† Outcome Adaptive Randomization with Delayed Survival Response — Huang et al. (2009, StatMed)

- **Delayed response** T**:** desired final response (for example, progression-free survival, PFS)

  $\star\star$ common in clinical trials, e.g., "GVHD within 100 days", any survival endpoint (PFS, OS etc.), "improvement in stroke score by week 13", etc.

  $\Rightarrow$ Investigators have to wait for the response of the currently recruited patients before being able to make a decision about sequential stopping or treatment allocation for the next patient.

- **Bayesian Paradigm:** always conditional on *observed* data, including censored response etc.

  $\Rightarrow$ Investigators can proceed on the basis of the partial information.

† Huang et al. (2009, StatMed) – contd

- **Early response:** increased efficiency by using available early response (S).

  For example, "improvement in stroke score by week 2"

- **Tumor response (S) and PFS (T):** Huang et al. use (early) tumor response (S) as early outcome in a trial with survival endpoint (T).

- **Tumor response vs. surival:** tumor response is usually used in phase II trials, although survival is the ultimate endpoint in phase III $\rightarrow$ many drugs fail in phase III.

- Jointly modeling S and T! $\Rightarrow$ Construction of more efficient clinical trial designs!

† Adaptive design with delayed survival response— Huang et al. (2009, StatMed)

- **Tumor response:** $S_i \in \{1, 2, 3, 4\}$,

  resistance to treatment or death ($S = 1$), stable disease (2), partial remission (3) or complete remission (4)

- **Survival:** $T_i = PFS$, $t_i = \min\{T_i, t - t_{0i}\}$ censored survival time (at time t for a patient recruited at t_{0i}), and $\delta_i = I(T_i = t_i)$

- **Sampling model:** joint model $p(S, T \mid x)$ under treatment $x \in \{A, B\}$

  $p_A = (p_{A1}, p_{A2}, p_{A3}, p_{A4})$ $\quad p_B$

$$p(S = j \mid x) = p_{xj} \quad \text{and} \quad T \mid S=j, x \sim \text{Exp}(\mu_{xj}),$$

where $\text{E}(T \mid S=j, x) = \mu_{xj} = 1/\lambda_{xj}$.

n_1

n_2

- **Prior:** conjugate

$$(p_{x1}, \ldots, p_{x4}) \sim \text{Dir}(\gamma_{x1}, \ldots, \gamma_{x4}) \quad \text{and} \quad \mu_{xj} \overset{indep}{\sim} \text{IG}(\alpha_{xj}, \beta_{xj}).$$

- **Posterior:** easy!

$j = 1, 2, 3, 4$

- **Adaptive allocation:** $\mu_x \equiv \sum_j p_{xj}\mu_{xj}$ mean survival under x. Allocate patients to arm A with probability $n_A > n_B$

$$p \equiv p(\mu_A > \mu_B \mid \mathbf{y})$$ = Post. Prob that A is better than B in terms of average survival

- **Early stopping for futility and for superiority:**
  - ⋆⋆ Early stopping for futility when $p < p_L$, e.g. $p_L = 0.025$ ∘ A is futile
  - ⋆⋆ Early stopping for superiority when $p > p_U$, e.g. $p_U = 1 - p_L = 0.975$ ∘ A is superior

i: cohort index

k: patient within a cohort index

† **Algorithm 4.6** Adaptive allocation with survival response.

- **0. Initialization:** Initialize $n = 0$ and calendar time $t = 0$. Initialize $p = 0.5$.
- **1. Next cohort:** If $n \geq N_{\max}$, continue with step 3. Otherwise,recruit the next cohort: $i = n + 1, \ldots, n + k$: Allocate treatment A and B with prob p and $q = 1 - p$.

  ⋆⋆ Increment calendar time by one week, $t = t + 1$.

  ⋆⋆ Record tumor responses, $S_{i,k}$, $k = 1, \ldots, 4$ for the new patients. Record the recruitment times $t_{0i} = t$.

  ⋆⋆ *When simulating* only (for the evaluation of OCs):

  generate simulation truth for the (future) PFS T_i for the new patients.

  ⋆⋆ Increment $n + k$.

