Choosing sample size in randomized experiments

Aleksey Tetenov (University of Bristol)

Cemmap masterclass
Statistical decision theory for treatment choice and prediction
May 30-31, 2017

Prevailing convention

Convention for determining the sample size of a randomized trial comparing a new treatment with a control:

- Assume that the outcomes will be used to perform a test of a specified null hypothesis (new treatment is not better) at a conventional test level (5%)
- Select a specific positive effect size MCID ("Minimum detectable effect", "Minimum clinically important difference")
- Compute sample size sufficient to limit Type II error probability by 10% or 20% at the effect size MCID, i.e. to reject the null with at least 80% or 90% probability.

Shortcomings of the prevailing convention

- Inattention to magnitudes of losses: A given error probability should be less acceptable when the magnitude of the effect is larger. 10% error probability at effect size MCID tells us little about expected welfare losses at other effect sizes.
- Use of conventional error probabilities:
  Why limit Type I error by 1% or 5%? (Which usually implies Type II error of 99% or 95% for infinitesimal positive effects)
  Why limit Type II error by 10% or 20% at MCID?
  Why are they different?
- Limitation to settings with two treatments:
  Even with multiple testing adjustments, the hypothesis testing framework is still about probabilities of Type I/Type II errors. They do not capture the welfare losses in the problem of choosing among K treatments.

Bayesian critique

Bayesian statisticians have long criticized the use of concepts in hypothesis testing to design trials and make treatment decisions.

Bayesian statistical decision theorists argue that the purpose of trials is to improve medical decision making and conclude that trials should be designed to maximize subjective expected utility in settings of clinical interest.

The sample sizes selected may differ from those motivated by testing theory.

The Bayesian perspective is compelling when one can place a credible prior distribution on treatment response, but agreeing on priors is difficult.

ε-optimality

Source: Manski and Tetenov (2016), "Sufficient trial size to inform clinical practice," *PNAS* 113(38), 10518-10523.

An ideal objective is to collect data that enable implementation of an optimal rule - one whose expected welfare equals the welfare of the best treatment in every state of nature.

Optimality is not achievable in general, but ε-optimal rules do exist when trials have large enough sample size.

An ε-optimal rule has expected welfare within ε of the welfare of the best treatment in every state. Equivalently, it has maximum regret no larger than ε.

Implementation of the idea requires specification of a value for ε.

The necessity to choose an effect size of interest when designing trials already arises in conventional practice, where the trial planner must specify the effect size at which power is calculated.

A possibility is to let ε equal the minimum clinically important difference (MCID) in the average treatment effect comparing alternative treatments.

There is suspicion that in practice MCID is often chosen ex post to formally justify sample size driven by other sample size constraints.

The setup

A planner must assign one of K treatments to each member of a treatment population, denoted J.

Denote the set of treatments by T.

Each individual $j \in J$ has a response function $u_j(\cdot) : T \to \mathbb{R}$ mapping treatments $t \in T$ into welfare outcomes $u_j(t)$.

The probability distribution $P[u(\cdot)]$ of the random function $u(\cdot) : T \to \mathbb{R}$ describes treatment response across the population.

We will later consider individual observable covariates $x_j \in X$, where X is finite.

A statistical treatment rule (STR) δ maps sample data ψ into a treatment allocation.

Q is the sampling distribution generating the data
Ψ is the sample space.

Let Δ denote the space of functions that map $T \times \Psi$ into the unit interval and satisfy $\sum_{t \in T} \delta(t, \psi) = 1$, $\forall \psi \in \Psi$.

Each δ is an STR. $\delta(t, \psi)$ is the fraction of individuals assigned to treatment t when the data are ψ.

Denote the mean outcome of treatment t by $\mu_t \equiv E[u(t)]$.

The planner wants to maximize additive population welfare

$$U(\delta, P, \psi) \equiv \sum_{t \in T} \delta(t, \psi) \cdot \mu_t$$

but P is unknown.

