Top Design Challenges

Top 10 Series

Paul J. Catalano

June 11, 2019

Outline

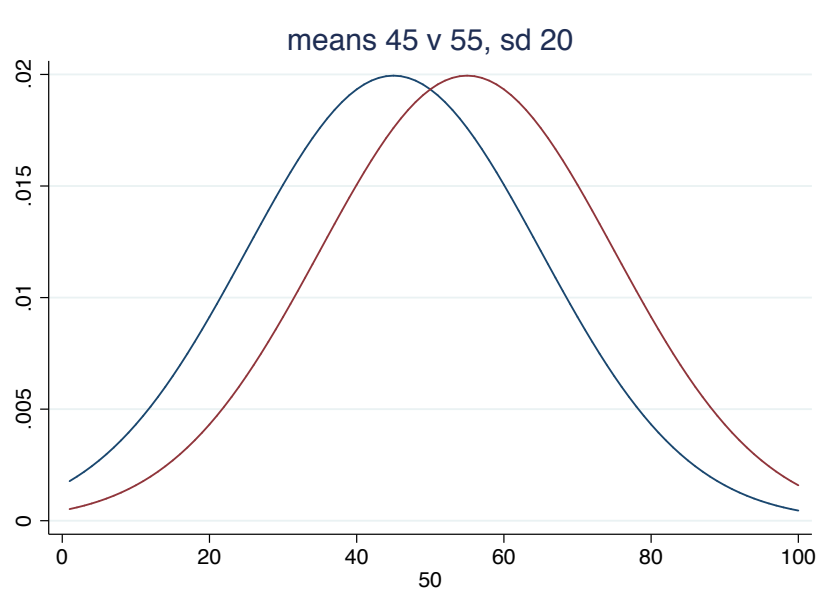
- Discuss some common design, sample size, power problems and related issues
- Intended audience: analysts, clinical researchers, translational and basic science researchers as well
- None of the math is complicated

- Making the right choices and having the right data are complicated
- Thoughts apply to prospective studies but also many retrospective, translational and basic science investigations as well

“The outcome is continuous but let’s just split it at the [median, mean, some value X]”

- Common attempt at trying to ‘simplify’ the situation
- Immediate loss of power and ability to understand the outcome in all but the simplest way

- Problematic (but commonly suggested) as a predictor as well

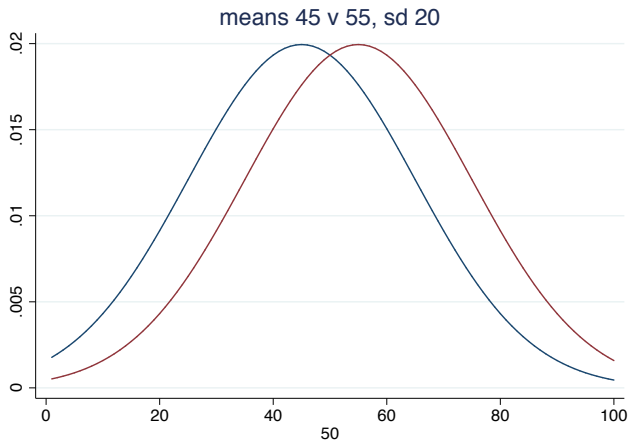


- Consider two groups with means 45 and 55 and common SD of 20
- Test of means can be done with 64 per group
- 80% power two-sided 5% level test (assumes normal)

- Maybe say 70 per group if considering the rank sum test
  (and still normal - otherwise we need to be more specific)

If we dichotomize at mean/median of 50

- Group 1 expected proportion above split: 40.1%
- Group 2 expected proportion: 59.9%
- Sample size for comparing two binomials: 99 per group
  - More than 50% larger sample size required



