Choosing an efficient multi-arm multi-stage clinical trial design

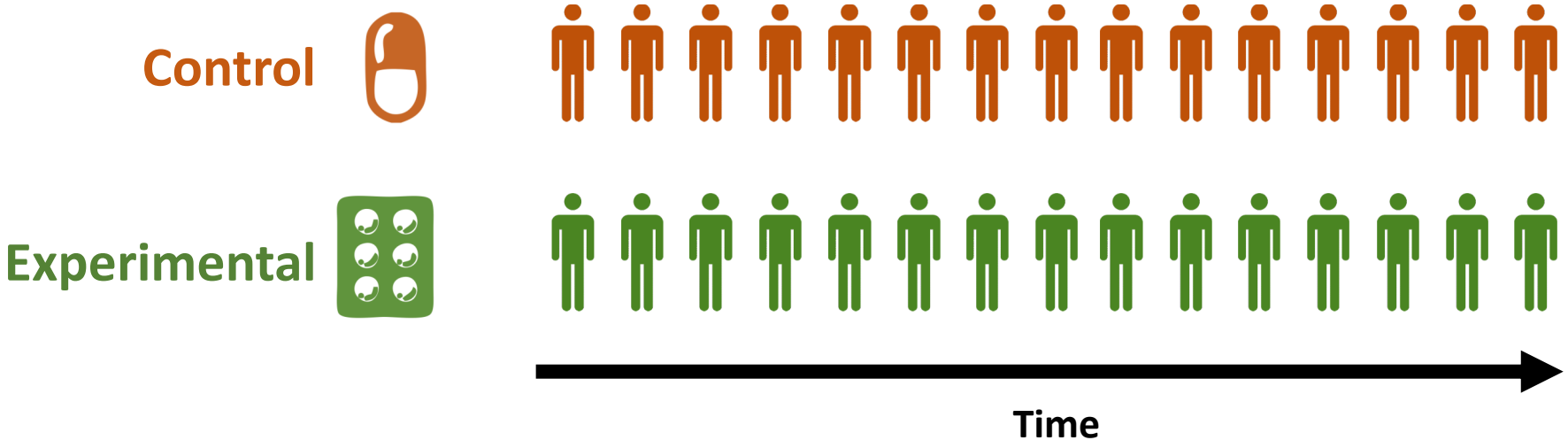
Michael Grayling

Population Health Sciences Institute Biostatistics Research Group, Newcastle University

www.newcastle-biostatistics.com

Conventional randomized controlled trials

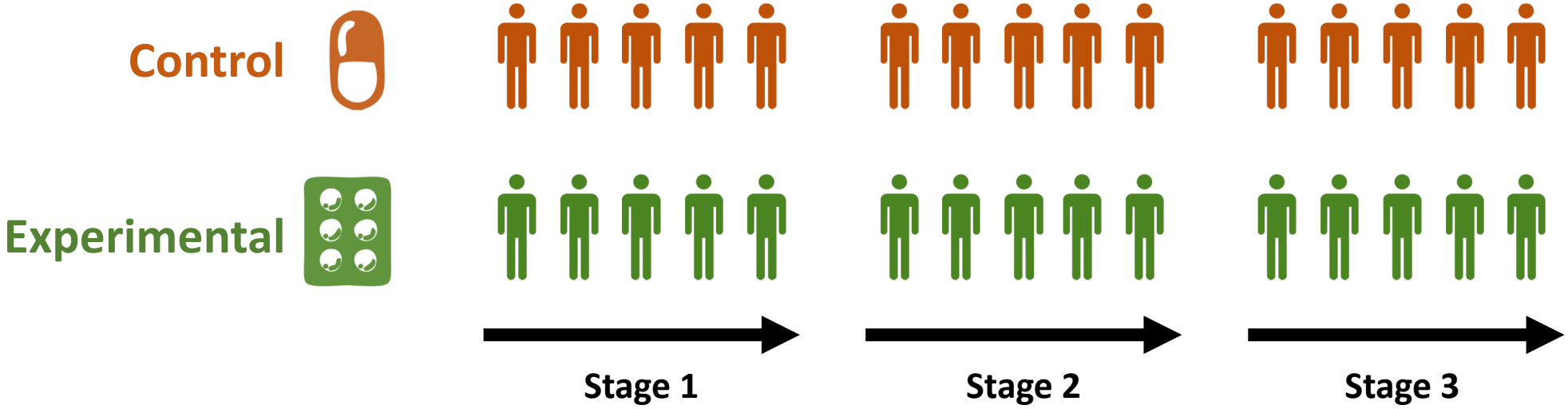
- Suppose we have a new experimental treatment and we want to determine *whether it provides a benefit* over the current standard-of-care
- Patients are randomly allocated to one of the treatments and their outcome data is compared



- Trials are **very** expensive…is this the best we can do? Can we make evaluation more efficient?

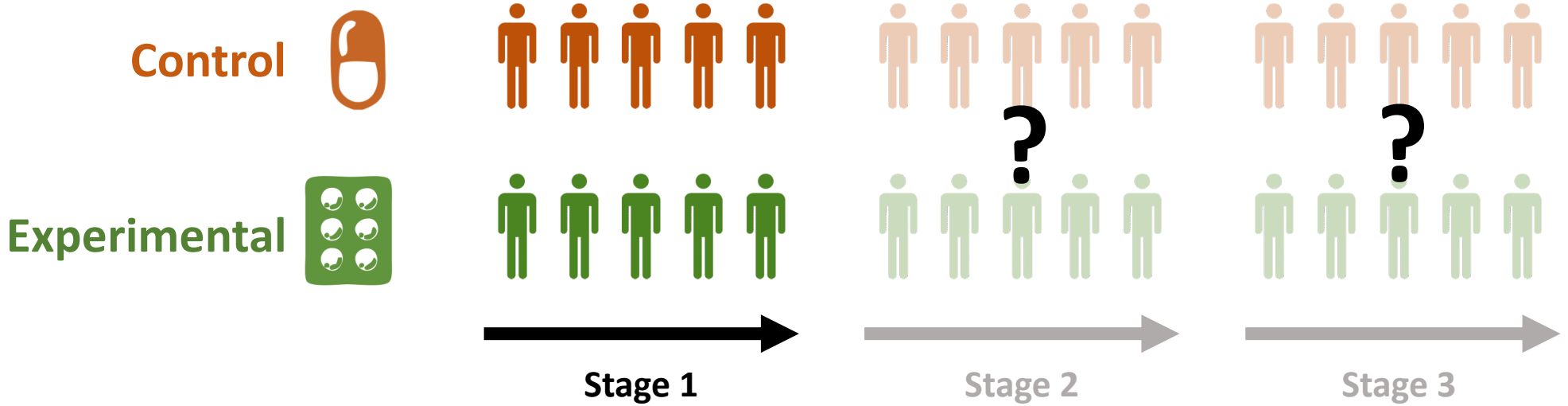
Group sequential trials

- Recruitment and outcome data collection **does not happen instantaneously** in an RCT
- We can potentially exploit this by including a series of *interim analyses* at which the trial may stop
- Reduces the expected required sample size compared to only analysing the data at the end of the trial