† **Algorithm 4.6** Adaptive allocation with survival response. – contd

- **2. Posterior updating:** Update the posterior parameters. Compute and record $p = p(\mu_A > \mu_B \mid \mathbf{y})$
- **3. Stopping:**

  ⋆⋆ If $p < p_L$, then stop for futility.

  ⋆⋆ If $p > p_U$, stop for efficacy.

  ⋆⋆ If $t > t_{\max}$ stop for the maximum horizon.

  Otherwise continue with step 1.

† **Example 4.7** – Adaptive leukemia trial with survival response – Huang et al. (2009)

- **Study:** phase II trial for AML.
- **Scenario:** simulation truth with higher response rates and longer response durations under B, i.e., PFS, under treatment B.

$$p_{Aj} = (0.2, 0.4, 0.1, 0.3) \text{ and } \mu_{Aj} = (4, 30, 75, 110), \text{ i.e. } \mu_A = 53$$
$$p_{Bj} = (0.1, 0.1, 0.2, 0.6) \text{ and } \mu_{Bj} = (6, 45, 112, 165), \text{ i.e. } \mu_B = 126$$

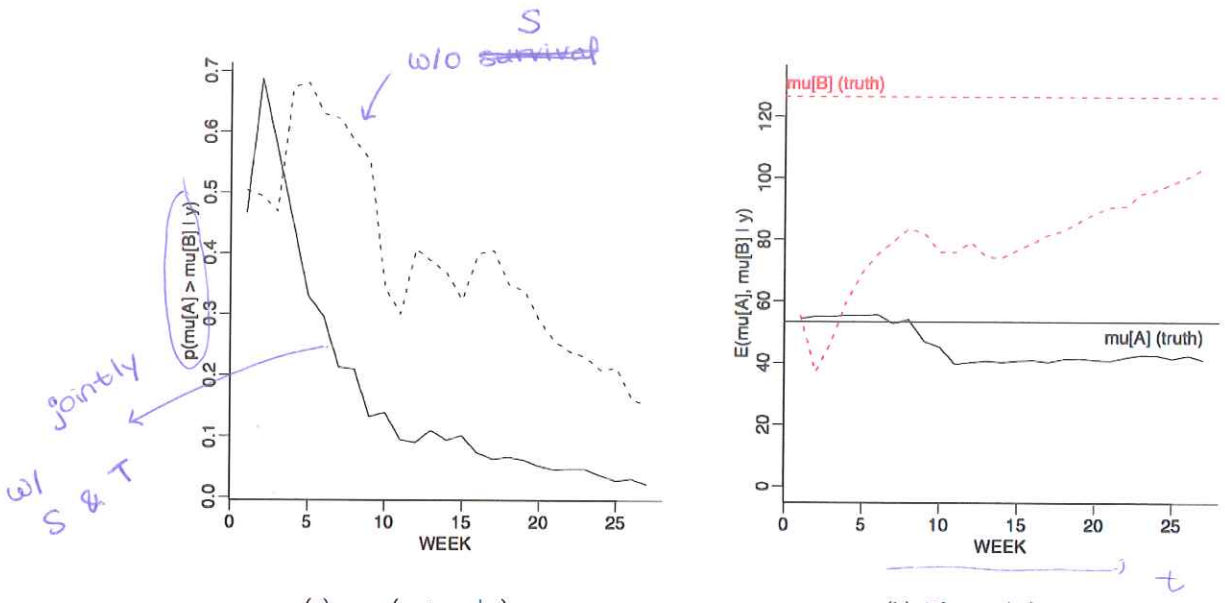
- **Design:** $n = 120$, $k = 1$ pat/week, $t_{\max} = 160$

$P = P_r(\mu_A > \mu_B \mid y)$ would be small if the design works well

$\Rightarrow \quad P < P_L$

$\Rightarrow$ Stop for futility

* Figure 4.6 Results: In (a) the dashed line shows without regression on S. TRUTH = B is better.



(a) $\pi = p(\mu_a > \mu_b \mid y)$

(b) $E(\mu_a, \mu_b \mid y)$

more than 50 types of sarcoma

Sarcoma: a rare kind of cancer
happen in a different kind of tissue
grow in connective tissue (bones, muscles, tendons...)

† Hierarchical models for phase II designs (BCLM Sec. 4.5)

Thall et al. (2003)— Borrowing strength across related phase II trials.

