OPTIMUM Simulations/Stats Notes

OPTimising IMmunisation Using Mixed schedules

Prepared by: James Totterdell

2019-02-14

Contents

1 Background and Rationale

The aim is to o assess the allergy-preventive benefit and the safety of using wP as the first infant pertussis vaccine dose, compared with using aP for all doses.

2 Primary Outcome

The primary outcome is challenge-proven IgE-mediated food allergy by age 18-months. The end-point is challenge-proven IgE-mediated food allergy at 18-months.

3 Sample size and accrual

The end-point is food allergy at 18-months. For simplicity, assume babies are enrolled and randomised at 0 months of age. So, no follow-up data is available until 18-months after the first infant is enrolled.

Suppose accrual is 20 infants per week, then, by the time we have follow-up on the first individual we will have enrolled 1,560 infants (78 weeks $\times$ 20 per week). Assuming the first analysis was at $n = 500$, we would have about 2,000 individuals enrolled, so 1,500 with missing information at the time of the first interim. Full follow-up would occur at about week 228 (Figure 1).

Suppose accrual is 10 infants per week, then, by the time we have follow-up on the first individual we will have enrolled 780 infants. Assuming the first analysis was at $n = 500$, we would have about 1,300 enrolled, so 800 with missing information at the first interim. Full follow-up would occur at about week 378.

The minimum acrrual rate needed to enroll 3,000 infants over 5 years is about 11.5 per week.

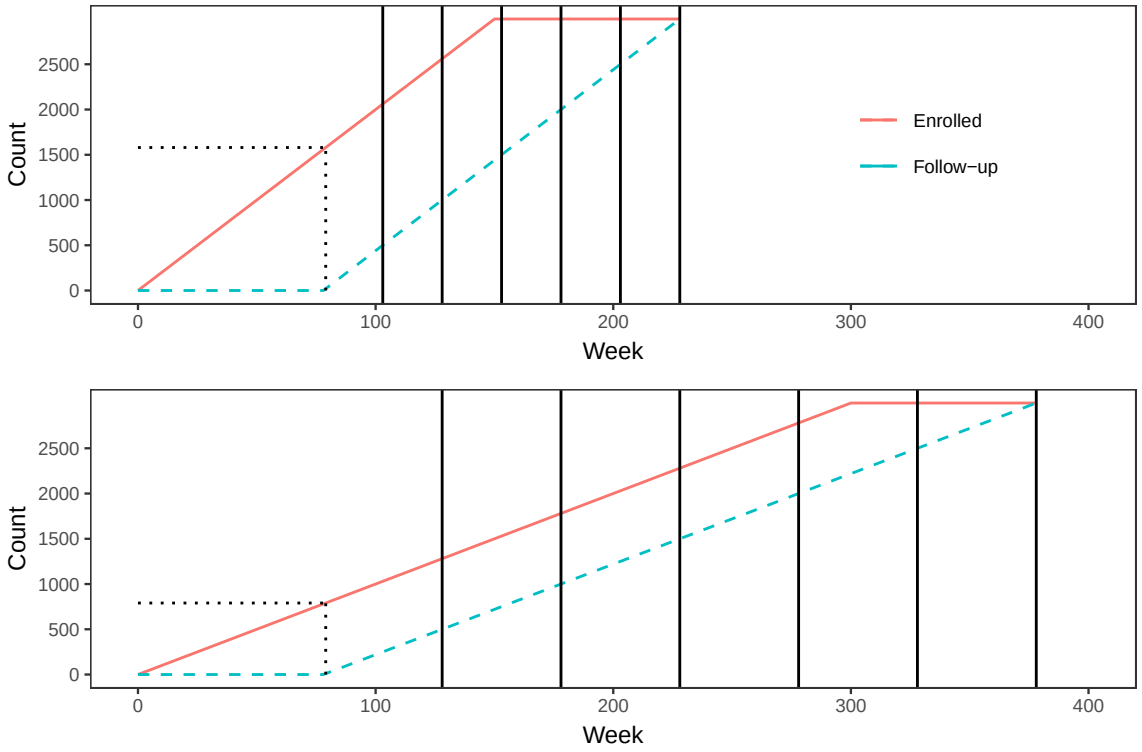


Figure 1: Assumed accrual rate and associated delay of information.