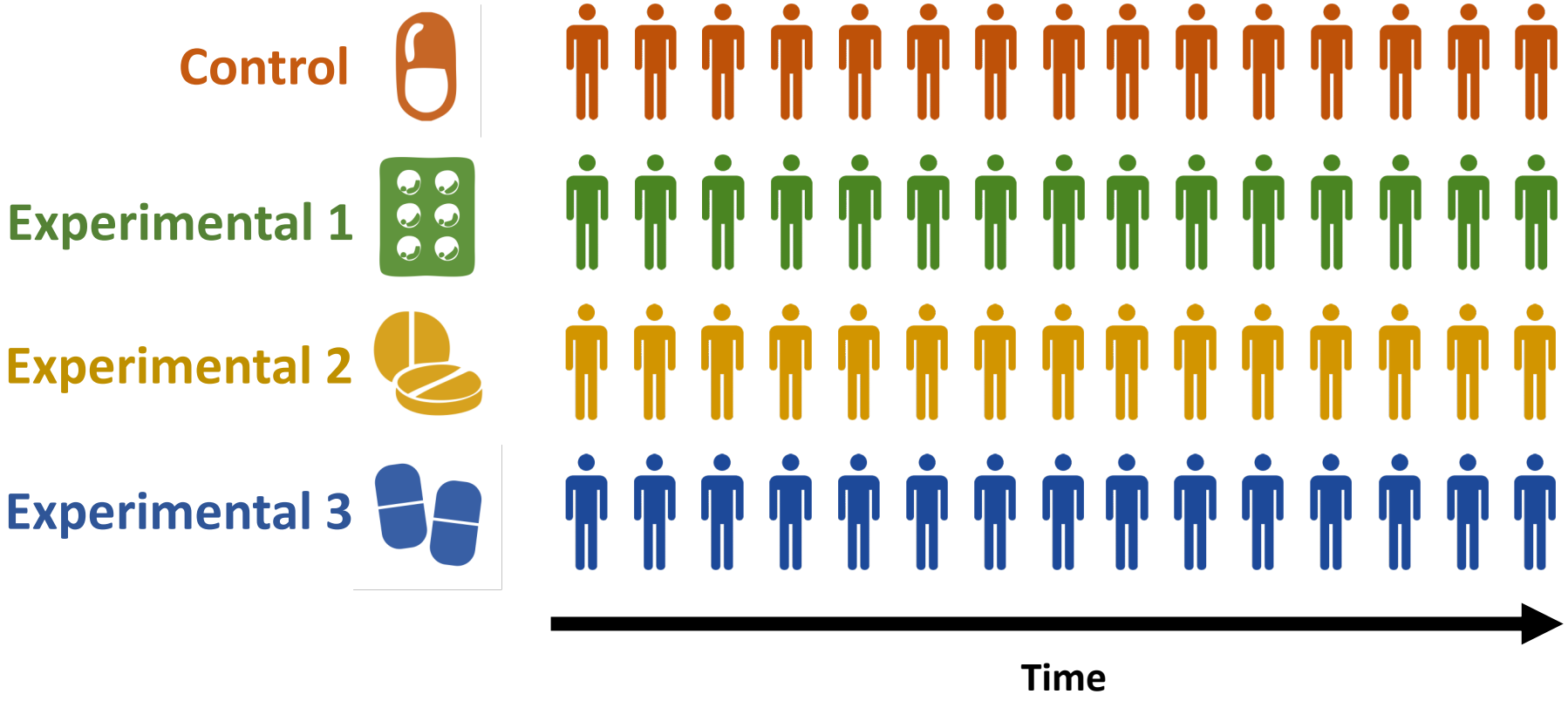
Group sequential trials

- Recruitment and outcome data collection **does not happen instantaneously** in an RCT
- We can potentially exploit this by including a series of *interim analyses* at which the trial may stop
- Reduces the expected required sample size compared to only analysing the data at the end of the trial



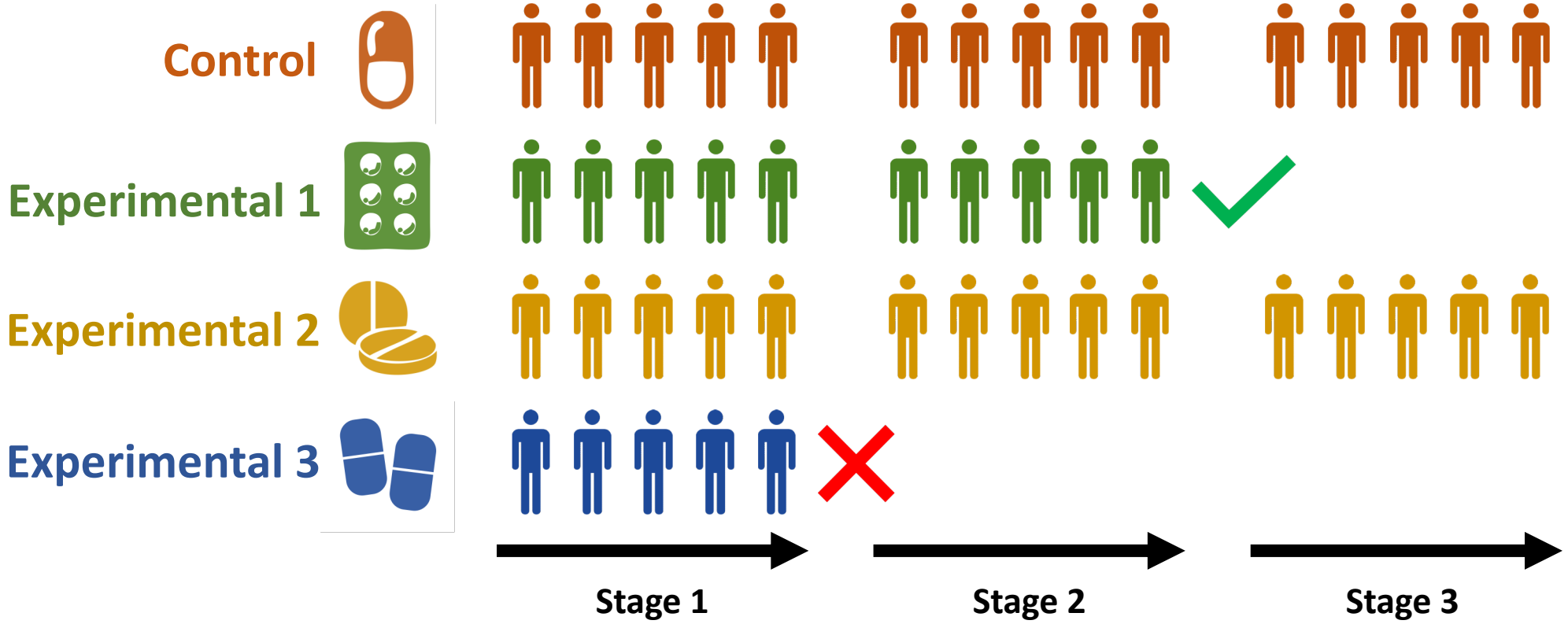
Multi-arm trials

- Compare several experimental treatments to a **shared control group**
- Requires *fewer patients* in total than doing a series of two-arm trials