- **Study:** multiple related phase II trials, e.g., $J = 12$ different sarcomas.

  Binary outcomes $p(X_{ji} = 1) = \pi_j$ for patient i in study j

- **Goal:** borrow strength across sub-populations.

  compromise of the extremes of pooling vs. separate studies

  Note: A practical limitation: very slow accrual. That is, only very small sample sizes (6 or fewer) $\Rightarrow$ only very small sample sizes making it impossible to run separate trials.

- **Prior Model:** Hierarchical model across J subpopulations $\theta_j = \log(\pi_j/(1-\pi_j))$

$$\theta_j \overset{iid}{\sim} \mathsf{N}(\mu, 1/\tau),$$

with population level pars $\eta = (\mu, \tau)$ and hyperprior

$$\mu \sim \mathsf{N}(m_\mu, s_\mu), \text{ and } \tau \sim \mathsf{Gamma}(a_0, b_0).$$

- **Posterior:** easy
- **Posterior summary:**

$$p_{30,j} = p(\pi_j > 30\% \mid \mathbf{y}^{(n)}).$$

Will use $p_{30,j} < 0.005$ for decision to stop accrual for subpopulation j.

Could use the same model and inference for alternative decision rules.

† **Algorithm 4.7** Phase II design for related subpopulations

- **1. Initialize:** Initialize $n_j = y_j = 0$ for each subpopulation j

- **2. Simulate patient cohort:** Generate a random cohort size $p(k_j = k) = a_{jk}$, $j = 1, \ldots, J$ and $k = 1, \ldots, K$ and responses $x_j \sim \text{Bin}(k_j, \pi_j^o)$ where π_j^o simulation truth.

  Increment $n_j \equiv n_j + k_j$ and $y_j \equiv y_j + x_j$.

- **3. Update posterior inference:** Evaluate $p_{30,j} = p(\pi_j > 30\% \mid \mathbf{y}^{(n)})$ and posterior moments $\text{E}(\pi_j \mid \mathbf{y}^{(n)})$.

  Computation requires several posterior integrations. See **Algorithm 4.8** for details.

$$p_{30,j} = P(\pi_j > 30\% \mid y^{(n)}) = \int I(\pi_j > 30\%) \; p(\boldsymbol{\theta}, \mu, \tau \mid y^{(n)}) \, d\boldsymbol{\theta}, \mu, \tau$$

† **Algorithm 4.7** Phase II design for related subpopulations – contd

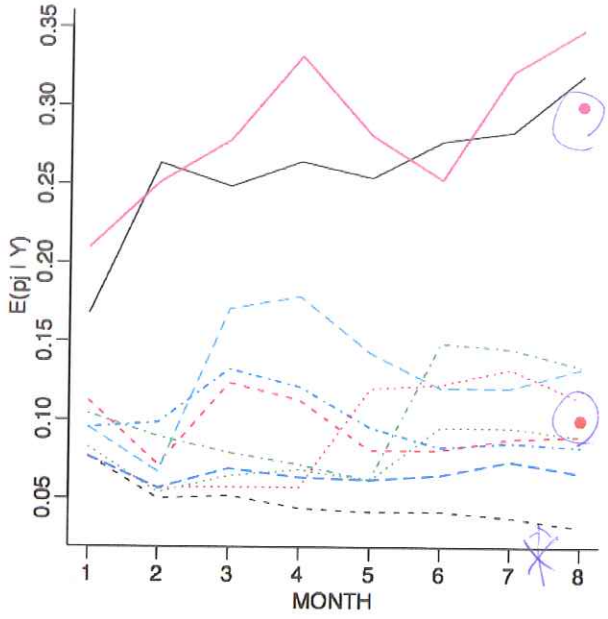
- **4. Drop trials**: Identify all subpopulations with $p_{30,j} < 0.05$ and exclude them from further recruitment.
- **5. Stopping:** If all subpopulations are dropped, then stop the trial for lack of efficacy in all subpopulations.
  - ⋆⋆ If $\sum n_j > N_{\max}$, stop for maximum enrollment.
  - ⋆⋆ Otherwise repeat with step 2.

