HEIDELBERG
UNIVERSITY
HOSPITAL

2018 Summer School Lambrecht

Phase II: Single-arm Trials and Optimization under Uncertainty

21/06/2018

Kevin Kunzmann
kevin.kunzmann@mrc-bsu.cam.ac.uk

imbi

Outline

1. Why phase-II?
2. Why single-arm?
3. Why optimize?
4. What to optimize?
5. How to optimize?
6. Demo
7. Remarks

Why Phase-II?

- phase-II serves as ‘gate-keeper’
- not necessarily demonstrate efficacy …
- … but filter futile candidate drugs early!
- phase-II closes the gap between phase-I (purely exploratory) and phase-III (randomized + confirmatory)

Consequence:

- wide range of objectives, wide range of methods - ‘phase-II’ trial rather an umbrella term
- Here: specifically phase-II designs in oncology (of course: anything said for phase-I and III might also apply!)

Why single-arm?

- Smaller sample size, thus
- ... less recruitment problems in small, targeted population
- ... potentially more attractive to participants (no lottery to treatment)!
- ... faster study completion
- ... lower cost
- In oncology: randomization potentially unethical if
  - no gold-standard treatment at all or ...
  - ... large effect size anticipated
  - cf. 'Right-to-Try bill' 2018

... and why maybe not single-arm?

- No randomization: no causal treatment effect estimable without further assumptions!
- Must compare to historical data (response rate)
- No blinding possible - additional bias potential
- Single-arm trials typically not sufficient for 'early-approval' by FDA (evidence from randomized trial required if possible)

Single- vs. two- or multi-arm trials

- Choice depends on study objective, feasibility and ethical considerations, more: cf. [3]
- Single-arm trials well-established in oncology
- Bayesian approach to planning a trial most important when uncertainty is high during planning (early phase-II)
- Most principles explained in the following are applicable to any trial design (and being worked on!)
- Principles are simpler to explain by example of single-arm trials

⇝ *continue with single-arm trial for the moment*

Two-stage designs

- Idea: perform interim analysis after n_1 patients and continue depending on this interim outcome
- Most importantly: binding early stopping for futility to protect against ineffective treatment
- Early stopping for efficacy: potentially further speeds-up development program
- More than two stages often infeasible (delayed response), little gains

Binary endpoint: two-stage more important

- In oncology often: binary endpoint (tumor response, early surrogate)
- **Question**: response rate ρ larger than historic rate ρ_0?
- Test null $\mathcal{H}_0 : \rho \leq \rho_0$ for given α and β at $\rho_1 > \rho_0$
- Single stage binomial test is simplest solution:
- $X \sim \mathrm{binom}(n, \rho)$, reject $\mathcal{H}_0$ iff $X > c$
- Choose $c(n)$ as the maximal c such that $\mathbf{P}_{\rho_0}\left[X > c\right] \leq \alpha$
- Pick minimal n such that $\mathbf{P}_{\rho_1}\left[X > c(n)\right] \geq 1 - \beta$
- Fixed-size test ineffective due to discreteness!

Ineffectiveness of binomial test

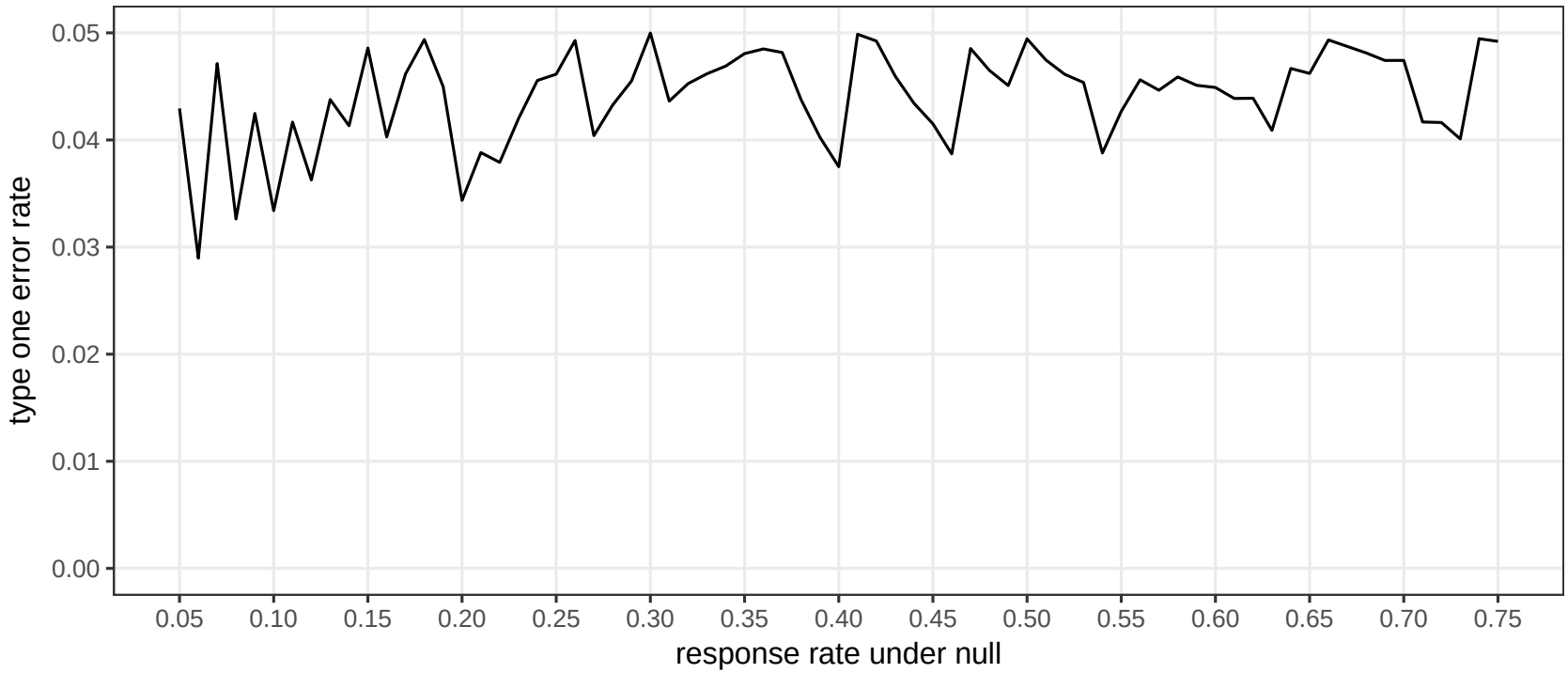


Figure: True type one error rate for $\alpha = 0.05, \beta = 0.2$, and varying response rate under the null (ρ_0) for the simple binomial test