Worse if we split somewhere else, e.g. say at 65

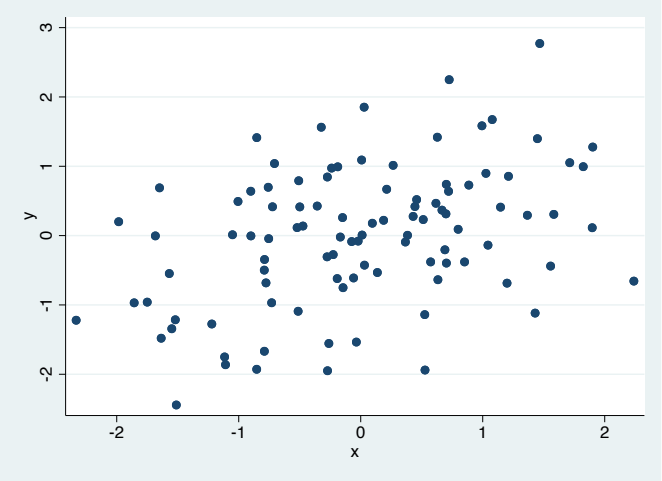
- G1 expected proportion: 15.9%
- G2 expected proportion: 30.9%
- Required sample size for 80% power: 124 per group

In defense of dichotomization

- Resulting sample size would be conservative (wasteful?)
- Distribution might not be well known (weak defense!)
- As a predictor, it is simpler to think about that way (2 groups)

Regarding predictors, probably better to test for trend or association than a shift in means

- True correlation of 0.40 requires only 47 subjects or samples total
- If we split X at the median and compared the means of Y in the 2 groups, we'd need 45 subjects *per group* (90 total)



“The cohort will contain different stage* patients, so let’s stratify on it”

* Or whatever

- One of the most confusing / misinterpreted terms in design
- Often translates (wrongly) into

  “we can learn something within the strata subsets”
- Designs incorporating stratification
  - Do *not* imply there is power within subsets
  - Do *not* imply there is power for any interaction
- This really needs to be clearly communicated

Stratification

- Does mean (randomized context) that groups will be *balanced* within the strata
- Addresses confounding or imbalance due to bad luck
  - Many methods to accomplish the balancing
- But…no power for subset analyses (what even will be proportions within strata?) and certainly interaction is not a goal of this design
  - Interaction can be analyzed later but is not the focus
- Stratified power/sample size calculations can be done
  - But need to know the strata proportions expected AND the event rates within the strata (common effect [OR / HR] is assumed though across strata)
- Stratified analyses have appeal
  - Especially for time-to-event since underlying hazards can differ
  - No direct measure of comparisons between strata though in primary analysis

“We don’t expect any responses, so let’s focus on PFS”

- Very tricky from lots of perspectives
- Single sample classical objective response $p_0 = 0.05$ vs $p_A = 0.20$ gives
  - Fixed sample size: N = 32 exact binomial w/ 90% power, 10% 1-sided α
  - Simon’s two-stage design, N = 37 with $N_1 = 12$ and 1 response needed
- Cytostatic agents, targeted therapies, immunotherapies, *etc* might not expect a ‘response’
- Focus on disease stabilization
- Translates often into look at time to progression or PFS

Statistical issues to grapple with…

- Problem is anchoring the null hypothesis
  - Here let's use the example of PFS
- Advantage: time-to-event endpoint better than binomial
- But, the PFS metric poses a problem in a dynamic landscape
  - Even vs standard therapy; worse when improving upon new combination
- Large literature on problems with using imprecise nulls in early therapeutic context … easy to go wrong either way
- Reported metrics and associated sample sizes are too small to have good faith in setting a proper null hypothesis (wide CIs)

Some examples…

- Set null median 6m anticipate 70% increase: alt median 10.2m
- Typical parameters of 10% 1-sided α, 90% power
- Single sample exp survival: N = 43
- Inference can go wrong quickly either way …
- If assumed medians too high by, say, 1.5m then power is only 70%
  - You are still assuming a 6m null but you're off on your alternative
- If your null was set too low, you're in for a type I error
- Problem in either case is too much reliance on historical data
  - One a number is published it is taken as gospel…

Solution?