† **Example 4.8**– Sarcoma Study (Thall et al. (2003))

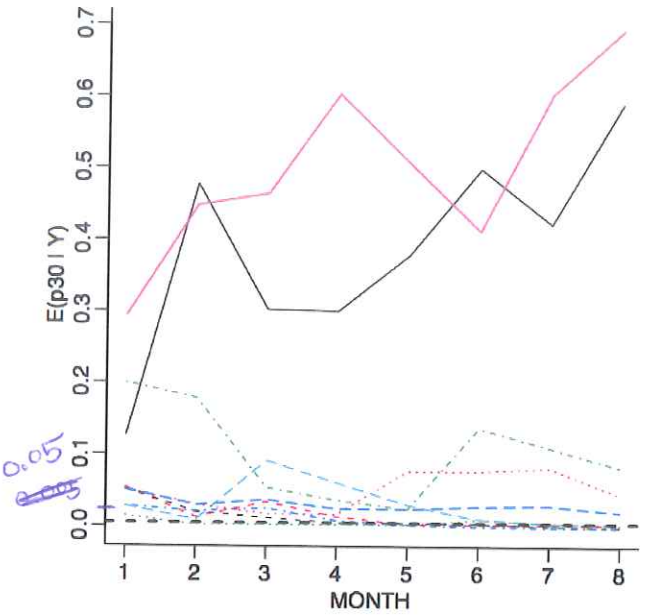
- **Study:** phase II trial of imatinib in $J = 10$ different sarcomas
- **Hyperprior:** $\mu \sim \text{N}(1.386, 10)$, mean matching logit(0.20)
  $\tau \sim \text{Gamma}(2, 20)$ with $\text{E}(\tau) = 0.10$.
- **Scenario:** $\pi^0 = (3, 1, 1, 1, 1, 3, 1, 1, 1, 1)/10$ ← TRUE
  accrual $\text{E}(k_j) = 5.5$ for $j = 1, \ldots, 5$ and $\text{E}(k_j) = 2$ for $j = 6, \ldots, 10$.

large cohort small cohort size

† **Example 4.8** Results (Figure 4.7)



(a) $E(\pi_j \mid \boldsymbol{y}^{(n)})$ = $\hat{\pi}_j$



(b) $p_{30,j}$.

† **Decision theoretic designs** (BCLM Sec 4.6)

Bayesian decision problem: probability model for observable data y and (unobservable) parameters (θ)

plus utility function $u(d, \theta, y)$ = worth of a hypothetical outcome y under assumed truth θ and action d.

- **Prob model:** $p(\theta, y) = p(\theta)p(y \mid \theta)$
- **Action space:** $d \in \mathcal{A}$, e.g.,

  ⋆⋆ sequential stopping

  $$d \in \{ \text{ stop for victory, stop for failure, continue} \};$$

  ⋆⋆ dose allocation $d \in \{1, \ldots, J\}$; treatment allocation, etc.

† **Decision theoretic designs** – contd

- **Utility function:** $u(d, \theta, \mathbf{y})$, relative preferences under assumed truth. Data $\mathbf{y}$ often includes observed and future data, $\mathbf{y} = (\mathbf{y}_0, \mathbf{y}_1)$.

- **Bayes rule:** rational decision maker maximizes utility,

$$U(d, \mathbf{y}_0) = \mathrm{E}\left(u(d, \theta, \mathbf{y}_0, \mathbf{y}_1) \mid \mathbf{y}_0\right)$$

  in expectation w.r.t. θ and $\mathbf{y}_1$, conditioning on $\mathbf{y}_0$.

  ⋆⋆ Bayes rule

$$d^\star = \arg\max U(d, \mathbf{y}_0)$$

† Decision theoretic design– caveats

- *Which* utility?
- Prob model and utility function *implicitly* define the optimal rule. There is no guarantee that the optimal rule looks sensible.
- *Sensitivity* of solution to prior and utility function.
- A rational decision maker would not randomize – of course that's crazy!
- Lack of *simplicity*, $d^* = \arg\max \int u(\cdot)dp(\cdot)$ is not usually closed form.
- But that said, in some cases a stylized decision theoretic solution is very useful.

