Introduction to Designing Adaptive Trials: Seamless 2/3

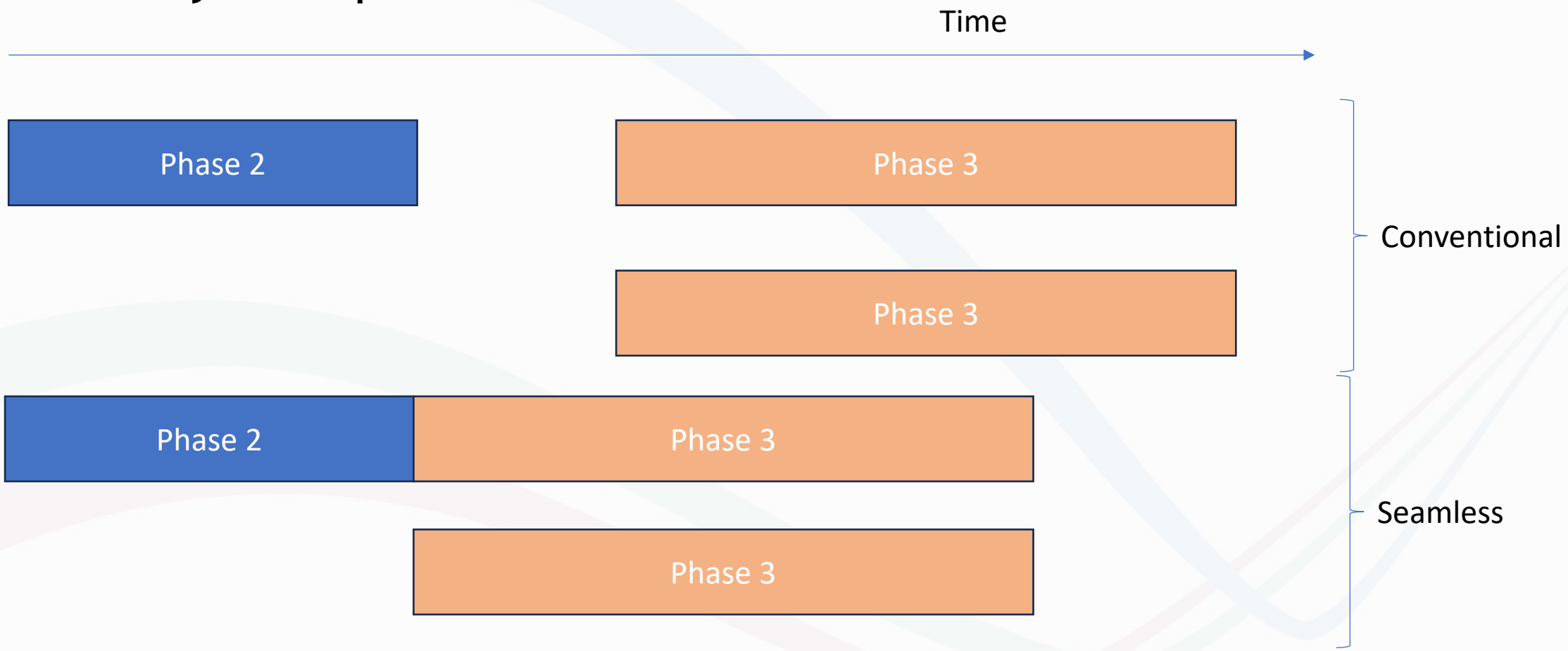
Berry Consultants
Statistical Innovation

Outline

- The Basics
- FACTS implementation
- Advanced Topics
- Regulatory

Seamless phase 2/3

- Essentially a simple idea:



Advantages

1. Save time
2. Possibly save Phase 3 sample size, if we can include Phase 2 data in the Phase 3 analysis

Complications

- Initial regulator reaction – “you call it deadtime, we call it thinking time”. Regulator has to approve running Phase 3 on the basis of the design, not seeing Phase 2 data.
  - How is trial integrity protected (use independent/firewalled team to produce interim analysis report)
  - The basis for dose selection
- Continue enrolment after phase 2 full enrolled, or pause until phase 2 fully followed up?
  - Use early endpoint in Phase 2?
  - Predict primary endpoint based on early visits?
  - If enrolment paused, can we re-start it, might we lose some sites?
- Use Phase 2 data in Phase 3 “inferentially seamless” or just “operationally seamless”.
- If using Phase 2 data, which data from Phase 2 to use in Phase 3?
  - All, just arms in Phase 3, pool all data on treatment arms in phase 2 onto Phase 3 arm, just control data, just incomplete subjects?

