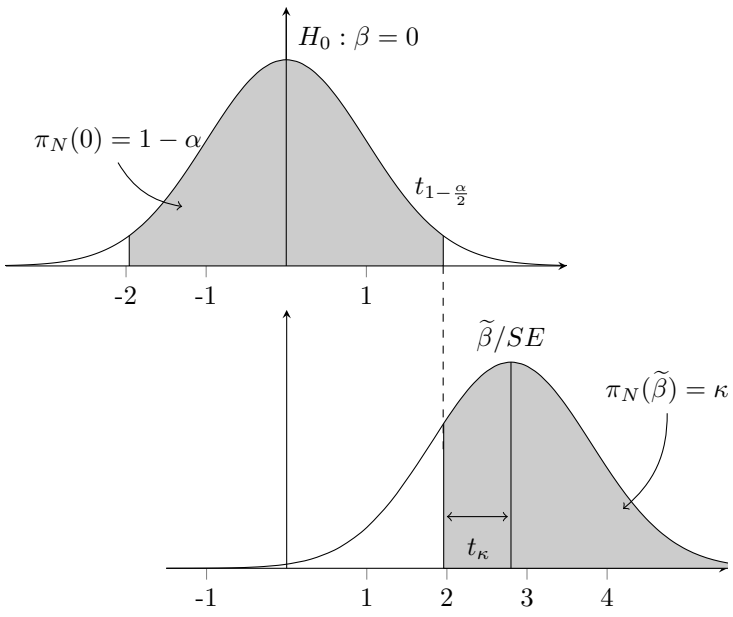
$$\pi_N(\beta) = P\{\text{reject } H_0 | \beta\} \tag{2}$$

For $\beta = 0$, this is α the probability of Type I error or *significance*: How likely are we to make a mistake? For $\beta = \widetilde{\beta} \neq 0$ this is *power*, the probability of rejecting the null: How likely are we to get it right? $\widehat{\beta}_{OLS}$ is $\sqrt{n}$-consistent, that is,

$$\sqrt{n}\left(\widehat{\beta}_{OLS} - \beta\right) \xrightarrow{D} N(0, V_{\widehat{\beta}})$$

where β is the true mean. Relying on large-sample asymptotics, we can visualize $\pi_N(\beta)$,

Figure 1.1: Power of a Test



A *parametric* approach to power estimates $\widetilde{\beta}$ given α, κ, N, SE and terms it the *minimum detectable effect*, MDE, or it estimates N given α, κ, SE, MDE.

2 Simulated Confidence Interval

However, it is possible to follow a *non-parametric* approach to estimating power. Note there are $C = \binom{N}{NP}$ ways to treat PN individuals. If we estimate $\widehat{\beta}_{OLS}$ for each $c = 1, \ldots, C$, then we would know the exact distribution of our estimator for the treatment effect under the null given the data. Thus we could compute an exact p-value and determine whether to reject the null.

Even for modestly-sized data, C will be intractably large. Hence we simulate K draws from the possible treatment-control arrangements, T_{ik} such that $\sum_{i=1}^{N} T_{ik} = PN$, and estimate

$$Y_i = \alpha + \beta_k T_{ik} + \gamma X_i + \varepsilon_i \tag{3}$$

Here $\widehat{\beta}_k$ will be distributed around 0 and a $1 - \alpha$ CI under the null is given by

$$\widehat{CI}_{1-\alpha} = \left(\widehat{F}^{-1}(\alpha/2), \widehat{F}^{-1}(1 - \alpha/2)\right) \tag{4}$$

with $\widehat{F}(\widehat{\beta}_k)$ the empirical cdf of $\widehat{\beta}_k$. This approach is appealing compared to a parametric approach because it naturally takes into account the correlation structure of the errors, whereas a parametric approach requires making an assumption about $V_{\widehat{\beta}}$. If we had historical data on our study population (we can think of historical data as data on a population where our treatment had no effect, or the counterfactual of what would happen were we to treat our population with no effect) then we can simulate the CI above, and say that we expect to be able to reject effects outside the confidence interval.

3 Simulated Power

Note, however, that this says nothing about power. In fact, power at either end of the confidence interval should be about 0.5 (if $\widehat{\beta}_{OLS}$ were to be symmetrically distributed). Typically we look for a power level of 0.8 or 0.9. We could assume that the true effect of T_i is $\widetilde{\beta}$, and estimate

$$\begin{aligned} \widetilde{Y}_{ik} &= Y_i + \widetilde{\beta} T_{ik} \\ \widetilde{Y}_{ik} &= \alpha + \beta_k T_{ik} + \gamma X_i + \varepsilon_i \end{aligned} \tag{5}$$