4 Statistical Analysis

4.1 Model

Let θ_a be the probability with food allergy amongst infants who receive the acellular pertussis vaccine at first dose, and θ_w the probability of food allergy amongst infants received the whole-cell pertussis vaccine at first dose. We are interested in estimating $\delta = \theta_w - \theta_a$ and investigating the statistical hypothesis

$$H_0 : \delta \geq 0$$
$$H_1 : \delta < 0$$

That is, that θ_w is no better than θ_a versus θ_w is lower than θ_a.

We could approach this in two ways:

- We might model θ_a and θ_w directly using independent Beta-Binomial models and integrate over the random variable δ to obtain posterior probabilities. The advantage is simplicity in the computations.
- We might model $\theta_a = 1/(1 + \exp(\beta_0)$, and $\theta_b = 1/(1 + \exp(-\beta_0 - \beta_1))$, that is as a logistic regression model, where $\delta = \beta_2$ is our parameter of interest (the difference in log-odds). The advantage we can incorporate more complexity (adjust for confoudners, partial pooling etc), and explore subgroup effects.

Given the lengthy delay in information, we will likely utilise posterior predictive probabilities at any interim analyses to impute the as yet unobserved data.

For the purpose of simplifying the simulations the independent Beta-Binomial models make sense. However, for the actual analyses, if covariates are to be included, logistic or logistic mixed models would be used. In any case, the Beta-Binomial model should reasonable approximate a main effects only logistic regression. Including confounders should only improve power.

4.2 Independent Beta-Binomial Models

Suppose that at each analysis $k = 1, ..., K$ we have data on n_k^i individuals with y_k^i responses for $i \in \{a, w\}$. We also assume that we have $m_k^i \geq n_k^i$ total enrolled but not all with data. The number without data is $\tilde{n}_k^i = m_k^i - n_k^i$. At an interim analysis we wish to impute the data for individuals enrolled but without follow-up. We denote these missing number of responses by $\tilde{y}_k^i$.

In addition to enrolled individuals with missing data, there are the yet to be enrolled individuals making up the maximum sample size. At stage K we have n_K^i individuals with y_K^i responses, and so for this end point we have $\tilde{n}_k^i = n_K^i - n_k^i$ data points missing. In either case, the posterior predictive will have the same parameters but with a different sample size parameter. Therefore in what follows we do not distinguish between the two, however, it is standard to use $\tilde{n}_k^i = m_k^i - n_k^i$ in deciding expected success and $\tilde{n}_k^i = n_K^i - n_k^i$ in deciding futility.

We specify the following model for $i \in \{a, w\}$ and $k \in \{1, ..., K\}$,

$$\begin{aligned}
\pi_0^i(\theta^i) &= \text{Beta}(\theta^i | a^i, b^i) \\
f_k^i(y_k^i | \theta^i) &= \text{Binomial}(n_k^i, y_k^i) \\
\pi_k^a(\theta^i | y_k^i) &= \text{Beta}(\theta^i | a^i + y_k^i, b^i + n_k^i - y_k^i) \\
P_k &= \mathbb{P}_{\Theta^a, \Theta^w | Y_k^a, Y_k^w}(\theta^w < \theta^a) \\
&= \int_0^1 \pi_k^a(\theta^a | y_k^a) \left[\int_0^{\theta^a} \pi_k^w(\theta^w | y_k^w) d\theta^w \right] d\theta^a \\
\tilde{f}_k^i(\tilde{y}_k^i | y_k^i) &= \text{Beta-Binomial}(\tilde{y}_k^i | \tilde{n}_k^i, a^i + y_k^i, b^i + n_k^i - y_k^i) \\
\tilde{\pi}_k^i(\theta^i | y_k^i + \tilde{y}_k^i) &= \text{Beta}(\theta^i | a^i + y_k^i + \tilde{y}_k^i, b^i + n_k^i + \tilde{n}_k^i - y_k^i - \tilde{y}_k^i) \\
\tilde{P}_k &= \mathbb{P}_{\Theta^a, \Theta^w | Y_k^a + \tilde{Y}_k^a, Y_k^w + \tilde{Y}_k^w}(\theta^w < \theta^a) \\
&= \int_0^1 \tilde{\pi}_k^a(\theta^a | y_k^a + \tilde{y}_k^a) \left[\int_0^{\theta^a} \tilde{\pi}_k^w(\theta^w | y_k^w + \tilde{y}_k^w) d\theta^w \right] d\theta^a \\
\text{PPoS}_k(q) &= \mathbb{E}_{\tilde{Y}_k^a, \tilde{Y}_k^w | Y_k^a, Y_k^w} \left[\mathbb{I}\left\{ \tilde{P}_k > q \right\} \right] \\
&= \sum_{i=0}^{\tilde{n}_k^a} \sum_{j=0}^{n_k^w} \mathbb{I}\left\{ \tilde{P}_k > q \right\} \tilde{f}_k^w(j | y_k^w) \tilde{f}_k^a(i | y_k^a)
\end{aligned}$$