Multi-arm multi-stage (MAMS) trials

- Include interim analyses in a multi-arm trial
- Can be a *highly efficient* approach to evaluating multiple experimental treatments



Many varieties of MAMS design now available

- Different types of outcome data
  - E.g., continuous, binary, survival[1]
  - Covariates[1]
  - Changing outcomes – "intermediate" outcome available at interim analyses[2,3]
- 'Separate' and 'simultaneous' stopping[4]
  - Do you terminate the whole trial as soon as one experimental treatment is found to be efficacious?
- Bayesian designs[5,6]
  - Particularly useful for inputting external information. E.g., COVID trials
- Sample size re-estimation[7]
  - Helps handle scenarios in which there is limited information available to help power the trial accurately
- Several varieties that are about targeting improved statistical efficiency in terms of either
  - Benefit to patients in the trial
  - **The required sample size/power**

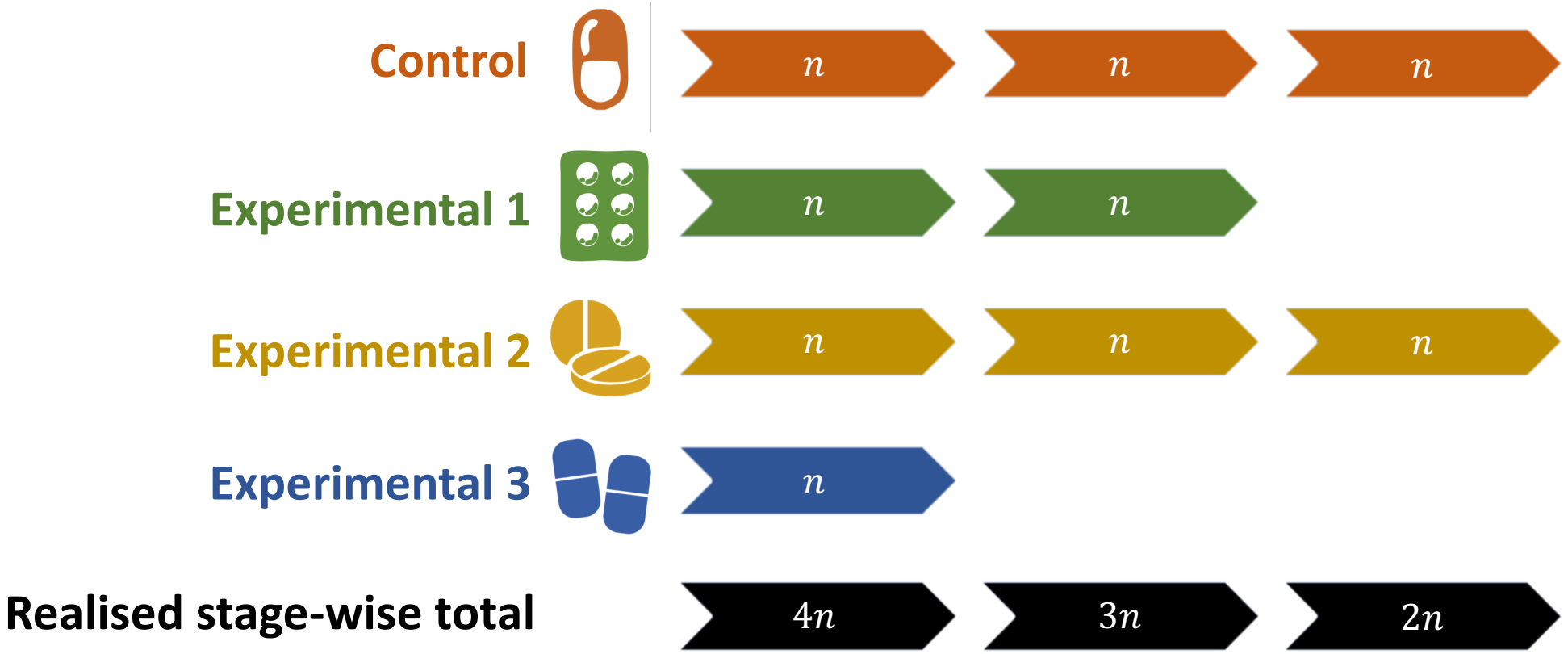
Sample size efficient designs

- Code available on `ssc` a few years ago now for this type of design
  - Adaptive design course run with Adrian Mander, David Robertson, and James Wason
- Provides a lot of flexibility in terms of the stopping rules
- *Little* flexibility in terms of the sample size per-stage
- Discuss **fixed vs. variable stage-wise sample size**
  - Relates to practical considerations in some recent trials

Variable stage-wise sample size

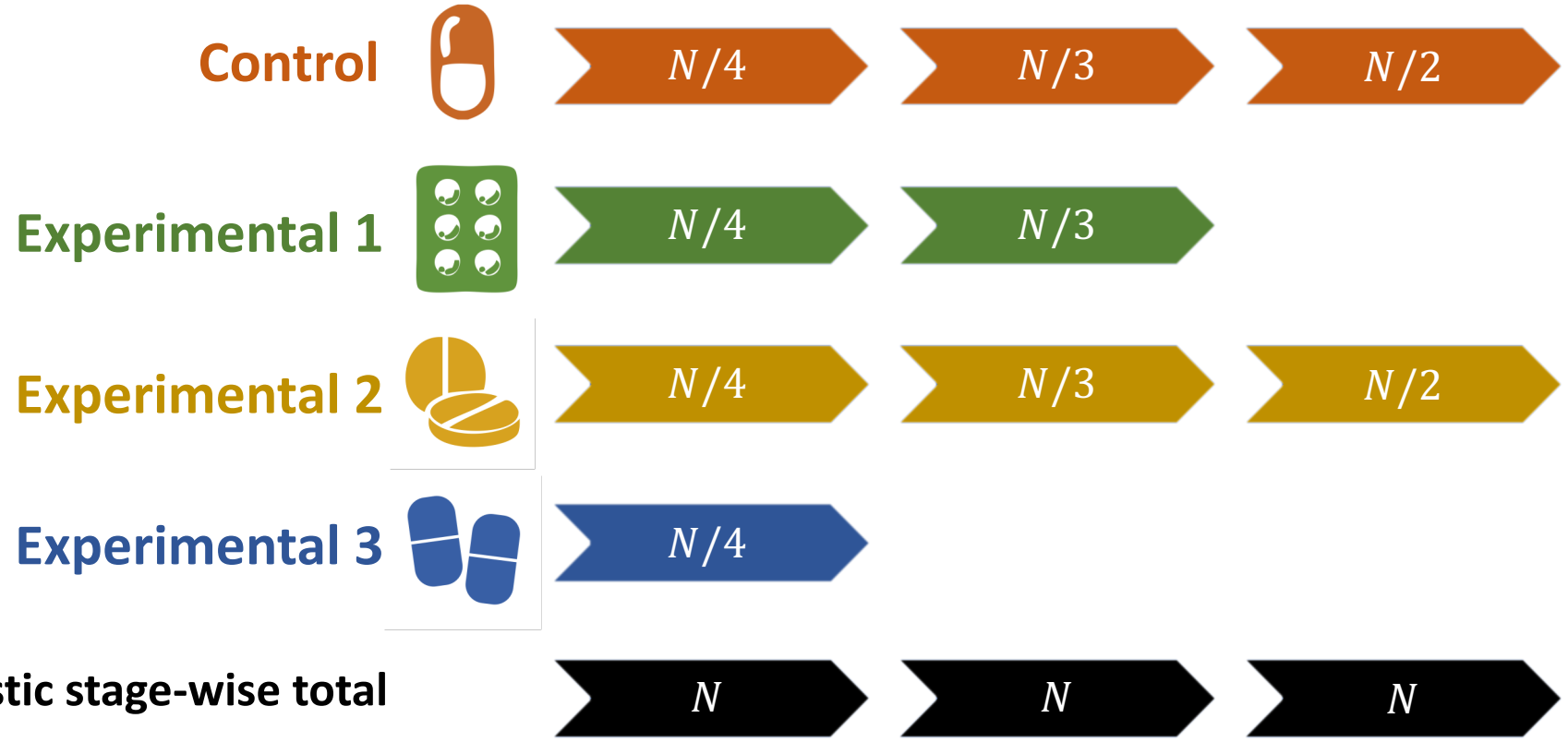
- Majority of MAMS literature assumes that a particular number of patients will be enrolled to an arm *if it is present* in the trial[8,9]
- Exact **stage-wise sample size is variable**
- E.g., 3 experimental arms and 3 stages allowed: 8 possible sample sizes
- Can be problematic in terms of costing the trial, knowing when the interim analyses will occur, knowing whether recruitment is going well