Complications #2

- How to select which arm(s) from Phase 2 to include in Phase 3
- How to select how many arms from Phase 2 to include in Phase 3
- How to size Phase 3.

Justification

- For the company the time saving can be very significant – 6 months to a year.
- To justify to regulators:
  - Earlier availability of new treatment (good if high unmet medical need)
  - V large sample sizes required for phase 3 – reduce by also using phase 2 data
  - V rare disease, difficult to recruit enough for phase 3 – so reduce by also using phase 2 data
  - V well understood disease and endpoints, little risk of “unknown unknowns” arising.

Example:

- Phase 2
  - 3 arms vs SoC
  - Continuous Endpoint, after 6 months, SD of endpoint ~2 points
  - Expected improvement on SoC: 1pt, Target improvement on Trt: 2pts
  - Accept 20% false positive (6.66% per arm) and require 90% power – requires 63 per arm .

- Phase 3
  - 1 arm vs Soc
  - 1pt treatment effect 0.025 single-sided alpha, 90% power, 85 per arm

Scenarios

- We will consider 4 scenarios, SD of endpoint 2pts in all scenarios. Mean change from baseline:

	Control	Dose 1	Dose 2	Dose 3
Null	1.00	1.00	1.00	1.00
Arm 1 only	1.00	2.00	1.00	1.00
Arm 2 peak	1.00	1.50	2.00	1.50
Arm 3 linear	1.00	1.33	1.67	2.00

Designs

- We'll compare designs that:
  - Don't use any phase 2 date in phase 3
  - Use all the phase 2 data from the arms continued into phase 3
  - Pool all the data on the different treatments arms in phase 2 and include it on the treatment arm in phase 3 (and use the phase 2 control arm data)
- For now we'll assume a very quick endpoint and ignore 'overrunning' issues. (i.e. all phase 2 participants complete at phase 2 analysis).

Enter into FACTS

- Create a FACTS core design of Phase 2
  - Enrolment is 6 per week
  - Allocation is 1:1:1:1
  - Select the arm with the greatest probability of having the maximum response
  - Success is p-value of this arm < 0.066 (α = 0.2/3)
- Create a FACTS staged design using the core design for the first stage
  - Fail to ‘graduate’ if p-value is > 0.066
  - Allocation is 1:1
  - Success in phase 3 if p-value < 0.025

Results

- Look at success
- Look at probability of correctly selecting the best arm

Comparing results

- Use only phase 3 data has an overall (phase 2 & phase 3) success rate in the Null scenario of 0.0039 (10,000 sims)
  - Theoretically the overall type 1 error should be 0.2 * 0.025 = 0.005, 0.0039 is within simulation error (~1.5SDs).
  - Power in the other scenarios is 0.81, 0.77 and 0.78
- Using phase 2 data, pooling all treatment arm data has an overall success rate in the null of 0.017
  - Power in the other scenarios is 0.70, 0.90 and 0.90, better when there is some treatment effect on the other arms, worse when there is no treatment effect on the other arms
- Using the phase 2 data of the retained arm has an overall success rate in the null of 0.0296 – type-1 error is inflated by the selection in phase 2
  - Power in the other scenarios is 0.90, 0.91and 0.92, these are much better, but will they still be after we've adjusted to remove the type-1 error inflation?

Adjusting the type-1 error control

- We want to fix the problem (of the inflated type-1 error when using the phase 2 data of the retained arm) by adjusting the alpha level of the test at the end of phase 3.
  - What should it be adjusted to?
    - So overall type-1 error is 0.025 in the Null?
    - So type-1 error if we go to phase 3 is 0.025 in the Null?
- Ppn late futility Phase 2 is 0.85 in the Null, (0.10-0.06 in the other scenarios).
- Looking at the simulations.csv file we can see that we could limit the overall success to 0.025 is we adjust the alpha to 0.02.
- But only ~1500 sims of the Null go to Phase 3 (Ppn Late Futility S1 = 0.85) so to estimate alpha level to limit type-1 error to 0.025 conditional to going to Phase 3 (which may or may not be a regulatory requirement) need to run 10,000 sims of the Null without any futility rule at the end of phase 2.
- From this we can see that a one sided level alpha of 0.013 is required.