In this case, $\widehat{\beta}_k$ will be distributed around $\widetilde{\beta}$ but the shape of the distribution would not have changed. Thus we can estimate power as

$$\widehat{\kappa} = \frac{1}{K} \sum_{k} 1\left(\widehat{\beta}_k \notin \widehat{CI}_{1-\alpha}\right) \tag{6}$$

and we we can search for $\widetilde{\beta}$ such that $\widehat{\kappa} \approx \kappa$ for some desired power level κ (note $\widehat{\kappa} \xrightarrow{P} P\left(\beta_k \notin CI_{1-\alpha}\right) = \kappa$). That is, we look for a $\widetilde{\beta}$ that causes us to reject the null κ portion of the time. This is more complicated if Y_i is binary. The approach works to obtain a CI under the null but the subsequent search does not map trivially. One idea is to randomly swap successes to failures (or the converse) based on $\widetilde{\beta}$. Consider

$$\begin{aligned} \widetilde{Y}_{ik} &= Y_i(1 - T_{ik}) + (Y_i + S_{ik})T_{ik} = Y_i + S_{ik} T_{ik} \\ \widetilde{Y}_{ik} &= \alpha + \beta_k T_{ik} + \gamma X_i + \varepsilon_i \end{aligned} \tag{7}$$

where S_{ik} is constructed as follows

- Let

$$T_k = \sum_i T_{ik} \qquad S_k = \sum_i T_{ik} Y_i$$
$$\widetilde{\beta} \in \left[-\frac{S_k}{T_k}, \frac{T_k - S_k}{T_k} \right]$$
$$S_k^1 = \widetilde{\beta} T_k$$
$$S_k^2 = \begin{cases} \dfrac{S_k}{T_k} - S_{1k} & \widetilde{\beta} > 0 \\ \dfrac{T_k - S_k}{T_k} - S_{1k} & \widetilde{\beta} < 0 \end{cases}$$

- Construct the set ς_k with S_k^1 entries equal to $1(\widetilde{\beta} > 0) - 1(\widetilde{\beta} < 0)$ and S_k^2 entries equal to 0.
- For i such that $T_{ik} = 1$, draw s from ς_k *without* replacement and set $S_{ik} = s$ ($S_{ik} = 0$ otherwise).

Note that $T_k^{-1} \sum_i S_{ik} T_{ik} = \widetilde{\beta}$, hence

$$\begin{aligned} E\widehat{\beta}_k &= E\left[\widetilde{Y}_{ik} | T_{ik} = 1\right] - E\left[\widetilde{Y}_{ik} | T_{ik} = 0\right] \\ &= E\left[Y_i + S_{ik} | T_{ik} = 1\right] - E\left[Y_i | T_{ik} = 0\right] \\ &= EY_i + E\left[S_{ik} | T_{ik} = 1\right] - EY_i \\ &= \widetilde{\beta} \end{aligned}$$

So $\widehat{\beta}_k$ will be distributed around $\widetilde{\beta}$. Now we can outline a general simulation procedure:

1. Estimate a $1 - \alpha$ CI for $\widehat{\beta}_{OLS}$ under $H_0 : \beta = 0$ using Equation (3) and Equation (4).
2. Choose a starting MDE, $\widetilde{\beta}$, and estimate $\widehat{\beta}_k$ for $k = 1, \ldots, K$ using Equation (5) if Y_i is continuous or Equation (7) if Y_i is binary.
3. Estimate power using Equation (6).
4. If $\widehat{\kappa} < \kappa$ then increase $\widetilde{\beta}$; if $\widehat{\kappa} > \kappa$ then decrease $\widetilde{\beta}$.
5. Continue until $|\widehat{\kappa} - \kappa| < \epsilon$ for ϵ small.

4 Monte Carlo Simulations

Consider some data-generating process (DGP) for $w_i = (y_i, x_i, T_i)$,

$$x_i \sim F_x \qquad T_i \sim \text{Bernoulli}(P) \qquad T_i \perp\!\!\!\perp x_i \qquad y_i | x_i, T_i \sim F_y$$

where

$$\beta = E(y_i | T_i = 1, x_i) - E(y_i | T_i = 0, x_i)$$

The aim is to compute the power of testing $\widehat{\beta}_{OLS}$ against our simulated confidence interval. For $m = 1, \ldots, M$:

1. Generate two draws from the DGP, w_{1m} with $P = 0$ and w_{2m} $P \in (0, 1)$.

2. Compute $\widehat{\kappa}_{1m}, \widehat{\kappa}_{2m}$ from our power simulation procedure and $\widehat{\beta}_{1m}, \widehat{\beta}_{2m}$ from OLS.

3. Construct a $1-\alpha$ confidence interval for β, $\widehat{CI}^{M}_{1-\alpha}$, using Equation (4) and compute

$$\overline{\kappa}^{M} = \frac{1}{M} \sum_{m} 1\left(\widehat{\beta}_{2m} \notin \widehat{CI}^{M}_{1-\alpha}\right)$$

4. It should be the case that $\widehat{\kappa}_{1m}$ are distributed around α and $\widehat{\kappa}_{2m}$ are distributed around $\overline{\kappa}^{M}$.