The quantity P_k cannot be calculated analytically but can be evaluated numerically or estimated using Monte Carlo methods. Although PPoS_k can be computed analytically (assuming we have calculated the relevant $\tilde{P}_k$) it may still be more efficient to estimate using Monte Carlo methods for large sample sizes.

4.3 Decision Rules

At the final analysis (full follow-up on all individuals), a terminal decision is made regarding the difference in response between the two vaccines. This decision rule declares $\theta_w < \theta_a$, $\theta_w \geq \theta_a$, or that the study was inconclusive.

$$\delta_K(y_K) = \begin{cases} a_0 \text{ if } P_k \leq \underline{c}_K & \implies \text{accept } H_0 \\ a_1 \text{ if } P_k \geq \overline{c}_K & \implies \text{accept } H_1 \\ a_2 \text{ otherwise} & \implies \text{inconclusive} \end{cases}$$

At each interim analysis, a decision is made whether the study should be stopped for futility, expected success, or to continue enrolment. This decision is based on $\text{PPoS}_k(q)$ which depends on the chosen q. Perhaps setting $q = \overline{c}_K$ makes the most sense, as this is the criteria which would be used in assessing success at the final analysis. The interim decision rule is then

$$\delta_k(y_k) = \begin{cases} a_3 \text{ if } \text{PPoS}_k(\overline{c}_K) < \underline{\kappa}_k & \implies \text{futile to continue} \\ a_4 \text{ if } \text{PPoS}_k(\overline{c}_K) > \overline{\kappa}_k & \implies \text{expect success at interim} \\ a_5 \text{ otherwise} & \implies \text{continue to enrol to } k+1. \end{cases}$$

If stop for futility, when remaining enrolled individuals are followed-up underatake another predicitve probabilitiy analysis of futility to the original full sample size.

If stop for expected success, assess terminal decision rule when complete follow-up on remaining enrolled individuals is avaialable.

4.4 Logistic Regression

Advantage is that we can incorporate additional covariates. However it complicates the simulations. Posterior probabilities are no longer analytically tractable, but instead must be approximated by Monte Carlo measures,

or by approximating densities.

Let X denote fixed level design, and Z the group-specific design, then we model

$$\begin{aligned}
\eta_i &= x_i^\top \beta + z_i^\top \gamma \\
p_i &= \text{logit}^{-1}(\eta_i) \\
f_k(y_{k,i}|\beta) &= \text{Bernoulli}(p_i) \\
\pi_k(\beta|y_k) &\approx N^{-1} \sum_{j=1}^{N} \delta_{\beta^j}(d\beta), \quad \beta^{1:N} \sim \pi_k(\beta|y_k) \\
\tilde{f}_k(\tilde{y}_k|\beta,\gamma) &\approx N^{-1} \sum_{j=1}^{N} \delta_{\tilde{y}_k^j}(\tilde{y}_k), \quad \tilde{y}_k^{1:N} \sim \tilde{f}_k(\tilde{y}_k|\beta,\gamma)
\end{aligned}$$

5 Simulations

We want to investigate the operating characteristics of the trial for varying θ^a and θ^w and determine appropriate values of the following trial parameterse:

- Prior parameters: a^i and b^i.
- $(\underline{c}_K, \overline{c}_K)$ - the bounds used at the final analysis for decisions
- q - the value used in PPoS(q) - should this just be set to $\overline{c}_K$?
- $(\underline{\kappa}_k, \overline{\kappa}_k)$ - the bounds used for determining futility and expected success at interim analyses
- The frequency of interim anlayses

Assuming a maximum sample size of 3,000 (1,500 in each arm) at a fixed final anlaysis we estimate the following probability of success.