AMS 276
Lecture 10: Phase III Studies

Fall 2016

† Phase III Studies: comparative trials that evaluate the effectiveness of the new treatment relative to the current standard care (BCLM Chapter 5)

- Typically randomized controlled multicenter trials on large patient groups (300–3000 or more)
- **Aim:** Definitive assessment of how effective the drug is in comparison with current "gold standard" treatment. ⇒ confirmatory trials.
- Possible control groups: historical control (subjects treated earlier using the control), concurrent control (subjects treated at some other sites or clinics), randomized control (subject treated at the same time under the same conditions).

† Phase III Studies — contd

- The confirmatory study is typically overseen and judged by a regulatory agency.

  $\Rightarrow$ desired result: the approval of a therapy for public use.

  $\Rightarrow$ create predetermined statistical thresholds ("hurdles") that must be met for the regulatory agency to approve the medical therapy for public use

  e.g. Get a statistically significant result at a specified Type I error level.

- In fact, the most important regulatory hurdle is the Type I error of the design.

- Extremely difficult to calculate for adaptive Bayesian designs.

- Earlier phase studies have "learning" goals. & Their definitions of statistically important effects are not regulated by agencies.

† Possible adaptive features in confirmatory trials.

- **Adapt the sample size to the results of the trial:** appropriate sample size to determine success of the treatment & the ability to determine when the treatment is unlikely to show trial success and stop.
- **Arm dropping:** Start with multiple trt arms with the desire that one of the arms may be dropped during the trial.
- **Seamless phase II/III trial**

  ⋆⋆ Say we have two experimental medical therapies.

  ⋆⋆ Construct a trial that "learns" which of the two is better. The better of the two continues to the confirmatory stage.

  ⋆⋆ All of the data on the selected arm are used in the confirmatory analysis.

† Are *Bayesian* Phase III trial designs commonly used in practice?
No, not yet. Still very rare.

- There is increasing need for Bayesian designs when the patient population is small. e.g. orphan drug and pediatric trials
- For device trials (CDRH: Center for Devices and Radiological Health), FDA approves Bayesian designs (but still not so common).

Guidance
for Industry and FDA Staff

Guidance for the Use of
Bayesian Statistics in
Medical Device Clinical Trials

Document issued on: February 5, 2010

The draft of this document was issued on 5/23/2006

† Bayesian adaptive confirmatory trials (BCLM Section 5.2)

* How does the adaptive feature of a Bayesian design affect the Type I error (the most important regulatory hurdle)?

* Let's look at **Example 5.1**.

- One-arm trial with binary outcome (success or failure)
- **Sampling model:** For n observations, the number of successes is

$$X \mid p \sim \text{Bin}(n, p)$$

- Assume that the regulatory agency agrees the trial must show $p > 0.5$ to demonstrate statistical success of the therapy.

- This can be formulated as a hypothesis testing problem

$$H_0 : p \leq 0.5 \quad \text{vs} \quad H_a : p > 0.5$$

Say, fixed sample size $n = 100$ and significance level at the one-sided 0.05-level.

- **Frequentist Approach:** the Type I error $\Leftrightarrow$ rejecting the null that is correct.

Suppose we reject H_0 if $X \geq x$. Then

$$\text{Type I error} = Pr(X \geq x \mid H_0 \text{ true }) = \sum_{X=x}^{100} \binom{100}{x} (0.5)^x (1-0.5)^{n-x}.$$

- $Pr(X \geq 59 \mid \text{ true } H_0) = 0.0443$ ($\alpha = 0.05$)
- $Pr(X \geq 58 \mid \text{ true } H_0) = 0.066$

$\Rightarrow$ Statistical success if $X \geq 59$!

- **Bayesian Approach:** $X \mid p \sim \text{Bin}(n, p)$
  - ⋆⋆ Assume $p \sim \text{Be}(1,1) \Rightarrow p \mid X \sim \text{Be}(1+X, 1+n-X)$.
  - ⋆⋆ **Rule:** statistical success (i.e., reject H_0) if $Pr(H_a \mid X) > 0.95$.
    - $Pr(H_a : p > 0.5 \mid X = 58) \geq 0.945$ ✗
    - $Pr(H_a : p > 0.5 \mid X = 59) \geq 0.964$ ✓
    - $Pr(H_a : p > 0.5 \mid X = 60) \geq 0.977$ ✓
      - $\Rightarrow$ Statistical success if $X \geq 59$ (accidentally coincide!)
      - $\Rightarrow$ The Type I error for this Bayesian rule is 0.044. $P(X \geq 59 \mid p=0.5)$ 0.1
  - ⋆⋆ In general, the question is how to find a Bayesian rule to have adequate Type I error characteristics.

