Randomized Experiments - Design

Costas Meghir

March 1, 2018

Designing Experiments

- Define the relevant population and the intervention that is to be evaluated
- Decide on the hypotheses to be tested and define clearly the outcome variables you are interested in.
- The next step relates to deciding on the sample size. The sample size will be determined by the power to detect effects of particular size.
- The estimator of the ATE ($\hat{b}$) in a simple randomized experiment is

$$\hat{b} = \bar{Y}_1 - \bar{Y}_0$$

where $\bar{Y}_k$ denotes the mean outcome in the treatment ($k = 1$) or control ($k = 0$) sample. Because we have randomized we do not need to worry about correlation of outcomes be treatment and control units.

Designing Experiments - Setting the sample size

- When randomization takes plce a tthe individual level the standard error of $\hat{b}$ is

$$se(\hat{b}) = \sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_0^2}{N_0}}$$

- where σ_k^2 is the variance in the k_{th} (1,0) group. The variance may differ because of heterogeneous treatment effects. N_1 and N_0 are the respective sample sizes.

Designing Experiments - Setting the sample size

- Now suppose we want to be able to detect an effect of at least a with power $p\%$. This means the if the true effect is a we want to choose our sample size so that we reject the null hypothesis that the effect is zero with probability $p\%$ when the significance level chosen is 5%
- For a two tailed test this means that we wish to choose the sample size so that

$$p(-t_a < t_b < t_a) = 1 - p/100$$

  when the impact is a, where t_b is the standard t-statistic $\hat{b}/se(\hat{b})$ and t_a is the critical value.
- We choose t_α (based on α) and p. We need to know in advance σ_1^2 and σ_0^2. We then choose N_1 and N_2 to be consistent with these choices.

Designing Experiments - Example for setting the sample size

- Intervention: Offer a school subsidy to children between 12 and 16 (conditional on attendance)
- The outcome variable is school attendance of children both in the eligible age group and younger: The effect is likely to be positive on the subject children but could be anything on younger children in the family.
- Suppose the proportion attending pre-intervention is 0.4. This implies that $\frac{\sigma_0^2}{N_0} = \frac{pr \times (1-pr)}{N_0} = \frac{0.24}{N_0}$.

Designing Experiments - Example for setting the sample size

- Policy makers have decided that an effect of 0.1 ($a = 0.1$ or 10%) would make the program worthwhile. If this occurred $\frac{\sigma_1^2}{N_1} = \frac{0.5^2}{N_1}$. This is a safe assumption because the variance of a proportion is maximized at $pr = 0.5$.We decide to design the sampling so that $N_0 = N_1 = N$. This maximizes the variance of the right hand side variable (variance of treatment = $p_T(1 - p_T) = 0.5^2$).
- We also decide that any hypothesis will be tested using 5% level of significance. Finally, we wish to reject the null with probability at least $p = 0.8$ if the true effect is larger than 0.1. To achieve this under the stated conditions we need to find the suitable sample size for each group N.

Designing Experiments - Setting the sample size

- If the true effect is $a = 0.1$ we have that asymptotically $\hat{b} \overset{a}{\sim} N\left(a, p\lim\left(se(\hat{b})\right)^2\right)$. We want to find N so that the following is true

$$p(-1.96 < \frac{\hat{b}}{se(\hat{b})} < 1.96) = 0.20$$

- In other words we want the t-statistic to be such that in 80% of the times we reject this false hypothesis.