Variable stage-wise sample size



Fixed stage-wise sample size

- **Share a fixed sample size** between the arms present in each stage
- Limit number of possible sample sizes



Remainder of talk

- Discuss some of the *key* statistical details behind the MAMS approach
- Go through an example: what are the **advantages/disadvantages** of fixing the stage-wise sample size compared to the more conventional approach?
- Overview and discussion of Stata implementation

Design framework

- Suppose there are K experimental arms, and we allow at most J stages
- Test the following hypotheses, through the series of analyses:

$$H_k : \tau_k \leq 0, \qquad k = 1, \ldots, K$$

- τ_k represents the effect of experimental treatment k relative to the control
- Use the following test statistics at stage j to test H_k:

$$Z_{jk} = \frac{\hat{\tau}_{jk}}{\sqrt{\mathrm{Var}\left(\hat{\tau}_{jk}\right)}}$$

Design framework

- Need to specify lower (futility) and upper (efficacy) stopping boundaries: $\boldsymbol{f} = (f_1, \ldots, f_J)$ and $\boldsymbol{e} = (e_1, \ldots, e_J)$
- E.g., decision rules
  - If $Z_{jk} > e_j$ then terminate the trial, rejecting H_k. Else:
  - Drop the k with $Z_{jk} \leq f_j$ for futility
  - If one has $f_j < Z_{jk} \leq e_j$, continue to next stage retaining those not dropped + control arm
- Control the *familywise error-rate* to level α when $\tau_1 = \cdots = \tau_K = 0$
  - Probability of at least one type-I error
- Power of $1 - \beta$ to reject H_1 when:

$$\tau_1 = \delta_1, \tau_2 = \cdots = \tau_K = \delta_0$$

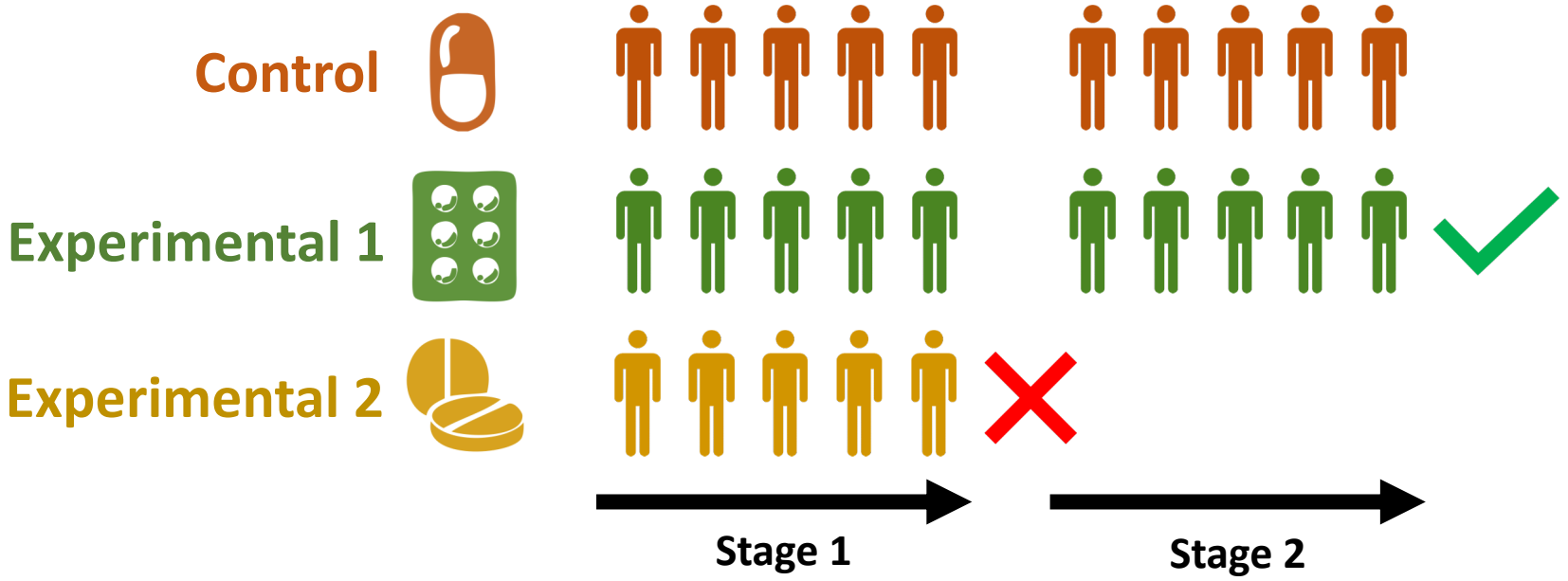
where δ_1 and δ_0 are *interesting* and *uninteresting* effects

Design determination

- Need to find suitable $\boldsymbol{e}$, $\boldsymbol{f}$, and required sample size
  - If allowing the stage-wise sample size to vary, search for n: sample size for each present arm in each stage
  - If fixing the stage-wise sample size, search for N: the total stage-wise sample size
- In practice the stopping boundaries are usually assumed to follow a simple functional form
  - E.g., Pocock boundaries: $e_1 = \cdots = e_J = f_J = C$, $f_1 = \cdots = f_{J-1} = -C$
- To find an efficient design, need to be able to evaluate statistical quantities of interest. In particular, for choices of C and n/N we would like to be able for any values of $\tau_1, \ldots, \tau_K$ to compute:
  - Expected, standard deviation, median, and modal sample sizes
  - Probability each null hypothesis is rejected
- Use fact that *joint distribution* of the test statistics is *multivariate normal*