Two-stage designs for binary endpoint

- Two-stage design: allow early stopping
- Three new parameters:
  - n_1 (stage-one sample size)
  - c_f (early futility boundary, stop if $X_1 < c_f$)
  - c_e (early efficacy boundary, stop if $X_1 > c_e$)
- More degrees of freedom: Designs can better exhaust the given α level (i.e. almost perfectly!)
- Final sample size no longer constant but function of X_1
- 'Smallest' sample size no longer well defined as random!
- Idea: minimize expected sample size [1]

Optimization problem

$$\underset{n_1,n,c_f,c_e,c}{\text{argmin:}} \quad \mathbf{P}_{\rho}\left[X_1 \in (c_f, c_e)\right] n + \mathbf{P}_{\rho}\left[X_1 \notin (c_f, c_e)\right] n_1 \tag{1}$$

$$\text{subject to:} \quad \mathbf{P}_{\rho_0}\left[X_1 > c_e\right] + \mathbf{P}_{\rho_0}\left[X_2 > c - X_1, X_1 \leq c_e\right] \leq \alpha \tag{2}$$

$$\mathbf{P}_{\rho_1}\left[X_1 > c_e\right] + \mathbf{P}_{\rho_1}\left[X_2 > c - X_1, X_1 \leq c_e\right] \geq 1 - \beta \tag{3}$$

- Technically easy to solve by exhaustive search

Different objective criteria

- $\mathbf{E}_\rho\big[n(X_1)\big]$ depends on unknown ρ
  (if $n \equiv n_{const}$, $\mathbf{E}_\rho\big[n(X_1)\big] = n_{const}$ for all ρ)
- $\rightsquigarrow$ 'optimal' two-stage design no longer unique!
- Mander and Thomson [2] studied optimal designs under both ρ_0 and ρ_1
- In the following: $\alpha = 0.05, 1 - \beta = 0.8$

Different objective criteria

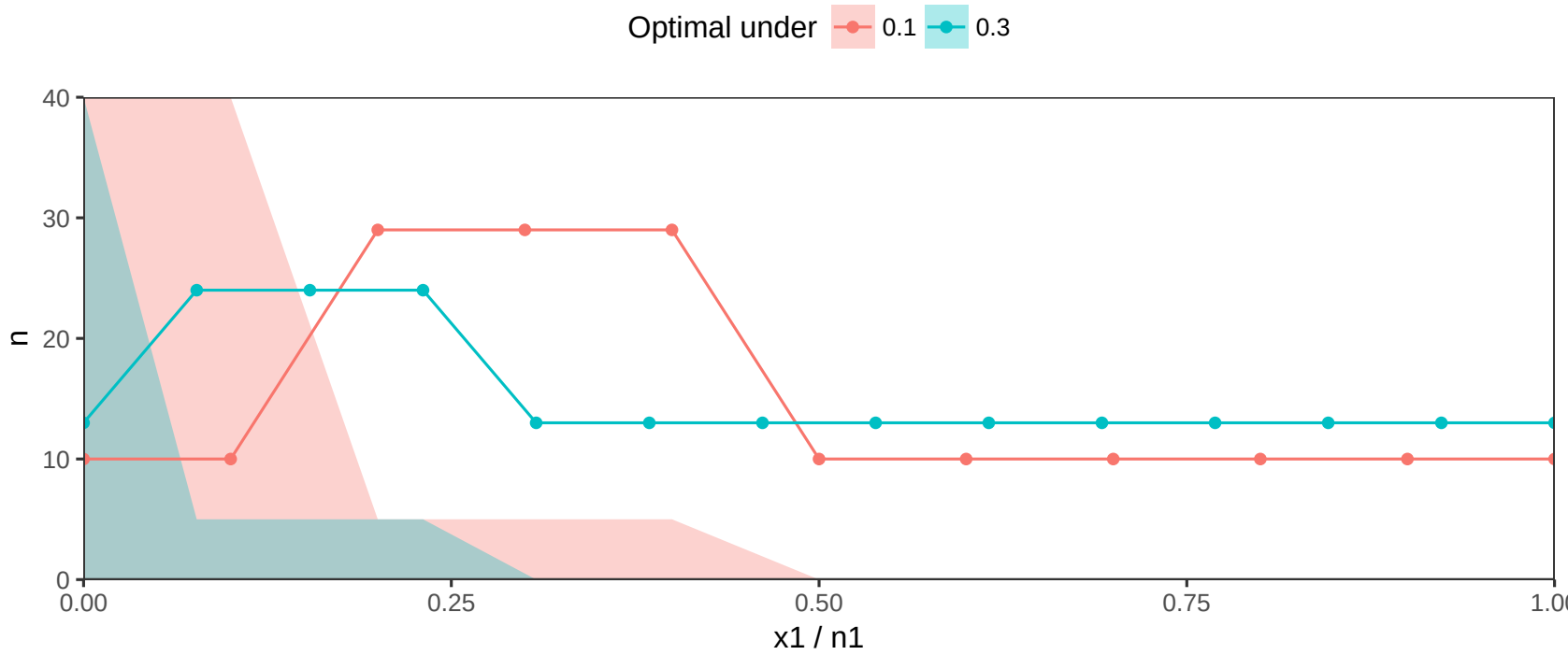


Figure: Visual comparison of the optimal two-stage designs of Mander and Thmason for expected sample size under the alternative ($\rho = 0.3$) or the null hypothesis ($\rho = 0.1$); shaded are is rejection region

Why should second-stage be constant?