Specify space S indexing possible states of the world. The treatment response distribution P_s and the sampling distribution Q_s depend on $s \in S$.

$\{(P_s, Q_s), s \in S\}$ - the set of feasible (P, Q) pairs.

Denote the mean response to treatment t in state s by μ_{st}.

The expected welfare (over repeated samples) yielded by rule δ in state s is

$$W(\delta, P_s, Q_s) \equiv \int_{\Psi} \left(\sum_{t \in T} \delta(t, \psi) \cdot \mu_{st} \right) dQ_s(\psi) = \sum_{t \in T} E_s[\delta(t, \psi)] \cdot \mu_{st}$$

The maximum welfare achievable is state of the world s is

$$U^*(P_s) \equiv \max_{t \in T} \mu_{st}$$

We call δ ε-optimal if for all $s \in S$

$$W(\delta, P_s, Q_s) \geq U^*(P_s) - \varepsilon,$$

i.e., if its maximum regret is no larger than ε:

$$\max_{s \in S} \left[U^*(P_s) - W(\delta, P_s, Q_s) \right].$$

We can consider two questions:

1. If a particular treatment rule (a hypothesis test rule or an Empirical Success (ES) rule) will be implemented, what sample size is needed to achieve ε-optimality?

2. If any treatment rule could be implemented, what sample size is sufficient to enable ε-optimal treatment assignment?

- We can obtain sufficient sample size in (1) by evaluating maximum regret of any candidate treatment rule (e.g., ES) if we do not know the exact minimax-regret rule.
- Rules that require fractional assignment (including the exact minimax-regret rule) may not be implementable, then we should consider implementable rules.
- Even if we cannot evaluate maximum regret exactly, an upper bound on maximum regret will give us sufficient sample size.

We use Empirical Success (ES) treatment rules to bound minimax regret.

Let $m_t(\psi) \equiv (n_t)^{-1} \sum_{j \in N(t)} u_j$ be the average outcome among n_t individuals assigned to treatment t in the sample.

An ES rule assigns all persons to treatment(s) that maximize $m_t(\psi)$ over T (treatments with the largest sample mean outcome).

ES rules are easily implementable and practical.
They are exactly or approximately minimax-regret in some settings with two treatments (Stoye 2009, 2012).
Upper bounds on regret of ES rules are analytically tractable.

Binary outcomes, two treatments, balanced design

With two treatments $T = \{a, b\}$, regret equals

$$U^*(P_s) - W(\delta, P_s, Q_s) = \max_{t \in T} \mu_{st} - \sum_{t \in T} E_s[\delta(t, \psi)] \cdot \mu_{st}$$

$$= \max(\mu_{sa}, \mu_{sb}) - E_s[\delta(a, \psi)] \cdot \mu_{sa} - E_s[\delta(b, \psi)] \cdot \mu_{sb}$$

If b is the new treatment and δ is a hypothesis test rule, then

$$= \underbrace{E_s[\delta(b, \psi)]}_{\text{P(Type I error)}} \cdot \underbrace{(\mu_{sa} - \mu_{sb})}_{\text{effect size}} \text{ if } \mu_{sa} \geq \mu_{sb},$$

$$= \underbrace{E_s[\delta(a, \psi)]}_{\text{P(Type II error)}} \cdot \underbrace{(\mu_{sb} - \mu_{sa})}_{\text{effect size}} \text{ if } \mu_{sb} \geq \mu_{sa}.$$

We compute maximum regret of candidate treatment rules in the case of binary outcomes $u_j(t) \in \{0, 1\}$, two treatments, and equal sample size for each treatment.