Design determination

- For example, suppose that $J = K = 2$ and you want to know the probability that H_1 is rejected at stage 2, and experimental drug 2 is dropped for futility at stage 1



$$\int_{f_1}^{e_1} \int_{e_2}^{\infty} \int_{-\infty}^{f_1} \phi\{(z_{11}, z_{21}, z_{12}), \mathbb{E}(Z_{11}, Z_{21}, Z_{12}), \mathrm{Cov}(Z_{11}, Z_{21}, Z_{12})\} \, \mathrm{d}z_{12} \mathrm{d}z_{21} \mathrm{d}z_{11}$$

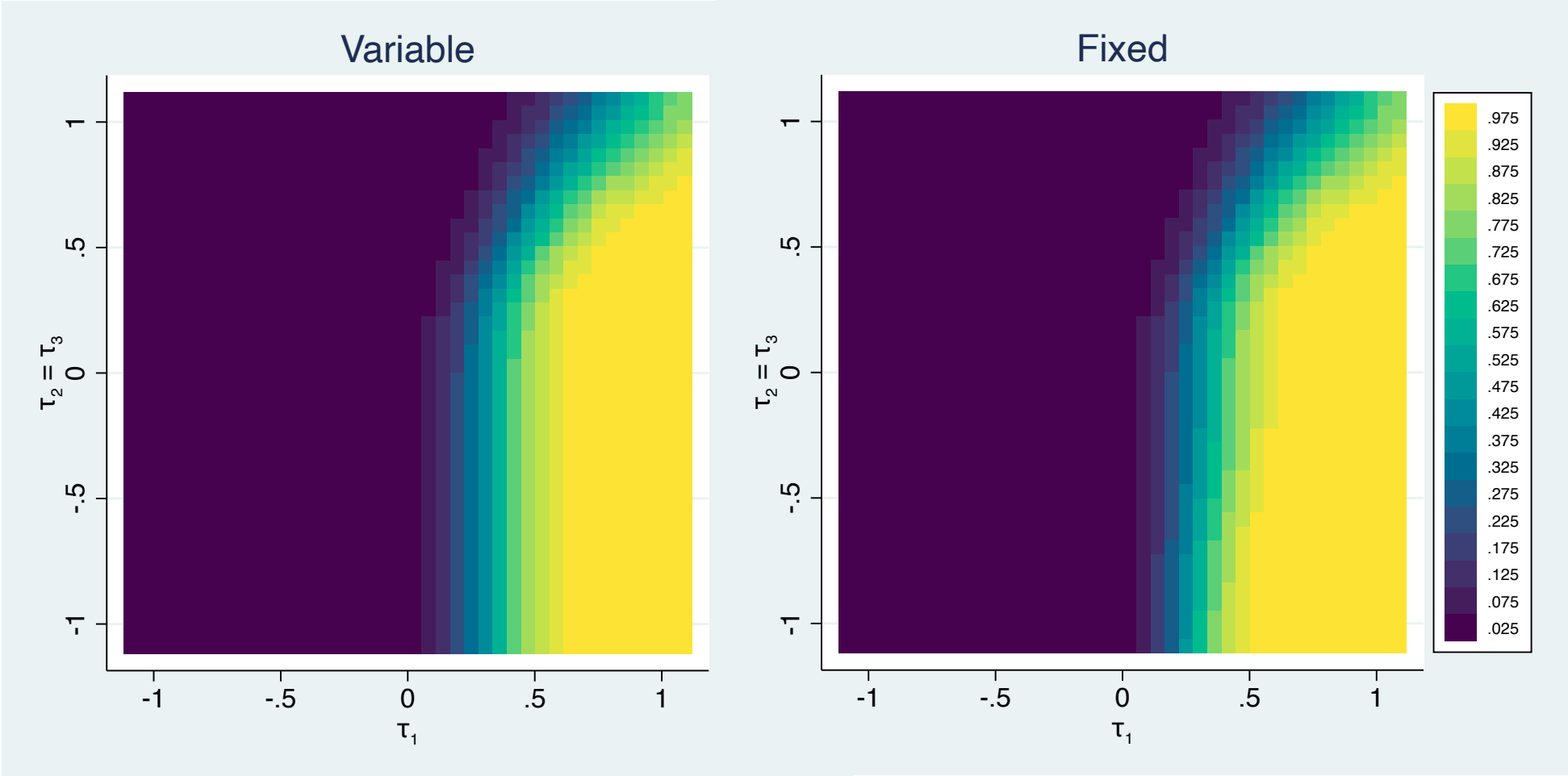
Design determination

- Two one-dimensional optimization steps: find C and then find n/N
- Speed therefore dependent on how fast you can
  - Evaluate multivariate normal integrals
  - Perform a one-dimensional search
- More on this later