- So far: second stage sample size fixed upon continuation, why?
- Rationale: larger x_1 give more evidence against $\mathcal{H}_0$ $\rightsquigarrow n(\cdot)$ could be decreasing in x_1
- More general definition of two-stage design: $(n_1, c(\cdot), n(\cdot))$ where $\mathcal{H}_0 : \rho \leq \rho_0$ is rejected iff $X_1 + X_2 > c(X_1)$, $X_1 \sim \text{binom}(n_1, \rho)$, $X_2|X_1 = x_1 \sim \text{binom}(n(x_1) - n - 1, \rho)$.
- $c(x_1) = -\infty \rightsquigarrow$ early efficacy
- $c(x_1) = \infty \rightsquigarrow$ early futility

Example

x_1 :	≤ 1	2	3	4	≥ 5
$n(x_1)$:	10	29	29	29	10
$c(x_1)$:	∞	5	5	5	-∞

Table: ρ_0-optimal design of Mander and Thomson for $\rho_0 = 0.1$, $\rho_1 = 0.3$, $\alpha = 5\%$, $1 - \beta = 80\%$.

Modified optimization problem

$$\underset{n_1, n(\cdot), c(\cdot)}{\text{argmin:}} \quad \sum_{x_1=0}^{n_1} \mathbf{P}_{\rho}\left[X_1 = x_1\right] n(x_1) \tag{4}$$

subject to:

$$\mathbf{P}_{\rho_0}\left[X > c(X_1)\right] \leq \alpha \tag{5}$$

$$\mathbf{P}_{\rho_1}\left[X > c(X_1)\right] \geq 1 - \beta \tag{6}$$

$$n(x_1) = n_1 \Leftrightarrow c(x_1) \in \{-\infty, \infty\} \tag{7}$$

$$c(x_1) = \infty \Rightarrow c(x_1 - 1) = \infty \tag{8}$$

$$c(x_1) = -\infty \Rightarrow c(x_1 + 1) = -\infty \tag{9}$$

- Much more difficult to optimize, still discrete
- Exhaustive search is futile, gradient descent does not work
- Details later

Where to minimize sample size

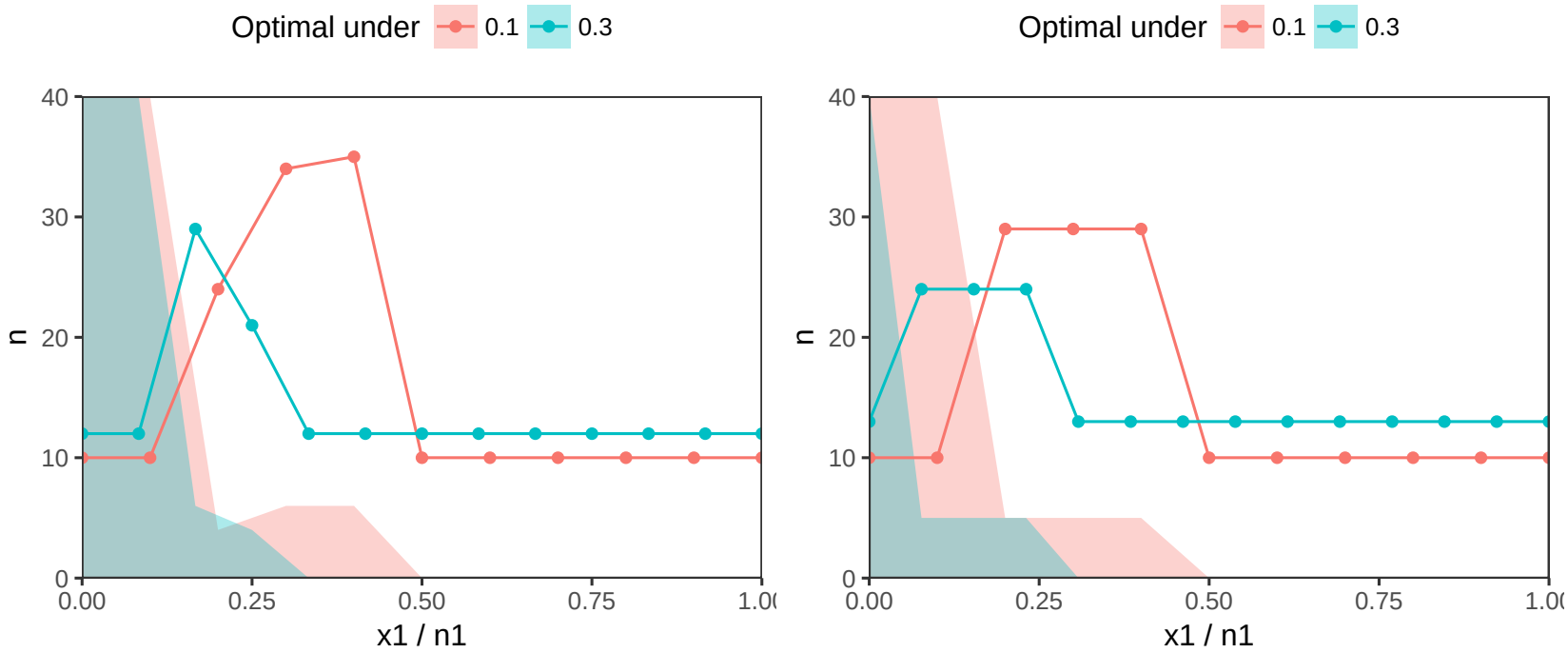


Figure: Visual comparison of the optimal two-stage designs with flexible second stage for expected sample size under the alternative ($\rho = 0.3$) or the null hypothesis ($\rho = 0.1$) (left) and constant second stage (right)

Performance gains?

When stage-two sample size is flexible:

- Difference between designs even more pronounced
- Performance gain over constant second stage is modest (< 0.5 patients on average under the respective objective)
- But: why not?