Comparing Results #2

- Using the phase 2 data of the retained arm with a final alpha level of 0.02 has an overall success rate in the null of 0.0249 – type-1 error just managed at the 0.025 level.
  - Power in the other scenarios is 0.90, 0.91 and 0.92, these still very good, hardly reduced at all. [The probability of success with the correct arm is 0.90, 0.81, 0.77]
- Using the phase 2 data of the retained arm with a final alpha level of 0.013 has an overall success rate in the null of 0.018.
  - Power in the other scenarios is 0.90, 0.90 and 0.91, these are still very good, and surprisingly similar to the results for a final level alpha of 0.02. [The probability of success with the correct arm is 0.90, 0.81, 0.76]

ADDING TIME TO ENDPOINT

- What if time to endpoint is 26 weeks?
- And also we include a 26 week (6 month) delay between the end of phase 2 and start of phase 3?
- How long is expected conventional phase 2 & phase 3?
- Average duration:
  - Null: 80 weeks
  - Alternative scenarios: 140-143 weeks.

26 week endpoint seamless phase 2

- Return to the FACTS core simulation of just the phase 2 part.
- Make it adaptive, and add a interim at 252 subjects enrolled
- Force the trial to stop at the interim by using complimentary success/futility criteria (e.g. Pr(Trt(Max) > Control) < 0.8, > 0.8
- Clearly there will be less information! The last 26 weeks of accrued subjects (6*26 on average) will not have final data. So we'll only have data on ~100 completers.
- (With the above rule) type-1 error is now 0.39; power is 0.82, 0.87, 0.88; probability of success & selecting the best arm is 0.77, 0.62, 0.61.

What if we allow phase 2 to recruit more?