$\theta_w^\star$	$\theta_w^\star$	$\mathbb{P}(\Theta_1 \mid y_K)$
0.10	0.070	0.904
0.03	0.015	0.874
0.28	0.210	0.996

For assessing interim analyses, we assume two accrual scenarios: 20 per week and 10 per week. The difference between the two (apart from the total study length required), corresponds to varying the delay in information from follow-up. As previously stated, In the 20 per week case, we expect individuals with follow-up to be about 1,500 behind the number of individuals enrolled at the time of the first interim, whereas for 10 per week we expect individuals with follow-up to be about 800 behind at the number enrolled at the time of the first interim. This effects the value of PPoS_k which will generally be closer to P_k the larger the dimension of the posterior predictive distribution considered.

5.1 Example

In the example we set $\theta_a = \theta_w = 0.1$, so that a result of success is a Type I error.

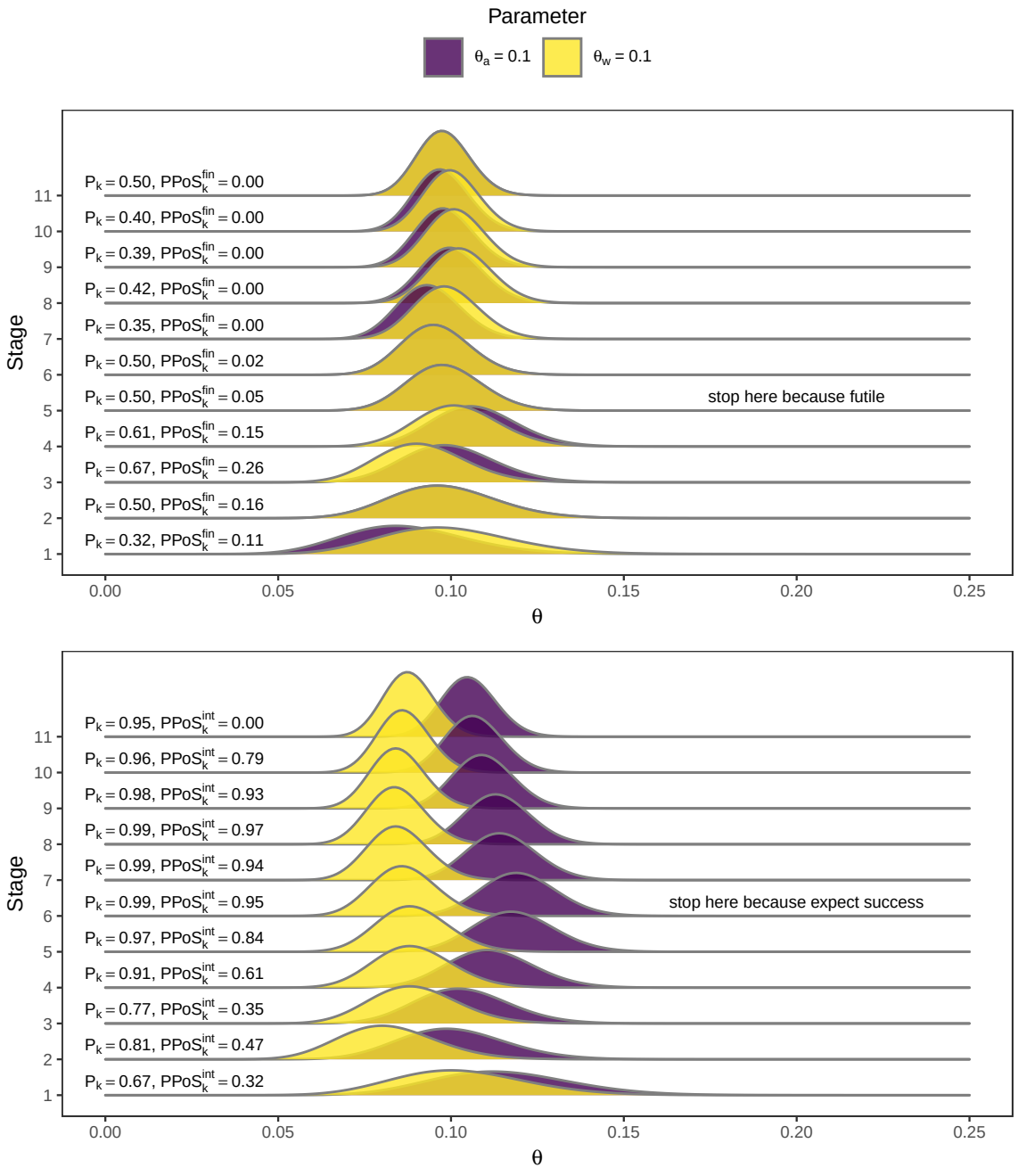


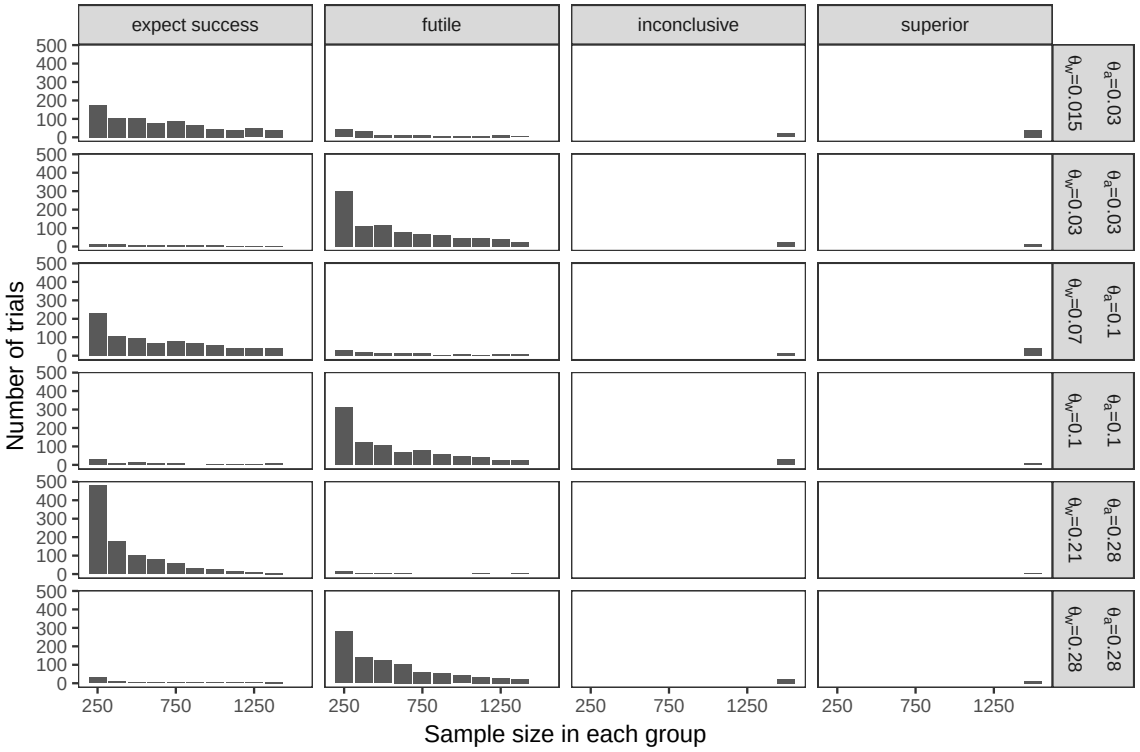
Figure 2: Two example trial procession of posteriors.

5.2 Accrual - 20 per week

Run-time was 37.4 minutes.