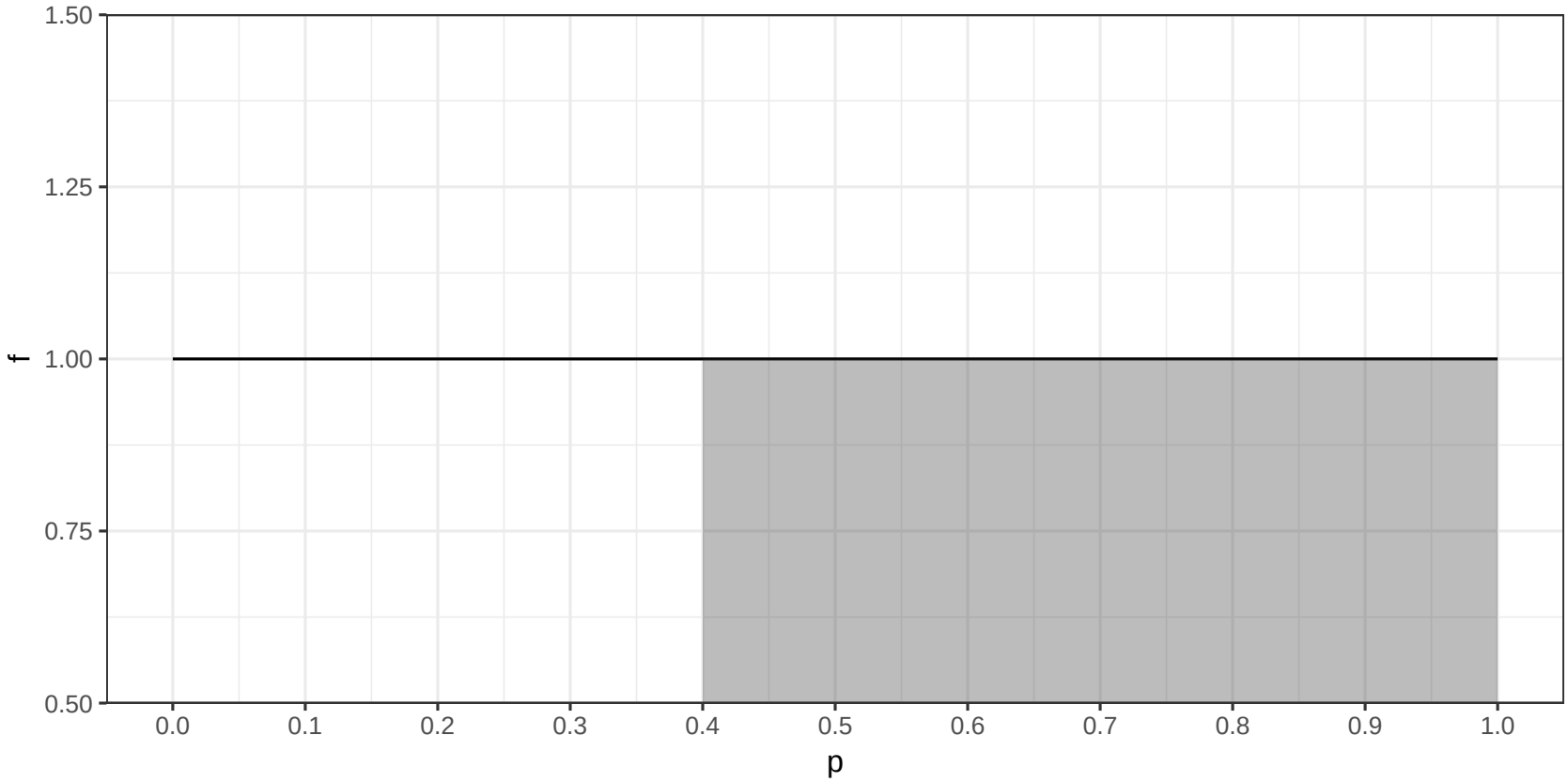
What to optimize

- Obviously, very different optimal designs for varying ρ
- Sponsors unhappy with minimizing $\mathbf{E}_{\rho_0}\left[n(X_1)\right]$
- Patients unhappy with minimizing $\mathbf{E}_{\rho_1}\left[n(X_1)\right]$
- Both positions are extreme in that they assume a single parameter value to be true, but ρ is uncertain during planning!
- Principled way of modeling planning uncertainty about ρ is by using a Bayesian approach
- Assume $\rho \sim F$ for prior distribution F with PDF $f(\rho)$

Example

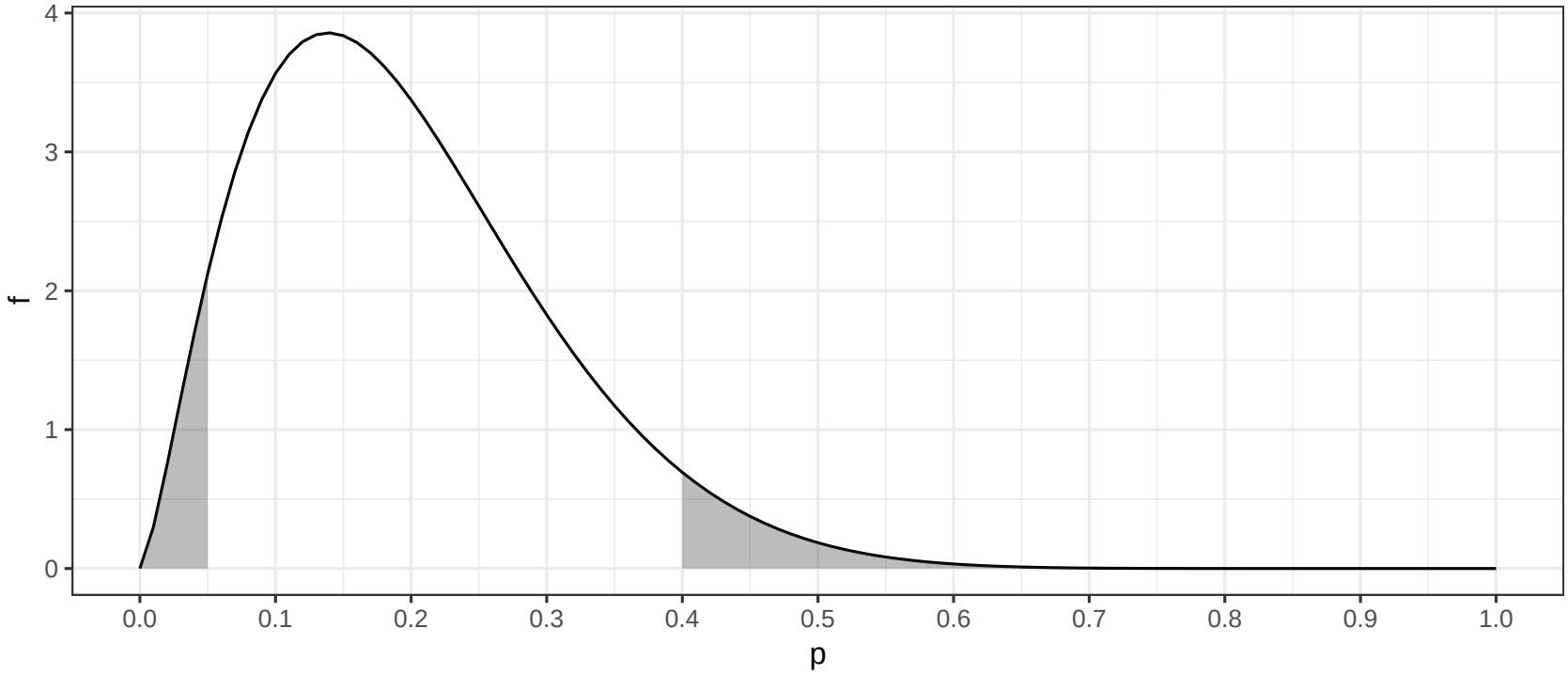
- Developing new anti-cancer drug
- 6 months RECIST objective response rate historically about 0.1
- Experts believe that response rate under new drug lies in $[0.05, 0.4]$
- maybe - phase-I data of 10 subjects is available with 4 responses
- How do we get a prior density f?