- Consider a concurrent control arm: randomized phase II context
- Again, control median 6m, experimental median 10.2m
- Log rank test with 90% power 1-sided 0.10 α yields total N = 110
- More than double single arm but buys **a lot** of protection

- Typical to have at least one efficacy/futility interim analysis
  - If one at 50% information time, minimal increase in sample size (5)
- Could consider 2:1 (or other) randomization schemes
  - 2:1 here in favor of experimental only requires total N = 125
- Could consider more liberal operating characteristics
  - Be careful with relaxing power
  - Be aware you will need to explain the large α

“We need to focus on patients positive for the biomarker…”

Very common today

- protocols, grants, correlative analyses, pilot studies, you name it…

Very tricky for design

- Tons of unknowns and/or critical quantities that need specificity to anchor the design

Issues similar to earlier PFS discussion but more challenging

“We need to focus on patients positive for the biomarker…”

In the prospective trial context…

- What proportion of subjects will form the target population?
  - And who pays to find out?
  - Target pop might be tiny so lots of inflation of sample size (AND TIME)
- What about the negative patients? Is there a research plan for them?
  - Might be the vast majority of patients screened for the study
- What are the expected event rates or outcomes in this subset?
  - This is the largest problem and usually requires contemporaneous control
  - What control treatment is reasonable or even ethical? Are there data for this subset?
- What level of effect/improvement is clinically or biologically reasonable?
  - Often investigators are overly optimistic about targeted agents in enriched population

Biomarker designs, cont'd

In the correlative/translational sphere…

- Samples to be analyzed for biomarker(s), from some patient cohort w fup
- Power/sample size often of interest, but …
  - How to attribute the marginal event rate to the biomarker subsets?
  - What difference or effect is expected across biomarker groups?
  - How sensitive is power to the expected biomarker proportions?
  - How to properly sample cases for analysis when the overall cohort is large?
    - Helps to have clinical endpoints already with responses, progressions, recurrences, *etc* known
- What is the biomarker goal? Purely 'biological'? Prognostic? Predictive?
- Main effect or interaction with treatment?
  - Power for an interaction needs to be carefully thought out
  - Sometimes have to 'back in' to the sample size available
- How were patients treated? On study, off study, retrospective 'database'?
- Is pooling of cohorts possible or does it even make sense?

Biomarker designs, cont'd

In the grant writing context…

- Similar problems vis-a-vis construction of design parameters
  - Often no or little data on the biomarkers … prelim data sparse
  - Recall the problem of guessing nulls in face of little (or no) data
    - "Tumor growth was reduced by 50% in 8 mice…"
    - "Of 7 patients, the 4 with positive receptors were progression-free with 9 months fup"

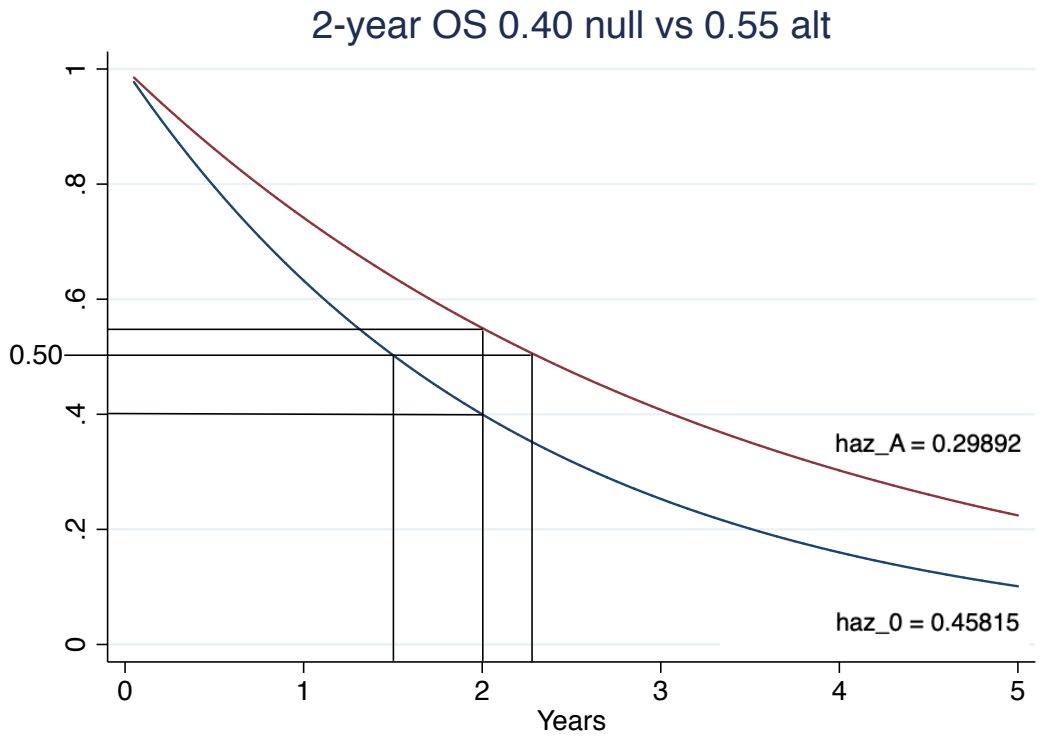
- Often need to also worry about multiplicity, α-adjustment
  - What are the primary objectives? What is exploratory?
  - Problem with 'co-primary' objectives…what does it mean?
    - Both significant? One or the other? Hierarchical?
    - What is an appropriate evaluation / testing strategy?