$P(p > 0.5 \mid n=75, X \geq 45) \geq 0.95$

- Now let's consider an adaptive design – add **interim analysis rules**.

  ⋆⋆ Suppose statistical success if at $n = 50$ or $n = 75$, there is at least a $P_{cut} = 0.95$ posterior probability of superiority.

  $\Leftrightarrow$ If $X_{50} \geq 31$, $X_{75} \geq 45$ or $X_{100} \geq 59$, claim success.

  $P(p > 0.5 \mid n=50, X \geq 31) \geq 0.95$

  ⋆⋆ What is the resulting Type I error?

  ↻ X_1, X_2 and X_3 are independent binomial random variables with sample sizes 50, 25, and 25, respectively. Then the Type I error is

$$1 - \sum_{i=0}^{30} \sum_{j=i}^{\min(25+i,44)} \sum_{k=j}^{\min(25+j,58)} Pr(X_1 = i) Pr(X_2 = j-i) Pr(X_3 = k-j).$$

  50 25 25

  $\Rightarrow$ The Type I error for this Bayesian rule is 0.09578662.

$P(\text{reject } H_0 \mid H_0 \text{ true}) = 1 - P(\text{do not reject} \mid H_0 \text{ true})$

- an adaptive design — contd

  ⋆⋆ We can compute the Type I error through simulation. Run a million simulated trials and summarize the proportion of rejected trials.

  From page 198,

```
> out
  Pr(win)     MeanSS      SD SS         50         75        100
0.095770  96.455625  12.247579  0.059165  0.023445  0.917390
```

  ⋆⋆ *Wait!* 0.0958 is too high. Find a Bayesian rule that has the Type I error less than the "required level."

  ⋆⋆ Make it less reject. $\Rightarrow$ raise P_{cut} (the posterior prob. threshold).

  ⋆⋆ How to find the right P_{cut}? Do simulation runs!

- an adaptive design — contd

  ⋆⋆ Table 5.1: The Type I error for different posterior probability levels of success (P_{cut})

P_{cut}	Type I error
0.95	0.0958
0.96	0.0692
0.97	0.0591
0.9725	0.0591
0.975	0.0532
0.976	0.0423
0.9775	0.0347
0.98	0.0347
0.99	0.0195

  ⋆⋆ *Wait!* How about the power and other operating characteristics with $P_{cut} = 0.976$?

- an adaptive design — contd

* Recall $H_0 : p \leq 0.5$ vs $H_a : p > 0.5$

p^{TRUE}	Pr(Win)	Mean SS	SD SS	Pr(50)	Pr(75)	Pr(100)
0.50	0.0421	98.9	6.9	0.017	0.011	0.972
0.55	0.217	94.7	14.2	0.077	0.058	0.864
0.60	0.578	84.1	21.0	0.237	0.162	0.601
0.65	0.889	69.0	21.1	0.504	0.229	0.266
0.70	0.989	57.0	14.2	0.780	0.160	0.060
0.75	0.999	51.5	6.53	0.944	0.051	0.005

Table 5.2 *Operating characteristics of the adaptive design with early stopping for success for Example 5.2.1.*

** *Note:* If a fixed trial of 69 subjects were conducted, the power under the alternative hypothesis of 0.65 would be 0.802

- Recall "sampling to a foregone conclusion" (repeated testing problem): Due to the repeated testing, Type I error rates can be seriously inflated.
- Bayesian: Likelihood principle – Inferences depend on the current study only through the data actually observed.
- Changing $P_{cut} = 0.95$ to $P_{cut} = 0.975$ for claiming success is so unnatural to Bayesian.
- Adaptive features (interim analyses), prior distribution, evaluation by frequentist properties... the paradigm clash!
- When the Bayesian design is required to control Type I error, it loses may of its philosophical advantages.

★★ "The paradigm conflict inherent in tweaking Bayesian designs using frequentist criteria" Read Section 5.5 for more.

† Futility analysis – Use predictive probabilities.