How to (not) construct a prior



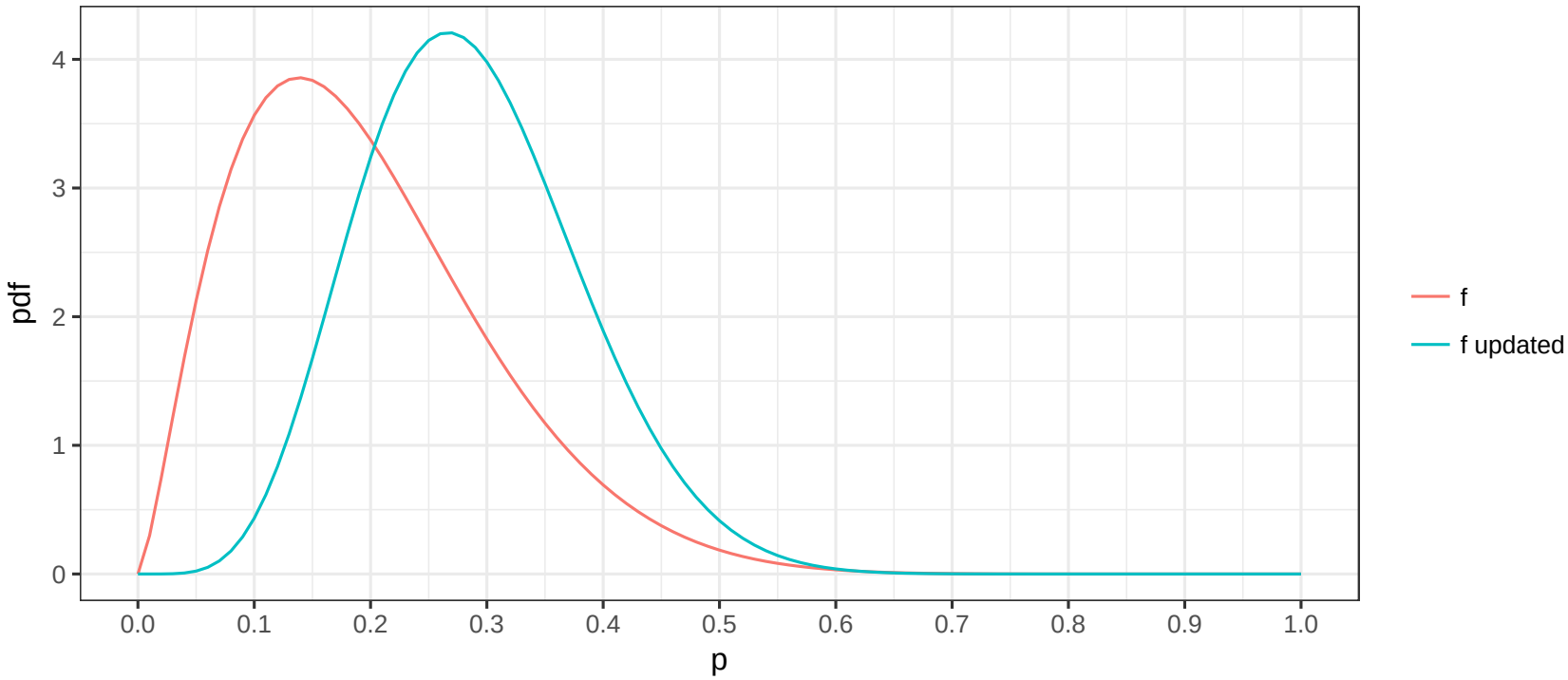
- Non-informative obviously nonsense: $\mathbf{P}\left[\rho > 0.4\right] = 0.6$

A quick informative prior



- Better: fit Beta distribution so that $\mathbf{P}\left[\rho < 0.05\right] = \mathbf{P}\left[\rho > 0.4\right] = 0.05$ - here $\mathrm{Beta}(2.45, 10.0)$

Incorporating previous results



- Advantage of using Beta: analytical update available (conjugate prior)
- Phase-I updated prior would be $\mathrm{Beta}(2.45 + 4, 10.0 + 6)$