Designing Experiments - Setting the sample size

In our example, under the alternative this is equivalent to

$$p(-1.96 - \frac{0.1}{\sqrt{\frac{\sigma_1^2}{N_1}+\frac{\sigma_0^2}{N_0}}} < \frac{\hat{b}-0.1}{se(\hat{b})} < 1.96 - \frac{0.1}{\sqrt{\frac{\sigma_1^2}{N_1}+\frac{\sigma_0^2}{N_0}}})$$
$$= p(z < 1.96 - 0.143\sqrt{N}) - p(z < -1.96 - 0.143\sqrt{N}) = 0.2$$

where $z \tilde{} N(0,1)$

Designing Experiments - Setting the sample size

- In practice $p(z < -1.96 - 0.143\sqrt{N}) \approx 0$.
- $p(z < 1.96 - 0.143\sqrt{N}) = 0.2$ implies $1.96 - 0.143\sqrt{N} = -0.84$ (from the Normal distribution tables)
- Thus, in this case we need to have that $1.96 - 0.143\sqrt{N} = -0.84$ which implies a sample size of approximately 384 observations for the treatment and 384 for the control. Note that in this case $p(z < -1.96 - 0.143\sqrt{N}) = p(z < -4.76) = 9.680e-07$. No adjustment need be made for the left tail.

Spatial Correlation and the estimator covariance matrix

- Then it is easy to show that the above general formula for the variance of the single β^{OLS} becomes

$$var(\beta^{OLS}) = \frac{\sigma^2}{(N_c M) Var(X)} (1 + (M-1)\rho_x \rho_u)$$

- Note that $v^{OLS} = \frac{\sigma^2}{(N_c M) Var(X)}$ is the variance of the OLS estimator if we assume no spatial correlation, i.e. the standard case.

Spatial Correlation and the estimator covariance matrix

- In a pilot study X may be the treatment. If the treatment is administered to a whole cluster (e.g. village) then $\rho_x = 1$. Moreover between the units of randomization the spatial correlation of X will be zero.
- In this case the formula becomes

$$var(\beta^{OLS}) = \frac{\sigma^2}{(N_c M)P(1-P)}(1+(M-1)\rho_u)$$

  where P is the probability of allocation to treatment. MAx power when $P = 0.5$.
- The sample size is $N_c M$. First note that for any *given* total sample size the variance *increases* with the size of the cluster. Thus it is much better to design a clustered sample so that there are many clusters with few members in them rather than the other way round (for a given sample size of course)

Power calculation with spatial correlation

- We need to know/make assumptions on the intra cluster spatial correlation ρ
- Fix the probability of allocation to treatment (in the sample)
- Define the minimum detectable effect a.
- We need to fix either M (individuals in cluster) *or* N_c number of clusters
- Power not very sensitive to M when spatial correlation is reasonably large. So fix M and find N_c by solving

$$p(-1.96-\frac{a}{\sigma}\sqrt{\frac{(N_cM)P(1-P)}{(1+(M-1)\rho_u)}}<z<1.96-\frac{a}{\sigma}\sqrt{\frac{(N_cM)P(1-P)}{(1+(M-1)\rho_u)}})=0.20$$

- In practice try alternative values of ρ and M to check sensitivity of the number of clusters.

Multiple Hypothesis testing - The Problem

- References
  1. Romano, J, M Wolf (2005) *Stepwise multiple testing as formalized data snooping*, Econometrica 73
  2. Joseph P Romano, Azeem M Shaikh, Michael Wolf *Formalized data snooping based on generalized error rates*, Econometric Theory, 24, 2008
- Suppose we wish to test S hypotheses of the form $H_s : \theta_s = \theta_{0s}$.
- When we test a hypothesis there is a probability we reject it even if it is true (Type I error)
- This probability is regulated by the choice of critical value depending on the distribution of the test statistic under the null
- For example if the test statistic is standard normal choosing critical values of ± 1.96 implies a type I error of 5%
- This does mean that in testing 100 true hypotheses we will on average reject 5 if they are independent.

The Problem

- However, we would really like to control for the Type I error for the set of hypotheses we are testing - the Familywize error rate FWE.
- So if the FWE is set at 5% this would then imply that the probability of rejecting at least one hypothesis (which means reject the joint set of hypotheses) is at most 5%.
- A simple adjustment is provided by the Bonferroni procedure: with k hypotheses being tested compare each test statistic to a critical value corresponding to a significance level of (α/k).
- So for example for 10 hypotheses and an overall size level of 5% we would test each parameter at the 0.5% level.