∗ For confirmatory trials, there are the strict hurdles for success.

- Let's create rules for stopping the trial for futility at the 50- and 75-subject look.
- **PP:** prob that the trial will be successful if it continues accruing subjects.

  ⋆⋆ Current posterior: $p \mid \text{data} \sim \text{Be}(\alpha, \beta)$

  ⋆⋆ The predictive distribution for the next n patients is a beta-binomial distribution,

$$f(x) = \frac{\Gamma(\alpha+\beta)\Gamma(n+1)\Gamma(x+\alpha)\Gamma(n-x+\alpha)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(x+1)\Gamma(n-x+1)\Gamma(n+\alpha+\beta)}$$

- Prior $p \sim \text{Be}(1, 1)$
- 25 successes in 50 subjects are observed. Be(26, 26)
- Our strict hurdle: statistical success if 47 of the first 75 or 60 of the 100 total subjects are successes.

$X_{75} \geq 47$
$\Rightarrow$ reject H_0

$X_{100} \geq 60$
$\Rightarrow$ reject H_0

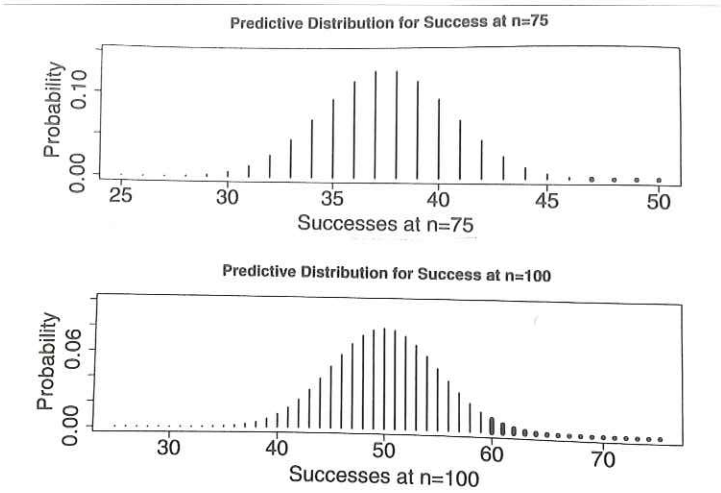


Figure 5.1 *The predictive distribution of the number of successes at the 75 and 100 subject looks for the trial in Subsection 5.2.2.*

★★ The PP of success at the $n = 75$ look is 0.00078 and is 0.0256 for the final analysis at $n = 100$.

- Let's include stopping rules for futility based on the predictive probability of success at the final analysis. $n=100$

  e.g. Stop the trial for futility if the prob of success at the final analysis is < 0.05.

- Then we have to check everything again. :-)

p	Pr(Win)	Mean SS	SD SS	Pr(50)	Pr(75)	Pr(100)
0.50	0.0407	64.3	18.2	0.016 0.555	0.011 0.275	0.014 ← stopping for success 0.129 ← for futility
0.55	0.215	74.1	20.7	0.078 0.283	0.059 0.253	0.078 0.248
0.60	0.569	76.1	21.1	0.238 0.099	0.161 0.122	0.170 0.210
0.65	0.882	67.3	20.1	0.506 0.021	0.227 0.028	0.148 0.069
0.70	0.987	56.8	13.9	0.782 0.003	0.158 0.003	0.048 0.008
0.75	0.999	51.5	6.4	0.945 0.000	0.050 0.000	0.005 0.000

Table 5.3 *Operating characteristics of the adaptive design in Subsection 5.2.2 with early stopping for success and futility. The two numbers in each of the last three rows represent the probability of stopping for success (top) and futility (bottom).*

- The predictive probability rule is driven by the current trial and the conclusions drawn in the current trial– not by some measure of learning or inference.
- Stopping a trial for futility implies
  - ⋆⋆ the evidence is conclusive
  - ⋆⋆ the medical therapy is detrimental
  - ⋆⋆ the therapy is conclusively inferior
  - ⋆⋆ the statistical hurdle is very unlikely to be met.

- More topics related to Phase III study designs are discussed in BCLM Chapter 5
- Special topics in Chapter 6 (incorporating historical data, equivalence studies, multiplicity and subgroup analysis).


Congratulations
dreams come true
You did it!