Scenario	$\theta_a^\star$	$\theta_w^\star$	$\mathbb{P}$(e.s)	$\mathbb{P}$(l.s)	$\mathbb{P}$(e.f)	$\mathbb{P}$(l.f)	$\mathbb{P}$(s)	$\mathbb{P}$(f)	$\mathbb{P}$(inc)	$\mathbb{P}$(s.e)
1	0.10	0.10	0.08	0.01	0.88	0	0.09	0.88	0.03	0.75
2	0.10	0.07	0.83	0.04	0.12	0	0.87	0.12	0.01	0.68
3	0.03	0.03	0.07	0.01	0.89	0	0.08	0.89	0.02	0.73
4	0.03	0.02	0.79	0.04	0.16	0	0.82	0.16	0.02	0.66
5	0.28	0.28	0.08	0.01	0.89	0	0.09	0.89	0.02	0.76
6	0.28	0.21	0.98	0.00	0.02	0	0.98	0.02	0.00	0.91

Scenario	$\theta_a^\star$	$\theta_w^\star$	$\mathbb{E}(N)$	$\mathbb{E}(\theta^a)$	$\mathbb{E}(\theta^w)$
1	0.10	0.10	1181.50	0.10	0.10
2	0.10	0.07	1330.00	0.10	0.07
3	0.03	0.03	1216.00	0.03	0.03
4	0.03	0.02	1385.25	0.03	0.02
5	0.28	0.28	1176.50	0.28	0.29
6	0.28	0.21	854.00	0.29	0.21

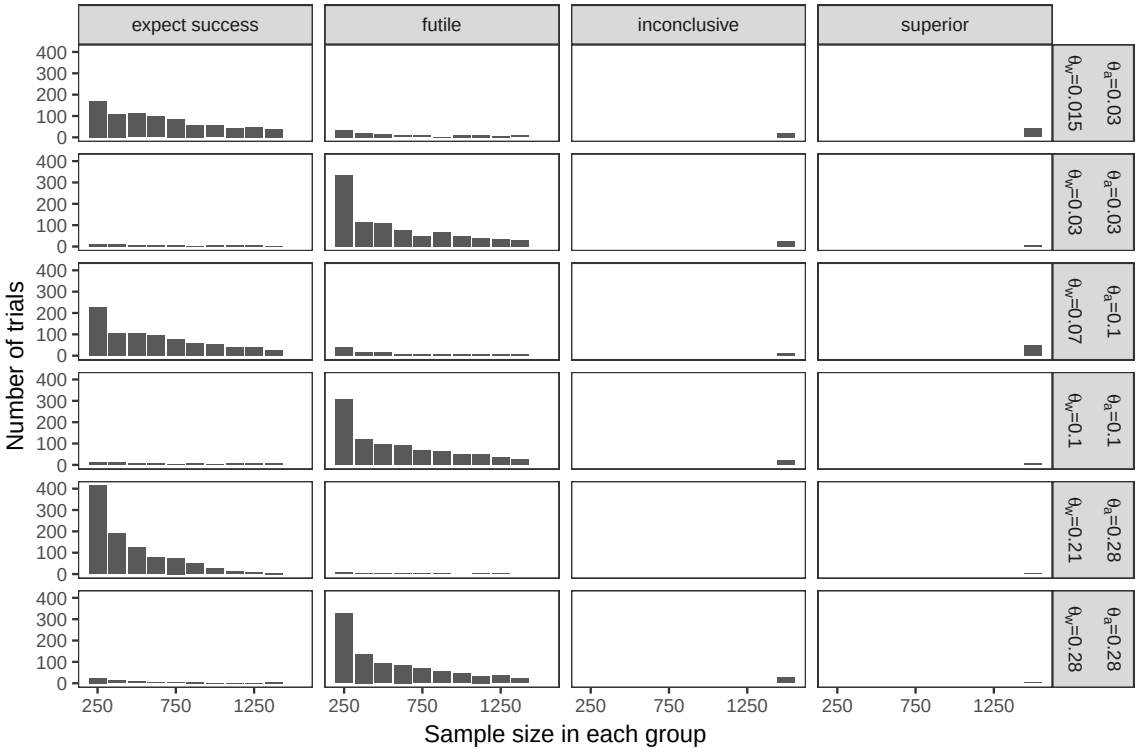


5.3 Accrual - 10 per week

Run-time was 32.3 minutes.