The Problem

- However, the Bonferroni is too conservative in the sense that the critical value for rejecting a hypothesis is for most cases far too large - it requires far too high t-statistics to conclude that something is significant.
- From Romano, Shaikh and Wolf (2008) we have that the Bonferroni procedure controls asymptotically for the Familywize error rate if the distribution of each p-value corresponding to a hypothesis being tested is stochastically dominated by a uniform distribution, i.e.

$$H_s \ \ true \implies lim_{N\to\infty} P(\hat{p}_{T,s} <= u) <= u$$

- The Bonferroni procedure ends up bing too conservative because it does not take into account the actual dependence structure between the hypotheses. Bonferroni assumes the worst case dependence structure.
- With modern simulation techniques we can do much better, exploiting the actual dependence structure.

Solutions to multiple hypotheses testing problem

- Start by defining the Family-wise error rate (FWE): The probability of rejecting one or more hypotheses falsely.
- The aim is to use a procedure that controls for the FWE allowing for the actual dependence structure among the test statistics
- We already know that we can test the hypotheses jointly using an F-statistic.
- This will test the joint hypothesis and control the Type I error at the desired level (say 5%)
- However, it will not be informative as to which hypotheses are responsible for rejection.
- In a more general sense we would like to derive adjusted p-values for each hypothesis separately, while controlling for the FWE
- The FWE is the probability that any of the set of true hypotheses is rejected
- We require that the FWE holds even when some hypotheses are false (strong control)

The Romano and Wolf Stepdown procedure

- Define the $k - FWE$ as the probability of rejecting at least k hypotheses falsely from among those that are true
- Usually we set k to 1.
- However, if there are that many hypotheses to test we may loose power. By increasing k we can regain some power at the expense of Type I error probability.
- A multiple hypothesis method controls the k-FWE (asymptotically) at the level α if $sup_{N->\infty} k - FWE \leq \alpha$, where N is the sample size.

The Romano and Wolf Stepdown procedure

- Take the standard case where $k = 1$, i.e. we tolerate no false rejections.
- Suppose the significance level is α. Take the case of a one-sided test.
- Suppose we can construct a critical value such that the probability of the largest test statistic being below c_1 is $1 - \alpha$ under the null
- Then in step one reject all hypotheses where the test statistic is above c_1.
- If we reject no hypotheses we stop.
- Suppose we reject R_1 hypotheses.
- We can now be confident that some of the hypotheses are false.
- Hence we can start again this procedure since false hypotheses do not affect the Type I error.
- We delete the rejected hypotheses and start again.
- for a k-FWE control we define the critical value in terms of the quantile of the the k-th largest statistic.

A bootstrap method for the stepdown procedure

1. Start by constructing a test statistic for each hypothesis. This could the t-statistics
2. Suppose we are carrying out the usual two-sided tests. Then take the absolute value of these tests
3. Sort them from largest to smallest.

A bootstrap method for the stepdown procedure

- Draw M (many) bootstrap samples (with replacement).
- For each sample:
  1. construct the same test statistic
  2. Center it by subtracting its value obtained from the orignal sample (to ensure the null is true in the bootstrap sample) and take the absolute value
  3. Find the Maximum of the absolute values of these centered statistics.
- Once we have simulated the distribution of the max absolute value find the $1-\alpha$ percentile
- Reject all the hypotheses for which the original test statistic is above this critical value
- If you reject at least one, delete these test statistics from the original set and from the bootstrap sample and repeat
- To derive p-values repeat this exercise with different values of α (0.005, 0.01, 0.015, 0.02, 0.03,...)

A view from the non-economics literature

- "In the tables reporting the p-values, the authors do something I have never seen before in a published paper. They report the uncorrected p-values, indicating those that are significant (prior to correction) in boldface, and then put an asterisk next to those that are significant after their (incomplete) correction." from *Does researching casual marijuana use cause brain abnormalities?* by Lior Pachter https://liorpachter.wordpress.com/2014/04/17/does-researching-casual-marijuana-use-cause-brain-abnormalities/