Result: less flexible design

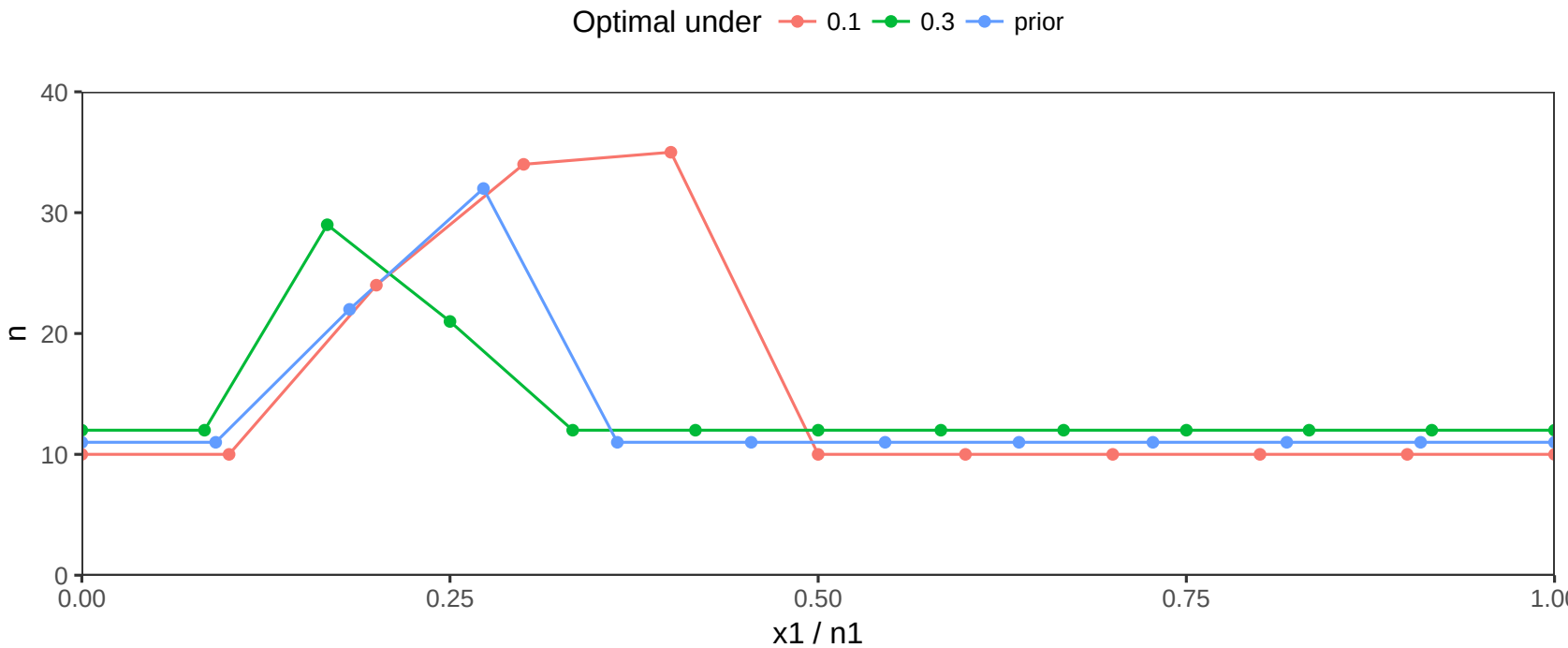


Figure: Effect of using prior $\mathrm{Beta}(2.45, 10.0)$ (assuming no phase-I data) on shape of optimal sample size function.

What about power?

- Incorporated planning uncertainty about ρ in objective (expected sample size) ...
- ... but not in constraints - why only power at ρ_1?
- Power is also a function of (unknown) ρ
- If only response rates $> \rho_1$ are of interest, powering solely on ρ_1 is fine
  (minimal clinically relevant response rate)
- If ρ_1 is considered the most likely effect, there is uncertainty and the power constraint should reflect that.

A 'Bayesian' Power constraint 1)

- Natural: integrate power curve with respect to f:

$$\int_0^1 \mathbf{P}_\rho\big[X > c(X_1)\big] f(\rho)\,\mathrm{d}\,\rho \tag{10}$$

- But if $\rho \leq \rho_0$ power is actually type one error rate! I.e. the part

$$\int_0^{\rho_0} \mathbf{P}_\rho\big[X > c(X_1)\big] f(\rho)\,\mathrm{d}\,\rho \tag{11}$$

  corresponds to the expected type one error rate
- Expected power should be conditional on a minimal relevant response rate - say $\rho_0 + \delta$ (typically $\delta = 0.05 - 0.2$):

$$\int_{\rho_0+\delta}^1 \mathbf{P}_\rho\big[X > c(X_1)\big] f(\rho \,|\, \rho > \rho_0)\,\mathrm{d}\,\rho \tag{12}$$

Is there an alternative?

- Expected power seems to be intuitive
- But how does it react when we ‘tweak’ the power curve of our design?
- Expected power is a functional of the power curve
- Functional derivative with respect to local variation in power curve can be computed

Functional derivative of expected power

$$\frac{\partial}{\partial \mathbf{P}_\rho\big[X > c(X_1)\big]} \int_{\rho_0+\delta}^{1} \mathbf{P}_\rho\big[X > c(X_1)\big] f(\rho \,|\, \rho > \rho_0 + \delta)\, \mathrm{d}\, \rho$$
$$= f(\rho \,|\, \rho > \rho_0 + \delta)$$

- Intuitive: increase power for most likely values of ρ
- But: ignores level of power, i.e. does not matter if power increases from 20% to 21% or 70% to 71%
- realistic?

Alternative to averaging power

- Don’t want correct power on average - doing the trial only once
- Rather: sufficient power (say 80%) in most cases (say 66%)
- I.e. want $\mathbf{P}\left[\,\mathrm{Power}_{\mathcal{D}}(\rho) > 0.8 \,|\, \rho > \rho_0 + \delta\right] \geq 0.66$
- Gives high a-prior chance (here 66%) of ending up with sufficient power (given favourable response rate) instead of getting correct power on average

Looks complicated but ...

- Note power curve is monotone!
- Let ρ^* be defined by $\text{Power}_{\mathcal{D}}(\rho^*) = 1 - \beta$

$$\begin{aligned}
&\mathbf{P}\left[\text{Power}_{\mathcal{D}}(\rho) > 1-\beta \,|\, \rho > \rho_0 + \delta\right] && \geq \eta && (13)\\
\Leftrightarrow\ &\mathbf{P}[\rho > \rho^* \,|\, \rho > \rho_0 + \delta] && \geq \eta && (14)\\
\Leftrightarrow\ &\mathbf{P}[\rho \leq \rho^* \,|\, \rho > \rho_0 + \delta] && \geq 1-\eta && (15)\\
\Leftrightarrow\ &\rho^* \text{ is } 1-\eta \text{ quantile of } f(\rho \,|\, \rho > \rho_0 + \delta) && && (16)
\end{aligned}$$

- Thus: compute ρ^* during planning and get optimal design subject to $\text{Power}_{\mathcal{D}}(\rho^*) \geq 1 - \beta$
- Just a more principled way of choosing ρ_1!
- Widely applicable without changing techniques!