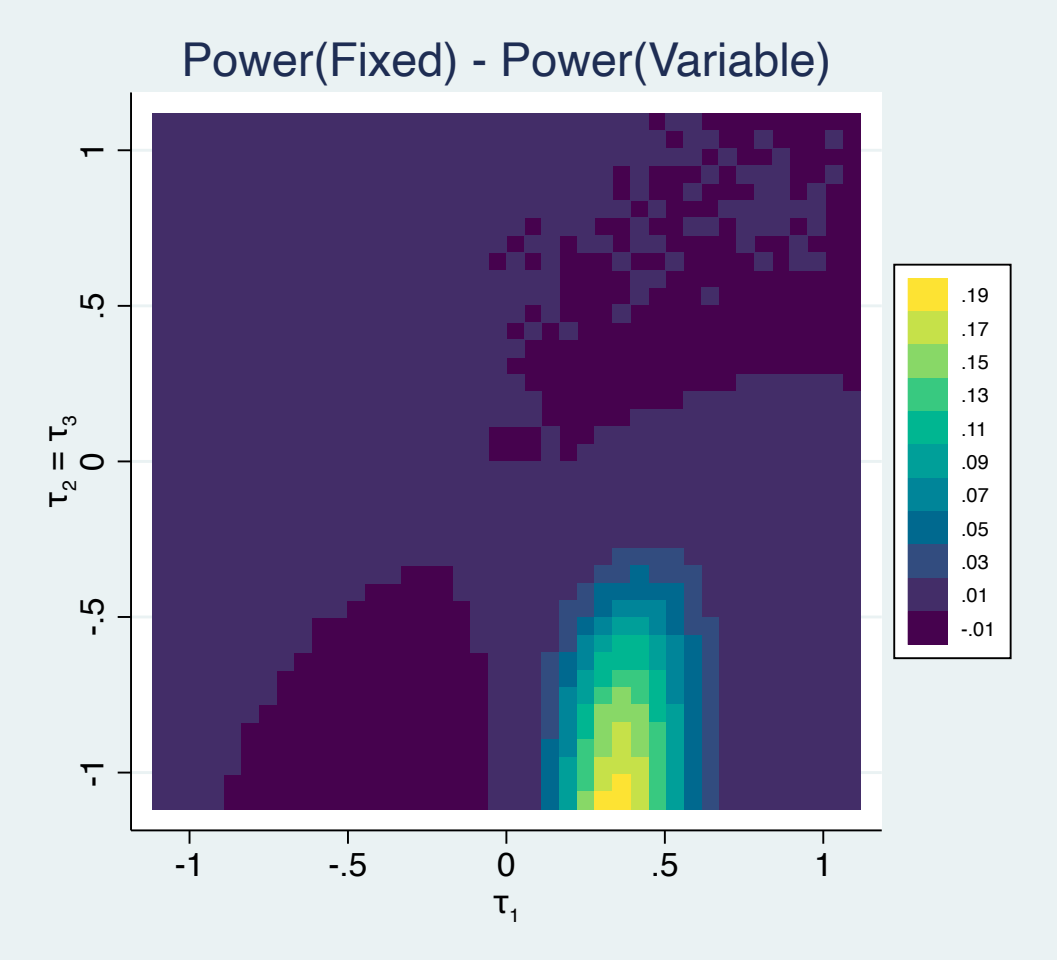
Example: TAILoR trial

- Trial assessing drugs for reducing insulin resistance in HIV-positive individuals on combination antiretroviral therapy[9]
- Use $K = 3$ and $J = 3$
- Also $\alpha = 0.05, \beta = 0.1\ , \delta_1 = 0.545, \delta_0 = 0.178, \sigma = 1, r = 1$, and O'Brien-Fleming stopping boundaries
  - Results not very sensitive to these choices
- Conventional MAMS design with a variable stage-wise sample size would need ~26 patients per-arm per-stage. 8 possible sample sizes
  - 105, 157, 183, 209, 235, 262, 288, 314
- Fixing the stage-wise sample size means you need 105 patients per-stage
  - 105, 209, 314
- By construction these are very similar!
  - Need to delve deeper to spot the differences