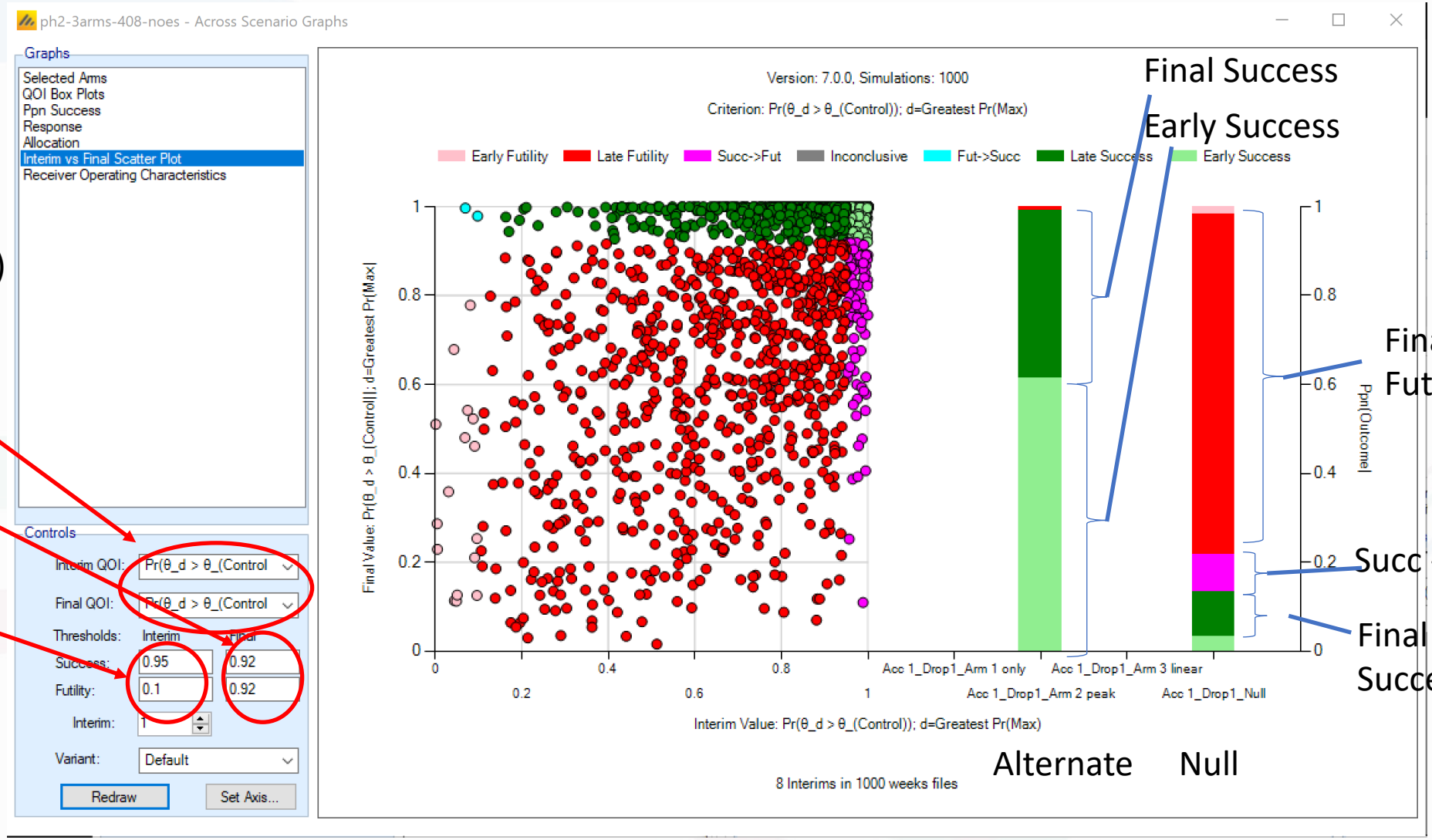
- Phase 2 is sample size is currently 252 participants.
- Accrual is ~6 per week. If we allow the first 252 subjects to complete, (another 26 weeks) we would accrue approximately another 156 subjects.
- So lets increase phase max N to 408, but add interims every 8 weeks after 252 recruited and allow adaptive allocation and early stopping.
- We simulate this without stopping rules and with 1,000 weeks files and then use the across scenario graph “Interim vs Final Scatter Plot”.

Interim v Final Scatter Plot

Using Pr(Trt(Max) > Control) for both final and interim decisions

Set final success/futility thresholds to 0.92

Set early success/futility thresholds to 0.95 / 0.1



Selecting Boundaries

- We can look interim by interim and see what thresholds allow some early stopping without more “incorrect” stopping than we can tolerate.
- We conservatively choose boundary thresholds of
  - $Pr(\theta_d > \theta_{Control}) > 0.99$ required for early stopping for success at each interim
  - $Pr(\theta_d > \theta_{Control}) < 0.1, 0.15, 0.2,$ required for early stopping for futility at interims 1, 2, 3
- Interim 4 is 8*3 weeks after enrolling 252 subjects, so will have nearly 252 complete and we decide to force a Go / No-go decision at this interim (success/futility is >< 0.92)
- Need to also ensure we’ve a clear enough decision on which dose is the ‘max’. We initially set this to Pr(Max) > 0.5. There are 3 treatment arms, so if one has a Pr(Max) > 0.5 it’s likely to have a significantly higher probability than the other two [unless there are 2 very good arms, and one very bad]

Options not explored here….

- More tuning could be done here.
  - Have and earlier interim (when ~50 complete),
  - have different end of phase 2 threshold (0.86 looks to keep type-1 error at ~ 0.2),
  - have more frequent interims?
  - Tune the required Pr(Max) threshold for success. As 0.5 has given reasonable OCs for dose selection we could try reducing it and checking the ppn trials successful and selecting a good arm.

Phase 2 results

- Using sample size of 408 requiring final $Pr(\theta_d > \theta_{Control}) > 0.92$ has an overall success rate in the null of 0.19 – type-1 error kept below the 0.2 level.
  - Power in the other scenarios is 0.91, 0.93 and 0.94, are the same. [The probability of success with the correct arm is 0.90, 0.76, 0.73 is slightly reduced]
- Original fixed design had an overall success rate in the null of 0.15
  - Power in the other scenarios of 0.91, 0.93 and 0.93 [The probability of success with the correct arm is 0.90, 0.81, 0.76]
- Compared to the fixed design (N = 252), adaptive design has E(N) ~ 325-332 (383 for the Null). But duration of adaptive is ~54-55 weeks (64 for the Null) compared to 68 for the fixed.
- The adaptive design is not stopping and following up, but continuing to enrol and stopping when it can.