Scenario	$\theta_a^\star$	$\theta_w^\star$	$\mathbb{P}$(e.s)	$\mathbb{P}$(l.s)	$\mathbb{P}$(e.f)	$\mathbb{P}$(l.f)	$\mathbb{P}$(s)	$\mathbb{P}$(f)	$\mathbb{P}$(inc)	$\mathbb{P}$(s.e)
1	0.10	0.100	0.07	0.01	0.90	0	0.08	0.90	0.02	0.72
2	0.10	0.070	0.82	0.05	0.12	0	0.87	0.12	0.01	0.69
3	0.03	0.030	0.06	0.01	0.90	0	0.07	0.90	0.02	0.72
4	0.03	0.015	0.81	0.04	0.12	0	0.86	0.12	0.02	0.66
5	0.28	0.280	0.07	0.00	0.90	0	0.07	0.90	0.03	0.76
6	0.28	0.210	0.98	0.00	0.02	0	0.98	0.02	0.00	0.90

Scenario	$\theta_a^\star$	$\theta_w^\star$	$\mathbb{E}(N)$	$\mathbb{E}(\theta^a)$	$\mathbb{E}(\theta^w)$
1	0.10	0.100	1226.25	0.10	0.11
2	0.10	0.070	1317.50	0.11	0.07
3	0.03	0.030	1191.25	0.03	0.03
4	0.03	0.015	1406.00	0.03	0.02
5	0.28	0.280	1148.00	0.27	0.29
6	0.28	0.210	904.50	0.29	0.21



6 Simulation Details

6.1 Beta Inequalities

In the two arm case we are generally interested in at least one of the following equivalent probabilities

$$\begin{aligned}
\mathbb{P}_{X,Y}(X > Y + \delta) &= \int_{\delta}^{1} \int_{0}^{X-\delta} f(y)dy f(x)dx \\
&= \int_{\delta}^{1} f_X(x) F_Y(x-\delta)dx \\
&= 1 - \mathbb{P}_{X,Y}(X < Y + \delta) \\
\mathbb{P}_{X,Y}(Y < X - \delta) &= \int_{0}^{1-\delta} \int_{Y+\delta}^{1} f(x)dx f(y)dy \\
&= \int_{0}^{1-\delta} f_Y(y)(1 - F_X(y+\delta))dy \\
&= 1 - \mathbb{P}_{X,Y}(Y > X - \delta)
\end{aligned}$$

where $X \sim \text{Beta}(a, b)$ and $Y \sim \text{Beta}(c, d)$ are independent Beta distributions. The probability of the event, $X > Y + \delta$, cannot be caluclated analytically, but can do done so using numerical integration over a univariate integral (for reasonable values of the parameters).

In the interest of speed we might alternatively approximate the Beta distributions by Normal distributions. The approximation should be satisfactory if $\frac{a+1}{a-1} \approx 1$ and $\frac{b+1}{b-1} \approx 1$ in which case

$$\text{Beta}(a, b) \sim N\left(\frac{a}{a+b}, \frac{ab}{(a+b)^2(a+b+1)}\right).$$

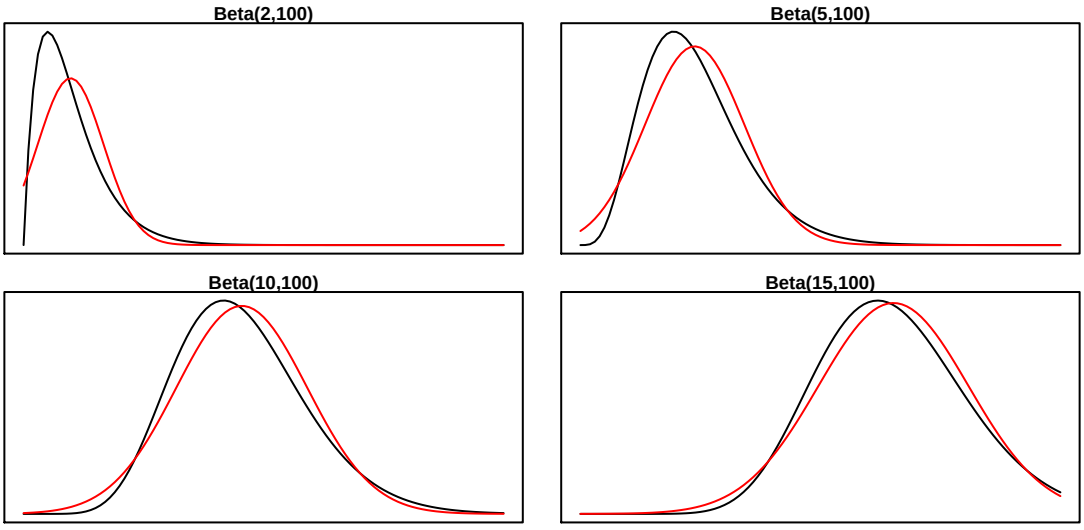


Figure 3: Example Normal approximation to Beta densities.