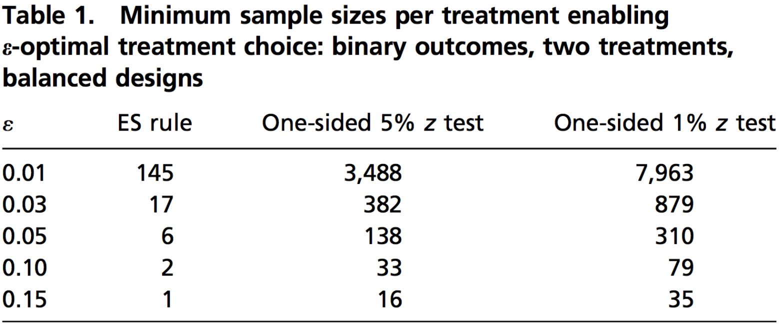
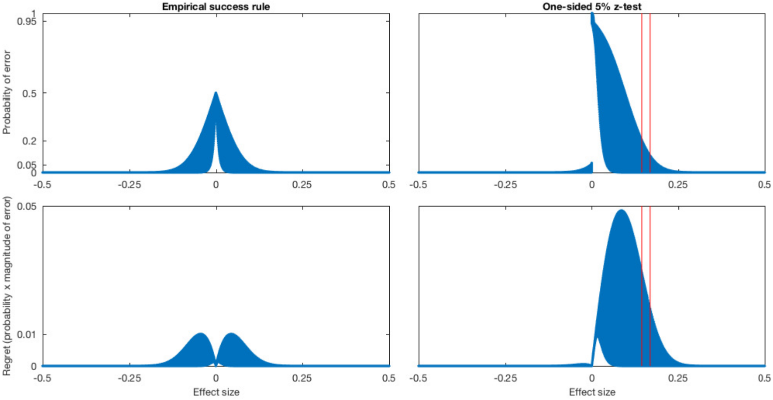


Table 1. Minimum sample sizes per treatment enabling ε-optimal treatment choice: binary outcomes, two treatments, balanced designs

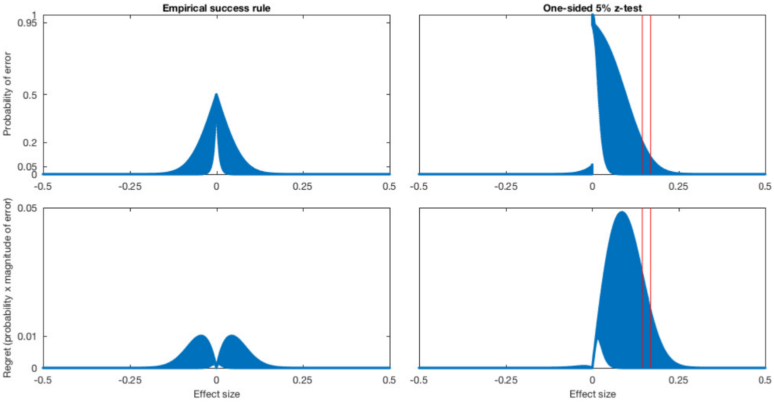
ε	ES rule	One-sided 5% *z* test	One-sided 1% *z* test
0.01	145	3,488	7,963
0.03	17	382	879
0.05	6	138	310
0.10	2	33	79
0.15	1	16	35

If hypothesis test rules are implemented, the minimum sample size required for ε-optimality is substantially larger.

For a given sample size, the maximum regret of a 5% one-sided hypothesis test rule is approx. 5 times larger than the maximum regret of an ES rule, which necessitates approx. 25 times larger sample for ε-optimality.



Red lines indicate effect sizes with P(Type II error) $=$ 10%/20%
If sample size is derived from a conventional power calculation, that's the MCID effect size.
Maximum regret $>$ MCID $\times$ P(Type II error at MCID)



Bounded outcomes, K treatments

We derive new upper bounds on the maximum regret of ES rules for bounded outcomes $u_j \in [u_l, u_h]$ with range $M \equiv u_h - u_l$ for any stratified sample sizes $(n_1, \ldots, n_K)$.