Phase 3

- Create new staged design importing our FACTS core design as the first stage (phase 2).
- Looking at the results of the FACTS core simulation of the phase 2, we can see that the average allocation to the best arm is ~85 subjects. Previously we had 63 + 85 = 148 subjects on control and selected arm by the end of phase 3. So we allow our seamless trial to have an additional 60 subjects on each arm in phase3. (408 + 120)
- But we do this by specifying a combined max of 528 subjects, if phase 2 transitions early we will be able to have more in phase 3.
- However we will add interims to phase 3 to allow it to stop early.

Using Predictive Probabilities

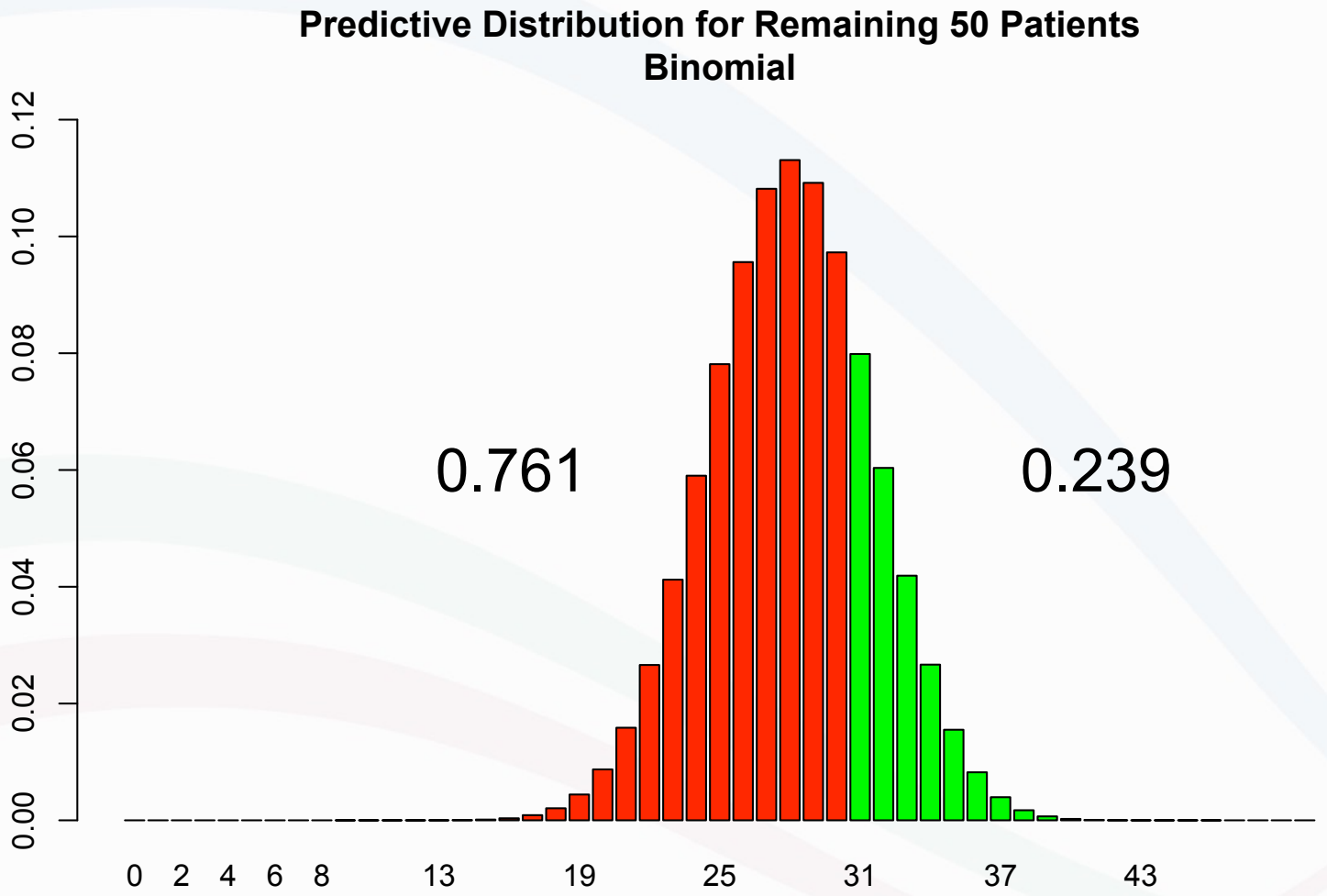
Simple Trial:

- Binary data. Observe $x \sim \text{Bin}(100, p)$
- Need to show $\Pr(p > 0.5 \mid x \text{ out of } 100) > 0.95$
- Assume $p \sim \text{Beta}(1,1)$ prior
- $\Pr(p > 0.5 \mid 59 \text{ out of } 100) = 0.963$
- $\Pr(p > 0.5 \mid 58 \text{ out of } 100) = 0.944$
- So need to see 59 successes

Observe data halfway through

- See 28/50 (56%) successes
- Need to see at least 31/50 (62%) to meet threshold
- What is predictive probability of trial success?

Using Predictive Probabilities



Using Predictive Probabilities

- At the interim, based on the interim data, predict the probability of success (e.g. p-value < 0.02) at the end of the trial.
- If this probability is too low (e.g. < 0.1) we can stop the trial for futility.
- At the interim, based on the interim data, predict the probability of success (e.g. p-value < 0.02) *if we follow up the currently enrolled subjects*.
- If this probability is sufficiently high (e.g. > 0.99) we can stop enrolment and complete the follow up of the current sample in the expectation of success.
- This results is a small inflation of type-1 error, and the final alpha level adjusting to compensate. This is done by simulation.

Phase 3 / Stage 2

- Keep the best phase 2 arm in phase 3, use all phase 2 data in phase 3.
- We specify interims in terms of number of subjects complete.
- First interim at 150 complete – (only counting subjects on arms in phase 3), note there will be ~26*6 = 156 subjects enrolled but not complete.
- Then interims every 50 subjects complete (~8 weeks).
- Initially simulate without early stopping bounds specified and as before use the Final vs Interim scatter plot to check boundaries.

Selecting Boundaries Phase 3

- We can look interim by interim and see what thresholds allow some early stopping without more “incorrect” stopping than we can tolerate.
- We conservatively choose boundary thresholds of
  - Pr(Final Success at current enrolment) > 0.99 required for stopping enrolment and following up in expection of success at interim 1, and > 0.98 at each subsequent interim
  - Pr(Final success at full enrolment) < 0.1 required for early stopping for futility at each interim

Phase 3 results compared to separate phase 2/3

- With 26 weeks to the final endpoint and accrual of 6 per week.
- Using the separate phase 2/3s has an overall success rate in the null of 0.0035 – type-1 error phase 3 managed at the 0.025 level.
  - Power in the other scenarios is 0.81, 0.77 and 0.78, [The probability of success with the correct arm is 0.81, 0.73, 0.69]
  - Max sample size is 422 across both stages. E(N) is 277 in the Null, 405-411 in the alternate
  - Duration, E(D) is 80 weeks in the Null, 141-143 in the alternate.
- Using the seamless phase 2/3 with the data of the retained arm with a final alpha level of 0.013 has an overall success rate in the null of 0.017.
  - Power in the other scenarios is 0.90, 0.90 and 0.91, these are still better than separate, and surprisingly similar to the results with an immediate endpoint. [The probability of success with the correct arm is 0.90, 0.76, 0.71]
  - Max sample size is 528 across both stages E(N) is 414, 439-443 in the alternate
  - Duration, E(D) is 74 weeks in the Null, 97-98 in the alternate, a saving of nearly a year.

To Conclude

- Thank You for attending
- Link to Recording will be sent out tomorrow
- Slides will be available via our website at the end of the series
- Any questions please contact us:
  - tom@berryconsultants.com
  - Kert@berryconsultants.com
  - facts@berryconsultants.com
- If you would like a demo and/or a free evaluation copy of FACTS – just ask.