Let's compare

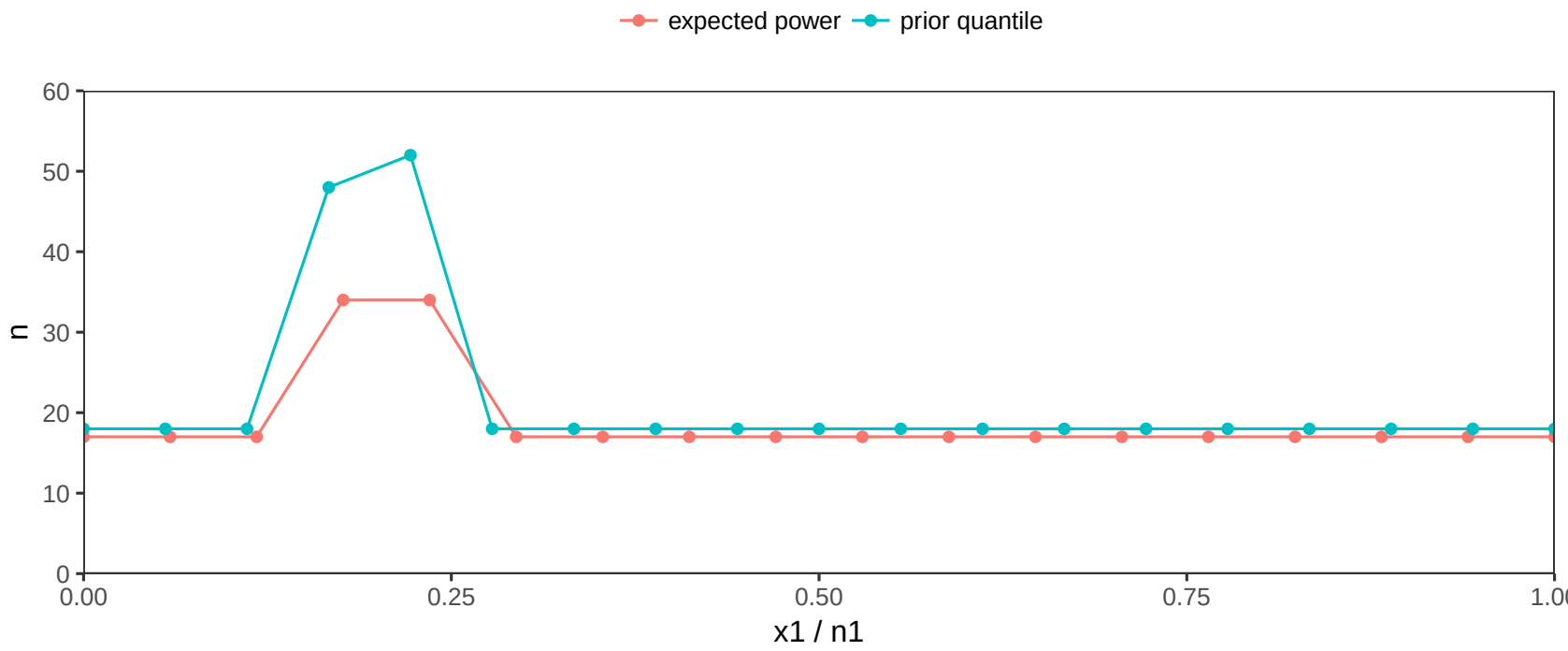
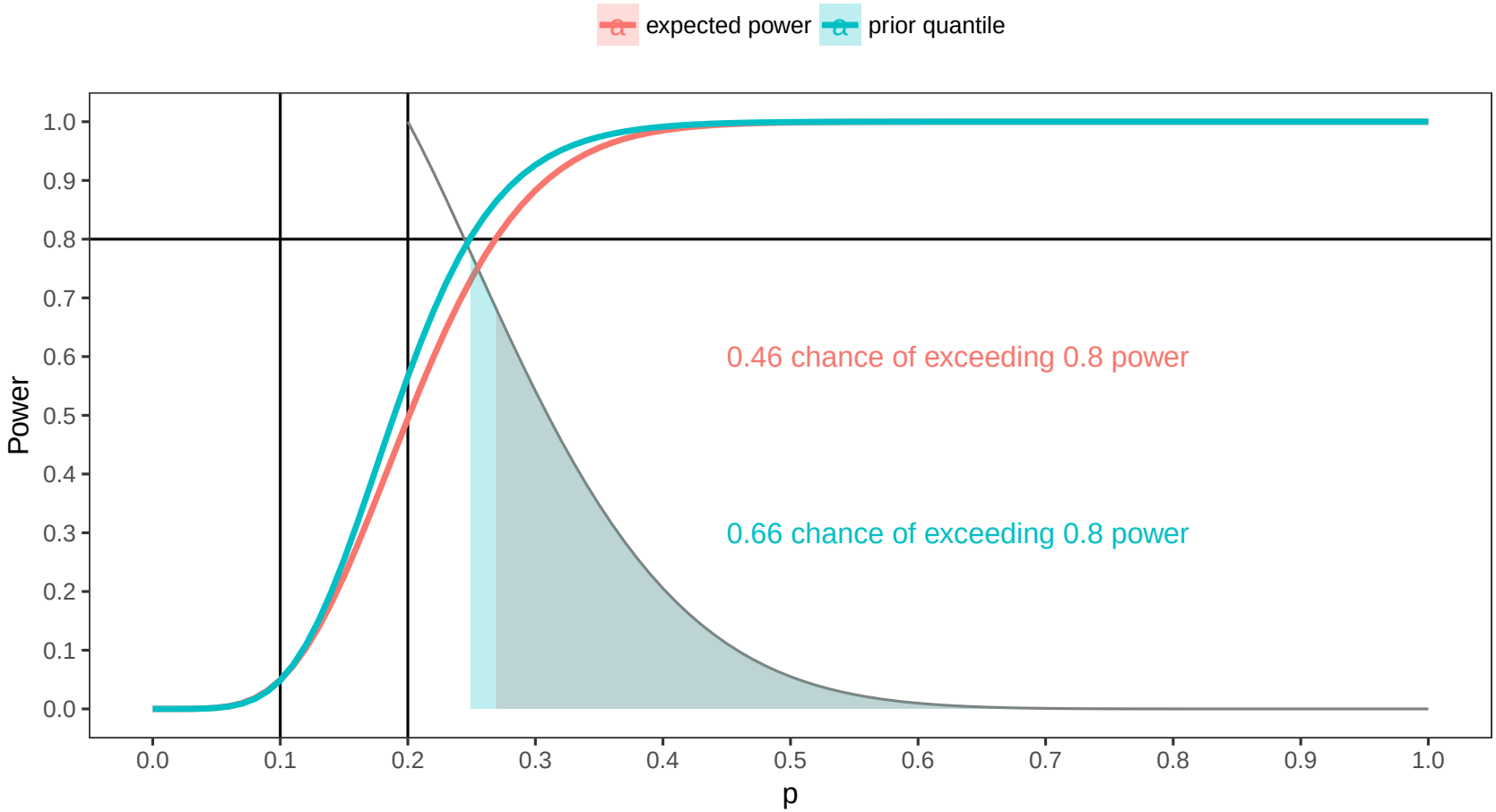


