Power and sample size calculations

Patrick Breheny

September 24

Introduction

- Last time we discussed testing whether two groups differ with respect to survival/hazard
- One reason such tests are useful is that they provide an objective criteria (statistical significance) around which to plan out a study: How many subjects do we need? How long will the study take to complete? This is our topic for today
- FYI: Our book doesn't really address this issue; today's lecture is largely derived from George and Desu (1974)'s classic paper on the subject

Exponential approximation

- The main idea behind George & Desu's approach is to assume constant hazards (i.e., exponential distributions) for the sake of simplicity
- Further work by other authors has indicated that the power/sample size one obtains from assuming constant hazards is fairly close to the empirical power of the log-rank test, provided that the ratio between the two hazard functions is constant
- Typically in a power analysis, we are simply trying to find the approximate number of subjects required by the study, and many approximations/guesses are involved, so using formulas based on the exponential distribution is often good enough

Special case: No censoring

- Let us begin with the special case of no censoring
- If $T_i \overset{\perp\!\!\!\perp}{\sim} \text{Exp}(\lambda)$ for $i = 1, \ldots, d$,

$$\begin{aligned} L(\lambda) &= \prod_i \lambda \exp(-\lambda t_i) \\ &= \lambda^d \exp(-\lambda V), \end{aligned}$$

  where $V = \sum_i t_i$
- Note that
  - V is a sufficient statistic
  - $V \sim \Gamma(d, \lambda)$

Type 2 censoring

- Now let's consider what happens in the case of type II censoring: in particular, that we have an initial sample size n and follow d subjects to failure
- In this case,

$$\begin{aligned} T_{(1)} &\sim \text{Exp}(n\lambda) \\ T_{(2)} - T_{(1)} &\sim \text{Exp}((n-1)\lambda) \\ &\dots \\ T_{(j)} - T_{(j-1)} &\sim \text{Exp}((n-j+1)\lambda) \end{aligned}$$

for $j = 1, \ldots, d$, with $T_{(0)} = 0$

Normalized spacings

- Alternatively, let $U_j = (n - j + 1)(T_{(j)} - T_{(j-1)})$
- Now $U_j \overset{\perp\!\!\!\perp}{\sim} \text{Exp}(\lambda)$, and

$$\begin{aligned} L(\lambda) &= \prod_j \lambda \exp(-\lambda u_j) \\ &= \lambda^d \exp(-\lambda V), \end{aligned}$$

  where $V = \sum u_j$
- Note that, once again, V is a sufficient statistic and follows a $\Gamma(d, \lambda)$ distribution

Remarks

- The exponential distribution, therefore, has the somewhat remarkable property that we arrive at the exact same inference if we follow d subjects until all have failed or if we follow some larger number n until d have failed
- Thus, we can carry out our calculations ignoring censoring, provided that we think of the sample size we obtain as the number of *events* that must be observed in order to achieve the desired power
- This is incredibly convenient for sample size planning, as it allows one to completely separate treatment effect concerns from censoring concerns

Exact vs. approximate results

- Note that because the exact distribution of V is known and easy to work with, it is possible to carry out exact power and sample size calculations
- However, one can obtain much simpler, closed-form expressions through a normal approximation
- Personal opinion: In an actual data analysis, exact results are quite desirable, but in a power analysis, the inaccuracy of the approximation is typically a minor concern compared to all other potential sources of error that go into the calculation

Central limit theorem

- The exponential distribution has mean $1/\lambda$ and variance $1/\lambda^2$
- Thus, by the central limit theorem,

$$\bar{X} \dot{\sim} \mathrm{N}\left(\frac{1}{\lambda}, \frac{1}{n\lambda^2}\right)$$

- This result, however, is not particularly satisfactory due to the λ term in the variance, which means we will have to solve a nonlinear equation to determine power/sample size