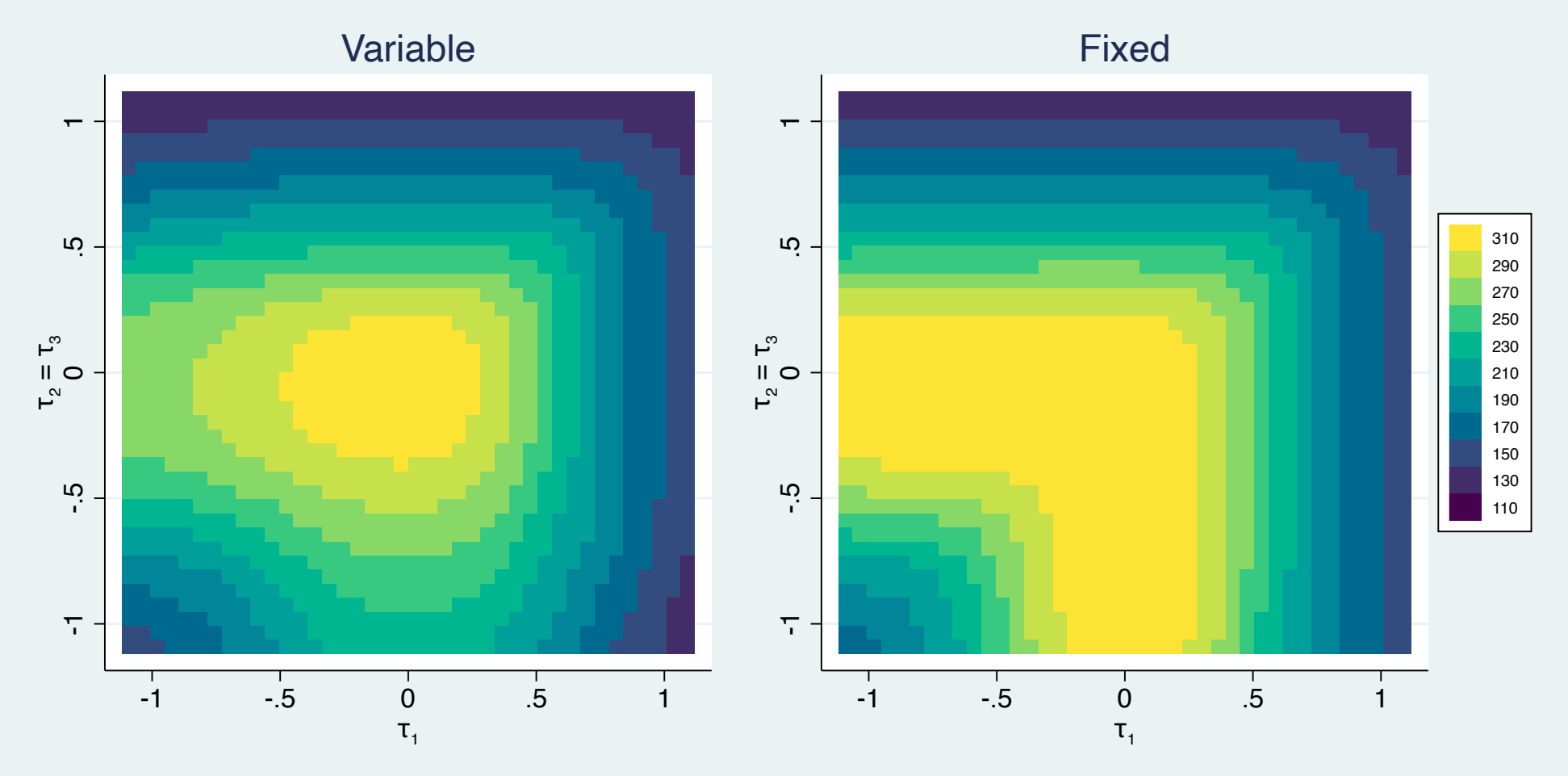
Probability we reject H_1



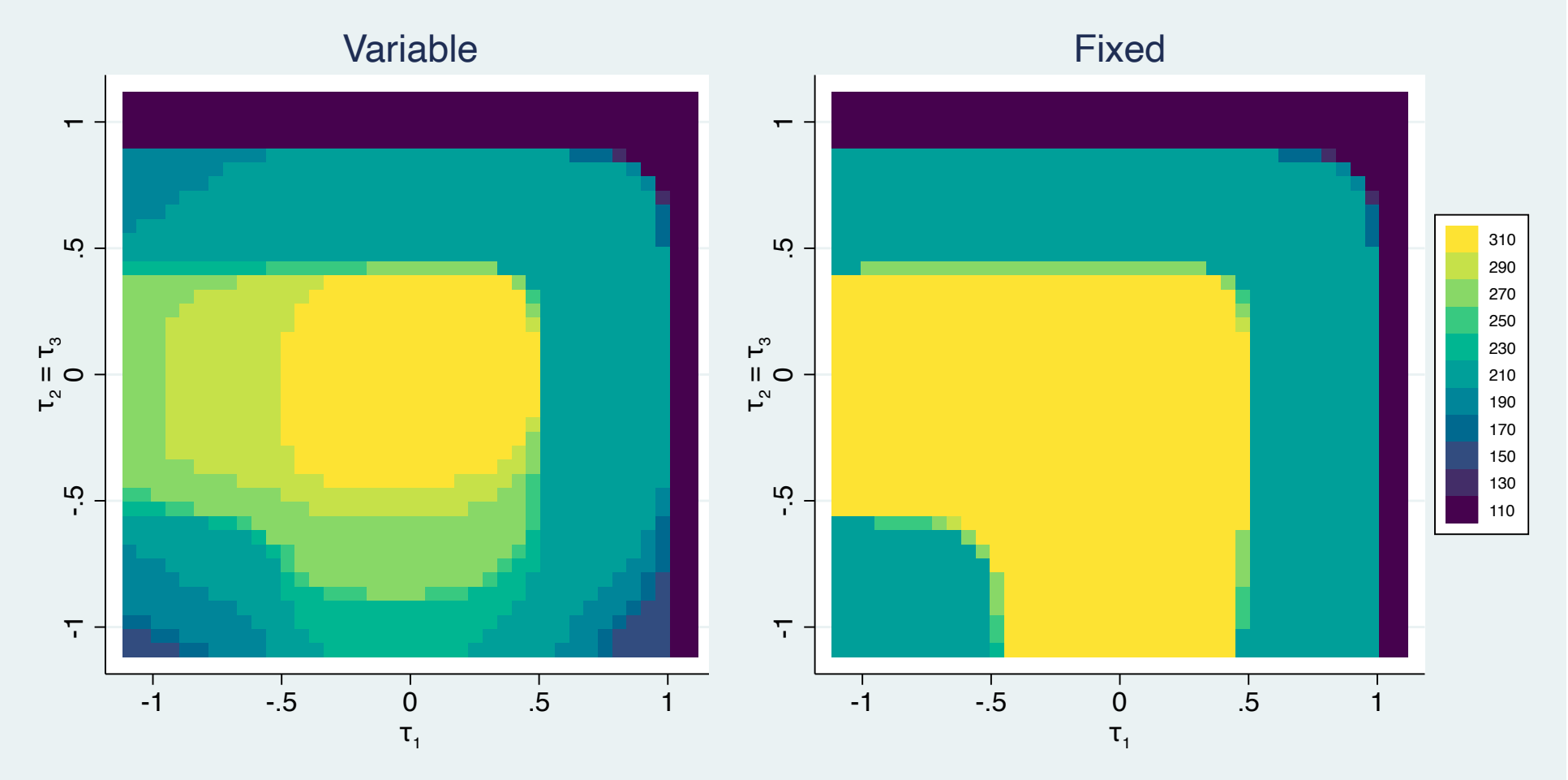
Probability we reject H_1



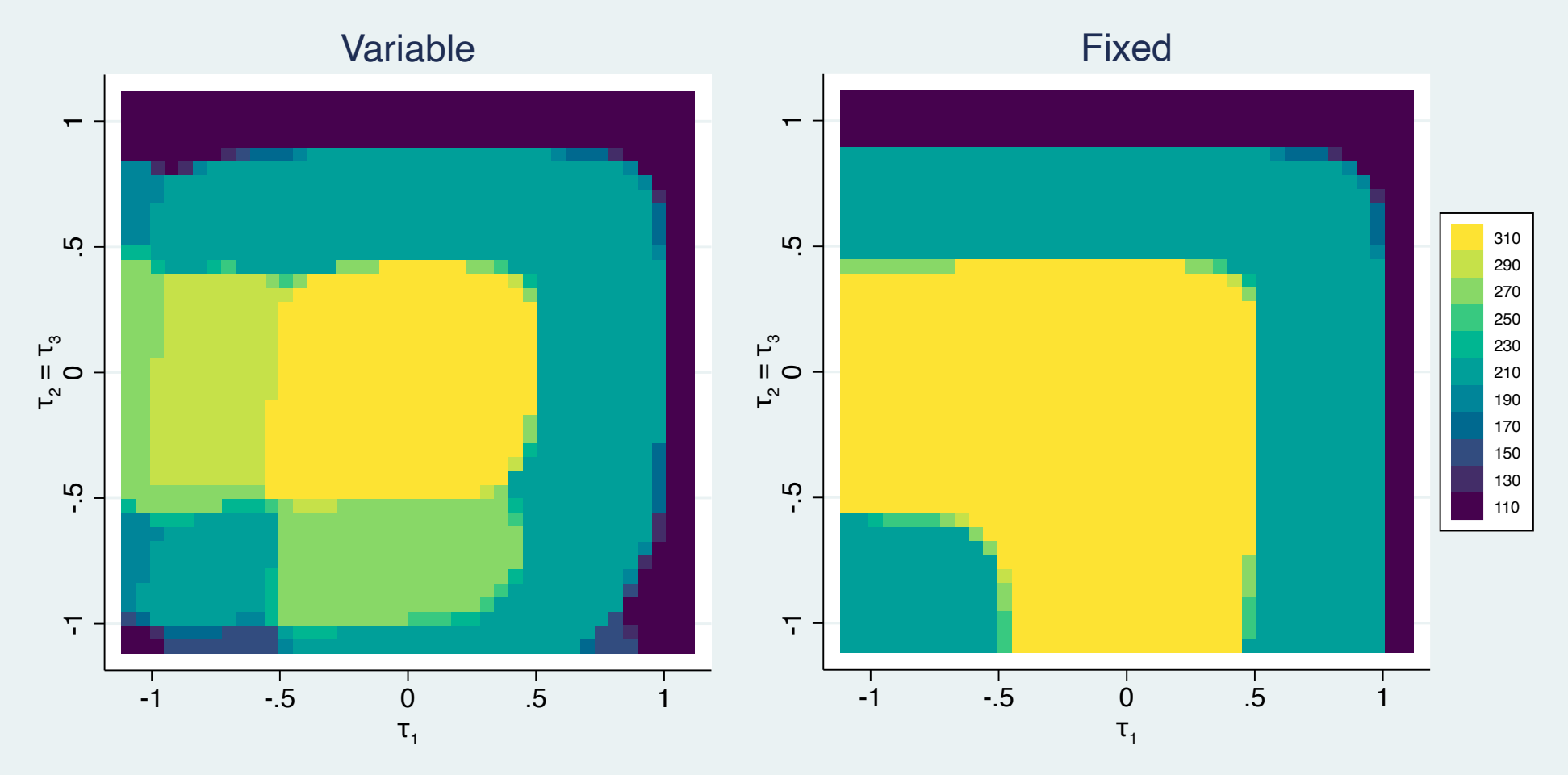
Expected sample size



Median sample size



Modal sample size



Discussion

- In the end, easy to summarise
  - All things being equal in terms of how error-rates etc. are controlled, performance of the two approaches often similar: in many cases you might not expect to see a difference
  - When they do differ, it is a question of what do you want to do more, minimize the sample size (variable) or maximise the power (fixed)

- Other considerations for the fixed approach
  - **Advantage:** Under (roughly) known recruitment rate, easier to predict timing of interim analysis
  - **Disadvantage:** Potentially more patients on the control arm

Stata implementation

```
des_mams, k(integer 3) j(integer 2) ALPha(real 0.05) beta(real 0.2)
          DELta1(real 0.5) delta0(real 0) sd(real 1) RATio(real 1)
          FSHape(string) ESHape(string) ffix(real 0) efix(real 2)
          SEParate FIXed
```

