Experiments

April 14, 2010

Power Analysis

- What sample size do I need in order to detect a certain departure from a null hypothesis?
- Power = 1 - Pr(Type II error)
- Inherently hypothetical.

The Model

- There are $n_1 + n_2$ subjects, with n_1 assigned to treatment, the rest to controls.
- We assume an infinite population model, from which we are sampling randomly:
    - Model the treatment data as observed values of n_1 IID random variables $Y_1, \ldots, Y_{n1}$.
    - Model the control data as observed values of n_2 IID random variables $X_1, \ldots, X_{n2}$.
- Assume different population means μ_1 and μ_2 with a common variance σ^2.

The Hypothesis Test

- The null hypothesis is $\mu_1 = \mu_2$
- Test: two-sided 5% test of the null against the alternative $\mu_1 \neq \mu_2$
- Given $\mu_1 = \mu_2 + k\sigma$ where $k > 0$ and the effect $k\sigma$ of interest.
- Where does σ come from?
  - Previous studies?
  - Theory?
- Assume n_1 and n_2 are large enough that the central limit theorem kicks in, and $\hat{\sigma} \doteq \sigma$, where $\hat{\sigma}^2$ is the usual pooled estimator of variance.
- Let $f = \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$.
- The test statistic is $z = (\bar{Y} - \bar{X})/s$, where $s = \hat{\sigma} f \doteq \sigma f$

Power

- Under the alternative hypothesis, z is asymptotically distributed as $N(k^*, 1)$, where $k^* = k/f$.
- If Z is $N(0, 1)$, asympotic power is then

$$P(Z > 1.96 - k^*) + P(Z < -1.96 - k^*)$$

  , where 1.96 is for the standard two sided 95% confidence intervals.
- The typical benchmark used is power of 80% which corresponds to a $k^* = 2.85$.

An example: Bhavani (2009)

- Context: Local elections in Mumbai, where some consituencies were randomly assigned be part of a political quota system for women. Seats that are reserved change from election to election.
- The question: After reservations are withdrawn, what is the effect of the reservation on the likelihood of electing a woman politician?
- There are 37 constituencies in treatment and 81 in control.
- What is the power of this experiment?

Block What You Can; Randomize What You Can't

- When designing experiments it can—in most cases—be quite advantageous prior to randomization, to group like with like in strata called "blocks".
- A block should contain units whose potential outcomes are as similar as possibe. In other words, you want to minimize the variation that's not attributable to treatment.
- Randomization occurs within these strata and as a consequence, the grouped structure of the randomization must be taken into consideration in the analysis.
- Generally, one wants to decrease within-block heterogeneity (for efficiency) and increase across-block heterogeneity (for generalizability).
- In most situations, blocking will increase power. The better the covariates you use in the blocking stage, the greater the power.

ATE and Variance with Matched Pairs

Suppose that there exists $2n$ units and n matched-pairs are formed based on the observed treatment characteristics. An indicator variable, Z_j, is randomized by the experimentor with equal probability $Pr(Z_j) = \frac{1}{2}$ and determines which unit receives the treatment within the jth matched pair, where $j = 1, 2, \ldots, n$. T_{ij} is the treatment indicator variable.

- The estimand: $\tau_m = \frac{1}{2n} \sum_{j=1}^{n} \sum_{i=1}^{2} (Y_{ij}(1) - Y_{ij}(0))$
- The estimator is: $\hat{\tau}_m = \frac{1}{n} \sum_{j=1}^{n} \sum_{i=1}^{2} (T_{ij} Y_{ij} - (1 - T_{ij}) Y_{ij})$

Variance Gains

- The variance estimator is $\sigma_m = \frac{1}{n(n-1)} \sum_{j=1}^{n} \{(T_{ij} Y_{ij} - (1 - T_{ij}) Y_{ij}) - \hat{\tau}_m\}^2$
- If you analyze the experiment as if it were completely randomized, your variance will be wrong!
- The difference between the variance of the ATE estimator under the completely randomized design ($\hat{\tau}_c$) and the matched pair estimator is:

$$Var(\hat{\tau}_c) - Var(\hat{\tau}_m) = \frac{2}{n} cov(Y_{ij}(1), Y_{i'j}(0))$$

More on blocking

- Typically, the best covariate to block on is the previous outcome.
- If one wants to block on many covariates, just adapt a matching algorithm and try to acheive the best balance possible.
- Often one wants to block on geography, but be careful about SUTVA violations.
- What if the number of units varies from block to block? Then weight each block by the number of units before summing for treatment effect and variance calculations.