Log transform

- Consider instead the variance-stabilizing transformation $g(x) = \log(x)$
- By the delta method,

$$\log \bar{X} \dot{\sim} \mathrm{N}\left(-\log \lambda, \frac{1}{n}\right)$$

- In addition to the convenience of linearity, variance-stabilizing transformations also typically lead to more accurate normal approximations

Two samples: Hazard ratio

- With these preliminaries out of the way, let's get to the actual business of comparing two samples
- Let $X_i \overset{\perp\!\!\!\perp}{\sim} \mathrm{Exp}(\lambda_1)$ and $Y_i \overset{\perp\!\!\!\perp}{\sim} \mathrm{Exp}(\lambda_2)$, with $X_i \amalg Y_i$
- We have

$$\log\left(\frac{\bar{Y}}{\bar{X}}\right) \dot{\sim} \mathrm{N}\left(\log\Delta, \frac{1}{n_1}+\frac{1}{n_2}\right),$$

where $\Delta = \lambda_1/\lambda_2$ is the hazard ratio

Power formula

- Thus, letting $Z = \log(\bar{Y}/\bar{X})/\sqrt{1/n_1 + 1/n_2}$, we have

$$\text{Under } H_0 : Z \dot{\sim} \mathrm{N}(0,1)$$

$$\text{Under } H_A : Z \dot{\sim} \mathrm{N}(0,1) + \frac{\log \Delta}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

- The critical value for Z is therefore $\mathrm{CV} = \Phi^{-1}(1 - \alpha/2)$, where α is the type I error rate and Φ is the CDF of the standard normal distribution
- Without loss of generality, we can take $\Delta > 1$, which yields

$$\text{Power} = 1 - \Phi\left(\mathrm{CV} - \log \Delta/\sqrt{1/n_1 + 1/n_2}\right)$$

Sample size formula

- In order to solve for the sample size(s) that yield a power of $1-\beta$, we must solve for the values of n_1 and n_2 that satisfy the following equation:

$$z_{1-\alpha/2} = -z_{1-\beta} + \log\Delta/\sqrt{1/n_1 + 1/n_2},$$

  where z_q is the qth quantile of the standard normal distribution

- In the special case of $n = n_1 = n_2$, we therefore have

$$n = 2\left(\frac{z_{1-\alpha/2} + z_{1-\beta}}{\log\Delta}\right)^2$$

  as the per-group sample size

Remarks

- Note that we do not even need to specify λ_1 and λ_2 to calculate power and sample size: we only need their ratio, Δ
- Furthermore, note that for the exponential distribution, the median survival time is $\lambda^{-1} \log 2$
- Thus, the effect size can be equivalently thought of as a ratio of median survival times, rather than a hazard ratio, which in my experience is convenient as non-statisticians typically prefer to think in terms of median survival times than hazards

NSCLC study: Background

- To illustrate how these formulas are used in practice, I'll discuss the planning of a study at the Holden Cancer Center here at the University of Iowa that I was involved in
- The study was looking at progression-free survival (PFS) in patients with refractory non-small cell lung cancer
- Historically, the median PFS for these patients is around 2.5 months
- The investigators hypothesized, however, that a novel combination of protein kinase inhibitors and a cytokines could extend PFS by 50%

Sample size

- A 50% increase in median PFS corresponds to $\Delta = 1.5$
- Thus, to achieve 80% power under 5% type I error rate control (these are typical numbers), we require

  $$\begin{aligned} n &= 2\frac{(1.96 + 0.84)^2}{(\log(1.5))^2} \\ &= 95.5 \end{aligned}$$

  events in each arm of the study
- The actual study, however, was only a "single-arm" study