- Set-up similarly to `power`
  - What you need and nothing more!
- `rclass`; returns the required sample size, the stopping boundaries, and prints a summary of the key operating characteristics
- Internally `des_mams` is broken down into modules and written in a very general way
  - Still know it's difficult to know it works correctly → limited results/software to compare to
  - So make `sim_mams` available as an internal check on results
  - Working on relating results to those from `nstage` where possible[10]

Stata implementation

```
des_mams, k(integer 3) j(integer 2) ALPha(real 0.05) beta(real 0.2)
          DELta1(real 0.5) delta0(real 0) sd(real 1) RATio(real 1)
          FSHape(string) ESHape(string) ffix(real 0) efix(real 2)
          SEParate FIXed
```

- Set-up similarly to `power`
  - What you need and nothing more!
- `rclass`; returns the required sample size, the stopping boundaries, and prints a summary of the key operating characteristics
- Internally `des_mams` is broken down into modules and written in a very general way
  - Still know it's difficult to know it works correctly → limited results/software to compare to
  - So make `sim_mams` available as an internal check on results
  - Working on relating results to those from `nstage` where possible[10]

Stata implementation

```
des_mams, k(integer 3) j(integer 2) ALPha(real 0.05) beta(real 0.2)
          DELta1(real 0.5) delta0(real 0) sd(real 1) RATio(real 1)
          FSHape(string) ESHape(string) ffix(real 0) efix(real 2)
          SEParate FIXed
```

- Set-up similarly to `power`
  - What you need and nothing more!
- `rclass`; returns the required sample size, the stopping boundaries, and prints a summary of the key operating characteristics
- Internally `des_mams` is broken down into modules and written in a very general way
  - Still know it's difficult to know it works correctly → limited results/software to compare to
  - So make `sim_mams` available as an internal check on results
  - Working on relating results to those from `nstage` where possible[10]

Stata implementation

```
des_mams, k(integer 3) j(integer 2) ALPha(real 0.05) beta(real 0.2)
          DELta1(real 0.5) delta0(real 0) sd(real 1) RATio(real 1)
          FSHape(string) ESHape(string) ffix(real 0) efix(real 2)
          SEParate FIXed
```

- Set-up similarly to `power`
  - What you need and nothing more!
- `rclass`; returns the required sample size, the stopping boundaries, and prints a summary of the key operating characteristics
- Internally `des_mams` is broken down into modules and written in a very general way
  - Still know it's difficult to know it works correctly → limited results/software to compare to
  - So make `sim_mams` available as an internal check on results
  - Working on relating results to those from `nstage` where possible[10]

Stata implementation

```
. des_mams, j(3) beta(0.1) delta(0.545) delta0(0.178) fshape("obf") eshape("obf")
-------------------------------------------------
3-stage 3-experimental treatment MAMS trial design
-------------------------------------------------
The hypotheses to be tested will be:

   Hk: τk = μk - μ0 ≤ 0, k = 1, 2, 3
…
Futility stopping boundaries, f, determined to be: (-3.64, -2.57, 2.1)
Efficacy stopping boundaries, e, determined to be: (3.64, 2.57, 2.1)
…
Required stage-wise group size in the control arm, n, determined to be: 27

The operating characteristics of the design are:

   r()       | Variable
   --------    --------
   P_HG      | P_HA(HG)  = .05,
   P_LFC     | P_H1(LFC) = .909,
   ESS_HG    | ESS(HG)   = 322.13,
   ESS_LFC   | ESS(LFC)  = 252.59,
   …
```

Stata implementation

- Algorithms for evaluating the performance of a candidate design have improved a lot[11]
- Still slow for (reasonably) large J and K
- Multivariate normal integrals done with an updated version of code from Grayling and Mander (2018)[12]
  - Similar in speed to `mvnormalcv()`
- Result in a key sub-routine called `power_mams(n,…)`

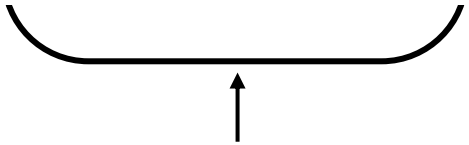
Stata implementation

- Current: One-dimensional root-solving done with our own implementation of Brent's algorithm

```
power_mams(n,…) – (1 - beta)
```

- Started with `optimize()`, re-framing as a minimization problem → convergence unreliable

```
(power_mams(n,…) – (1 - beta))^2
```



True required n

- Then moved to a `while` loop → too slow
- Then `mm_root()` → nearly there…own code allows us to strip out anything not needed

```
ssc install desma
```

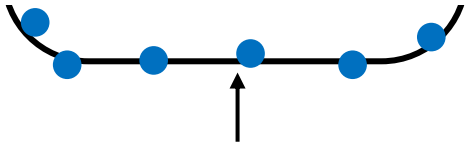
Stata implementation

- Current: One-dimensional root-solving done with our own implementation of Brent's algorithm

```
power_mams(n,…) – (1 - beta)
```

- Started with `optimize()`, re-framing as a minimization problem → convergence unreliable

```
(power_mams(n,…) – (1 - beta))^2
```



- Then moved to a `while` loop → too slow
- Then `mm_root()` → nearly there…own code allows us to strip out anything not needed

```
ssc install desma
```

References

1. Jaki T, Magirr D. Considerations on covariates and endpoints in multi-arm multi-stage clinical trials selecting all promising treatments. *Stat Med* 2013;**32:**1150-63

2. Royston P et al. Designs for clinical trials with time-to-event outcomes based on stopping guidelines for lack of benefit. *Trials* 2011;**12:**81

3. Bratton DJ et al. A multi-arm multi-stage clinical trial design for binary outcomes with application to tuberculosis. *BMC Med Res Methodol* 2013;**13:**139

4. Urach S, Posch M. Multi-arm group sequential designs with a simultaneous stopping rule. *Stat Med* 2016;**35:**5536-50

5. Jacob L *et al*. Evaluation of a multi-arm multi-stage Bayesian design for phase II drug selection trials – An example in hemato-oncology. BMC Med Res Methodol 2016;16:67

6. Ryan EG *et al*. Bayesian adaptive designs for multi-arm trials: An orthopaedic case study. *Trials* 2020;**21:**83

7. Li Y *et al*. Sample size re-estimation for confirmatory two-stage flexible multi-arm trial with normal outcomes. *J Stat Comp Sim* 2020;**90:**157-78

8. Wason JMS, Jaki T. Optimal design of multi-arm multi-stage trials. *Stat Med* 2012;**31:**4269-79

9. Magirr D *et al*. A generalized Dunnett test for multi-arm multi-stage clinical studies with treatment selection. *Biometrika* 2012;**99:**494-501

10. Barthel FMS *et al*. A menu-driven facility for sample-size calculation in novel multiarm, multistage randomized controlled trials with a time-to-event outcome. *Stata J* 2009;**9:**505-23

11. Grayling MJ *et al*. Efficient determination of optimised multi-arm multistage experimental designs with control of generalised error-rates. *arXiv* 2018;1712.00229

12. Grayling MJ, Mander AP. Calculations involving the multivariate normal and multivariate t distributions with and without truncation. *Stata J* 2018;**18:**826-43