"Two-year survival is 40% and the new treatment should bring it up to 55%"

- Common context for evaluation of improvement in time-to-event outcomes
- Lots of confusion over metrics, however
- Almost always, a specific time point analysis or focus is *NOT* of interest
- Need to orient investigators and probe a little
  - Are other metrics available? Median survival? Other N-year survival? K-M curve?
  - Actual hazard rates are almost never available

Let's be more specific about the design parameters

- Stick with 90% power, α = .1 for log rank test
- Assume exponential dist
  Implies medians of 1.51y, 2.32y
- And 2-yr OS of .40 and .55 correspond to HR of 0.65 or 53% increase in median OS
- 15% absolute increase in 2-yr OS but a 53% increase in median survival



Survival: Design versus Metrics

- For the log rank test, requirement would be N = 200 total
  - Full comparison of the two survival distributions of course (need 143 events)
- For the fixed 2–yr time point analysis (2 binomials), N = 288 total
  - 44% larger and assumes at least 2 years of fup; worse if allowing for censoring
- Even worse yet if we tried to design around 1-yr survival
  - Now compare 0.63 null vs 0.74 alt at 1-yr
  - Binomial approach would require N = 466
  - Only an extra year of fup but stopping then would not be advised

Survival: Design versus Metrics

- Need to clarify the design to investigators or in write-up
- Often an educational opportunity
- Analysis is *NOT* focusing on a particular time point
  - Also does not focus on comparing medians either
  - Unfortunate consequence of the way we report survival

In defense of the fixed time point analysis…

- Sometimes a given time point itself *is* of interest
- Or want to assess 'long term survivors'
- Or worried about proportional hazards or some other assumption
  - For example, cure rate … if so, then use a cure rate model

Some design software resources

- R: lots of packages (e.g., pwr)
  - Don’t forget about Bob Gray’s desmon package (it’s great!)
- SAS: PROC POWER (complex but very interesting options)
- STATA: power command (don’t forget STATA is on Unix too!)
- PC (via arcem remote desktop)
  - nQuery
  - EaSt
- Local Unix: many resources
  - phase2, twolook, twostg, twocon, b2p, confin
  - seqopr6, seqpwr6, sequse6
  - stplan

Software resources, continued

- Web resources
  - Simon's two-stage design
    cancer.unc.edu/biostatistics/program/ivanova/SimonsTwoStageDesign.aspx
  - SWOG's Statistical Tools
    stattools.crab.org
  - MGH Biostatistics Center - Power and Sample Size Tools
    hedwig.mgh.harvard.edu/sample_size/size.html
  - I'm sure you have your favorites too…