Single arm study

- In a single-arm study, one assigns all patients to the experimental therapy, with the intention of comparing it to historical controls
- The use of historical controls is clearly subject to all sorts of biases, and a randomized trial would be preferable
- However, single arm studies like this one are common in what is called “Phase II” of clinical trial research
- The goal of a Phase II study is to learn about the clinical efficacy of a treatment; if it appears promising, one would then continue on to a fully randomized trial in Phase III
- Note that for a single-arm study (treating the control group as a known constant), the number of events in the experimental arm is cut in half (i.e., the total sample size is cut by 3/4)

Censoring and accrual

- In this study, since these are patients with very poor prognosis and a median PFS of only 2.5 months (or ≈ 4 months, if the treatment is effective), we anticipated that only a small fraction of patients would remain censored at the end of the study
- Specifically, we made an assumption of 20% censoring, and included the following language in the proposal:

  > *Power calculations indicate that to achieve 80% power to detect a 50% increase in median PFS with a 5% type I error rate, 48 events must be observed. Allowing for a 20% censoring rate, we therefore plan to enroll 58 patients.*

Study duration

- The duration of a study is also an important concern in planning a study with a time-to-event outcome
- In the NSCLC study, the accrual rate was anticipated to be approximately 50 patients per year
- We therefore made the conservative estimate that we could enroll our 58 patients in 18 months, and that we should be able to conclude the whole study within 2 years

Formal approach

- But is this really an adequate amount of time in which to observe 48 events?
- To address this question, let's work through how to calculate the expected duration of a study
- To start, let $(0, T]$ denote the "entry" or "accrual" period of the study, and $(T, T+\tau]$ denote the follow-up period

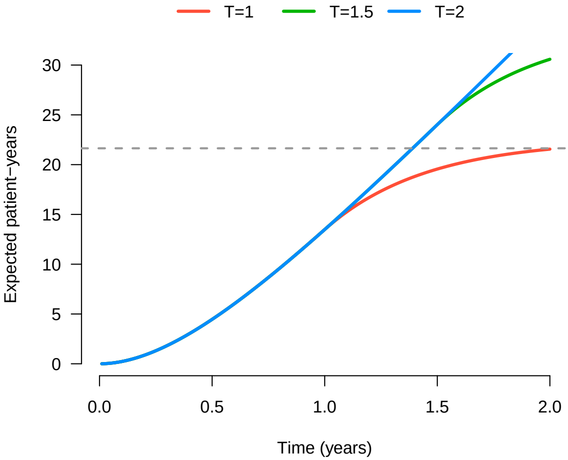
Formal approach (cont'd)

- One widely used approach (which is also the approach used by George & Desu) is to use the fact that the expected number of patient-years necessary to observe d events is d/λ
- Furthermore, letting $Y(t)$ denote the number of patient-years accumulated by time t and a denote the average accrual rate,

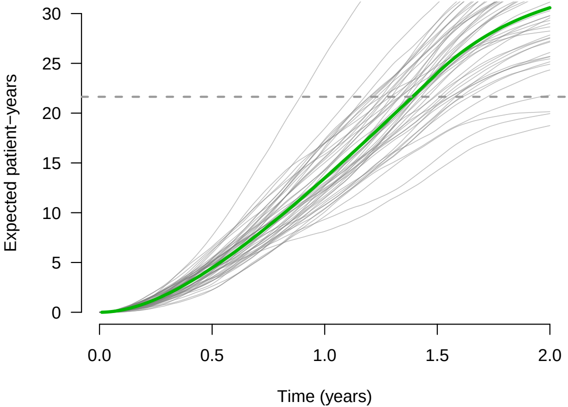
$$\begin{aligned}\mathbb{E}Y(t) &= a\int_0^{t^*}\int_0^{t-v} S(u)\,du\,dv \\ &= \frac{at^*}{\lambda}\left\{1-(\lambda t^*)^{-1}e^{-\lambda t}(e^{\lambda t^*}-1)\right\}\end{aligned}$$

where $t^* = \min(T, t)$

NSCLC study duration: Accrual 50 / year



50 random instances at 50 / year, $T = 1.5$


Expected patient−years
Time (years)