Balanced designs $n_1 = \cdots = n_K = n$ yield the lowest bounds:

Proposition 1:

$$(2e)^{-1/2} \cdot M \cdot (K-1) \cdot n^{-1/2}$$

Proposition 2:

$$M \cdot (\ln K)^{1/2} \cdot n^{-1/2}$$

(and a sharper bound that has to be evaluated numerically)

The bound in Proposition 2 is lower for $K \geq 4$

The bounds on maximum regret of ES rules imply simple bounds on sufficient sample size that guarantee ε-optimality:

Corollary to Proposition 1: (for $K = 2, 3$)

$$n \geq (2e)^{-1} \cdot (K-1)^2 \cdot \left(\frac{M}{\varepsilon}\right)^2$$

Corollary to Proposition 2: (for $K \geq 4$)

$$n \geq \ln K \left(\frac{M}{\varepsilon}\right)^2$$

These are only simple sufficient conditions for ε-optimality.

The best approach would be to bound maximum regret computationally, which seems challenging in the space of all possible bounded distributions of $u(t)$.

ε-optimality with observable covariates

Suppose that persons have observable covariates taking values in a finite set X and that the planner can execute a trial with (treatment, covariate)-specific sample sizes $[n_{t\xi}, (t, \xi) \in T \times X]$.

There are at least two reasonable ways that a planner may wish to evaluate ε-optimality in this setting.

One may want to achieve ε-optimality within each covariate group.

This interpretation requires no new analysis. The planner should simply define each covariate group to be a separate population of interest.

The design that achieves group-specific ε-optimality with minimum total sample size equalizes sample sizes across groups.

Alternatively, one may want to achieve ε-optimality within the overall population, without requiring that it be achieved within each covariate group.

The design that achieves ε-optimality with minimum total sample size does not equalize sample sizes across groups. Neither does it set the sample sizes proportional to group sizes.

With a balanced design assigning n_ξ individuals from covariate group ξ to each treatment, the maximum regret of an ES rule is bounded above by

$$(2e)^{-1/2} \cdot M \cdot (K-1) \sum_{\xi \in X} P(x = \xi) n_\xi^{-1/2},$$

$$M \cdot (\ln K)^{1/2} \sum_{\xi \in X} P(x = \xi) n_\xi^{-1/2}.$$

Given a predetermined maximum total sample size N, minimizing these bounds is achieved by choosing $(n_\xi, \xi \in X)$ to minimize

$$\sum_{\xi \in X} P(x = \xi) n_\xi^{-1/2}$$

If one treats $(n_\xi, \xi \in X)$ as continuous variables rather than as integer sample sizes, then the relative sample sizes for any pair (ξ, ξ') of covariate values should have the ratio

$$\frac{n_\xi}{n_{\xi'}} = \left[\frac{P(x = \xi)}{P(x = \xi')} \right]^{2/3}$$

Covariate-specific sample size increases with the prevalence of the covariate group in the population, but smaller groups are "oversampled" relative to their size.

Conclusion

Choosing sample sizes to enable ε-optimal treatment rules would align trial design with the objective of informing treatment choice better than the conventional practice of choosing sample size to achieve specified statistical power in hypothesis testing.

There are numerous directions for further research.

1. We considered trials drawing fixed number of subjects for each covariate group and treatment.

An alternative class of designs specifies a probability distribution for drawing subjects and assigning them to treatments. Our results could be used as an "inner loop" for evaluating probabilistic designs.

2. We let the state space contain all distributions of treatment response.

This assumption yields generally applicable findings.

However, it is unduly conservative when some credible knowledge of treatment response is available.

Given credible assumptions, it may be valuable to impose them. One could restrict feasible distributions $P[u(t)|x = \xi]$ or impose cross-covariate restrictions.

As the state space shrinks, the minimum sample needed to achieve ε-optimality logically cannot increase and may decrease.