Figure: Comparison of different Bayesian power constraint; using expected power is less strict (smaller design) than prior quantile approach ($\delta = 0.1$, $\eta = 0.66$, target power 80%, expected power $\geq 80\%$).

Let's compare II



Questions?

- Two-stage designs arise natural in situations with high uncertainty
- Require the choice of optimality criterion as there is no 'canonical' one
- Bayesian planning is a principled way to avoid over-fitting to particular values of ρ
- Bayesian designs much less variable (sometimes even constant second stage!)
- Theoretically appealing framework for making planning assumptions explicit
- Not restricted to phase-II

How to actually optimize such designs?

$$\underset{n_1, n(\cdot), c(\cdot)}{\text{argmin:}} \quad \sum_{x_1=0}^{n_1} \mathbf{P}\left[X_1 = x_1\right] n(x_1) \tag{17}$$

$$\text{subject to:} \quad \mathbf{P}_{\rho_0}\left[X > c(X_1)\right] \leq \alpha \tag{18}$$

$$\mathbf{P}_{\rho_1}\left[X > c(X_1)\right] \geq 1 - \beta \tag{19}$$

- Both $n(\cdot)$ and $c(\cdot)$ are discrete functions
- Gradient-based approaches infeasible (integer problem)
- Problem of finding optimal values is NP-hard!
- For relatively small $n_1 << 50$ and $n_2 << 150$, still feasible with dynamic programming

Let others do the heavy lifting...

- Either use custom solution algorithm or leverage existing technology by transforming the problem into an integer linear program
- can the be solved by standard software

Getting to an integer linear problem [4]

- Introduce binary auxiliary variables $y[x_1, n, c] \in \{0, 1\}$
- Define $y[x_1, n, c] = 1$ iff $n(x_1) = n$ and $c(x_1) = c$
- Require $\sum_{n,c} y[x_1, n, c] = 1$, i.e. exactly one 'selection' is active in the solution
- Many functionals can the be written as linear functions in $y[x_1, n, c]$:

$$\mathrm{ESS}(f) = \sum_{x_1,n,c} n \, \mathbf{P}_f\big[X_1 = x_1\big] y[x_1, n, c]$$

$$\mathrm{power}(\rho) = \sum_{x_1,n,c} \mathbf{P}_\rho\big[X_2 > c - x_1\big] \mathbf{P}_\rho\big[X_1 = x_1\big] y[x_1, n, c]$$

Implementation

- Use specialized integer linear programming software like Gurobi or Cbc to solve the problem
- Only need to compute the coefficients for $y[x_1, n_2, c_2]$ for the various constraints - most expensive part, R badly suited for this (slow for loops)
- Implementation in Julia is available at https://github.com/kkmann/BinaryTwoStageDesigns but still experimental

Time for a demo!

Further Comments

Is this ‘truly’ Bayesian?

- Arguable whether methods presented are truly ‘Bayesian’
- Do not invoke Bayes Theorem at all!
- Only use prior as weight function of potential scenarios (sensible even without the Bayesian interpretation!)
- Still, objective functions derived in a Bayesian spirit

Post trial inference

- No time to talk about the (extremely) interesting topic of inference in two-stage designs
- Details on point estimation in [5], in short
  - slightly tweaked Bayesian mean a posteriori estimator substantially improves MSE over maximum likelihood estimator...
  - ... but is not unbiased
  - ... and allows the definition of a sensible p value
- Confidence intervals described in [6]

Beyond binary endpoint & single-arm designs

- Exact methods for discrete outcomes do not generalize well to multiple arms (search space grows exponentially)
- But: Multi-arm trials usually larger, therefore approximations viable (Z-test)
- E.g., test for rate difference δ via stage-wise z-statistics

Beyond binary endpoint & single-arm designs

$$\underset{n_1,\, n(\cdot),\, c(\cdot)}{\text{argmin:}} \quad \int n(z_1)\, \phi(z_1)\, \mathrm{d}\, z_1 \tag{20}$$

$$\text{subject to:} \quad \mathbf{P}_{\delta=0}\big[Z > c(Z_1)\big] \leq \alpha \tag{21}$$

$$\mathbf{P}_{\delta=\delta_1}\big[Z > c(Z_1)\big] \geq 1 - \beta \tag{22}$$

- n and c are now functions in a continuous variable z_1
  $\rightsquigarrow$ variational problem!
- Actively worked on but not ready yet
- Can also be extended to incorporate prior uncertainty

Thank you!

Slides and code available at
https://github.com/kkmann/Sommerschule2018

References I

- Simon R. Optimal two-stage designs for phase II clinical trials. *Controlled Clinical Trials* 1989; **10**(1):1–10.
- Mander, AP, Thompson, SG. Two-stage designs optimal under the alternative hypothesis for phase II cancer clinical trials. Contemporary Clinical Trials 2010; **31**(6):572–578.
- Grayling, MJ, Mander, A P. Do single-arm trials have a role in drug development plans incorporating randomised trials?. Pharmaceutical statistics 2016; **15**(2):143–151.
- Kunzmann K, Kieser M. Optimal adaptive two-stage designs for single-arm trials with binary endpoint. *arXive.org* 2016.

References II

- Kunzmann, K, Kieser, M. Test-compatible confidence intervals for adaptive twoâĂŘstage single-arm designs with binary endpoint. Biometrical Journal 2018; **60**(1):196–206.
- Kunzmann, K, Kieser, M. Point estimation and pâĂŘvalues in phase II adaptive two-stage designs with a binary endpoint. Statistics in medicine 2017; **36**(6):971–984.