Then we estimate the inequality by

$$
\begin{aligned}
m_X &= \frac{a}{a+b} \\
s_X^2 &= \frac{ab}{(a+b)^2(a+b+1)} \\
m_Y &= \frac{c}{c+d} \\
s_Y^2 &= \frac{cd}{(c+d)^2(c+d+1)} \\
z &= \frac{m_X - m_Y - \delta}{\sqrt{s_X^2 + s_Y^2}} \\
\mathbb{P}_{X,Y}(X > Y + \delta) &\approx \Phi(z)
\end{aligned}
$$

expression	min	mean	median	max	itr/sec
beta_ineq_approx(3, 100, 13, 90)	6.56us	8.08us	7.66us	105us	123740.90
beta_ineq_sim(3, 100, 13, 90, sims = 1000)	408.07us	444.55us	424.11us	875us	2249.49
beta_ineq(3, 100, 13, 90)	478.45us	541.36us	512.73us	999us	1847.21

Approximation is also reasonably accurate for most parameter settings, in the worse case, it is no worse than the error which may occur using Monte Carlo estimate with 1,000 particles.

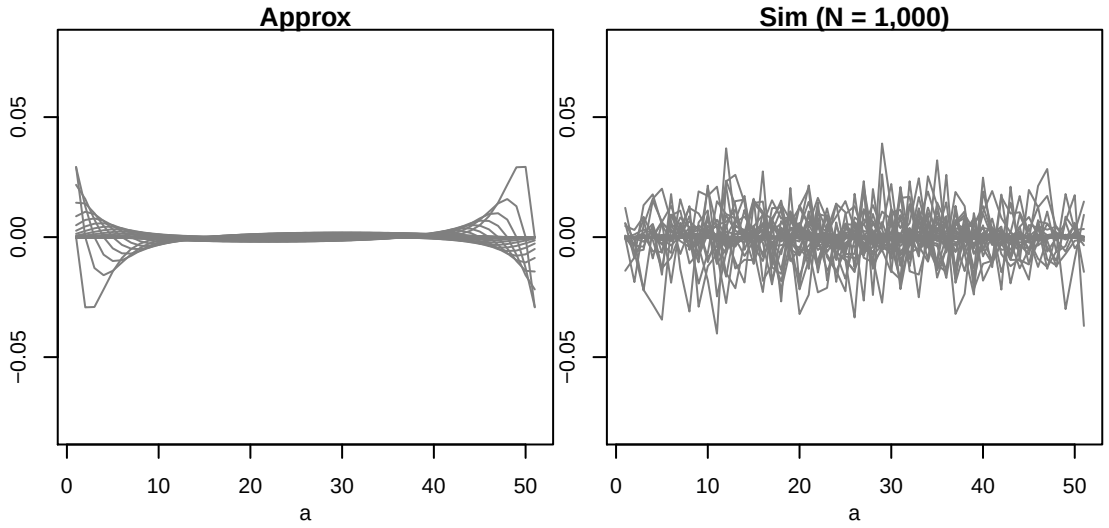


Figure 4: Deviation from exact value (adaptive quadrature) of $\mathbb{P}(X > Y + \delta)$.

A trade-off may be to use exact or simulation methods for parameter values where the approximation is known to be poor, and use the approximation otherwise.

6.2 Posterior Predictive Probabilities

To compute the predictive probability of success we take the expectation of an indicator function with respect to the posterior predictive distribution of the joint outcomes. In the two arm case this is a double summation over the domain.

We can:

- Compute exactly by enumerating over all $0 : \tilde{n}_k^a$ and $0 : \tilde{n}_k^w$ and compute $\tilde{P}_k$ for every value, however for large n_k^i this becomes computationally intensive.
- Use Monte Carlo estimates by drawing $\tilde{y}_k^{i,j} \sim \tilde{f}_k(\tilde{y}_k^i | y_k^i), j = 1, ..., N$ for each i and average the values of $\mathbb{I}\{\tilde{P}_k > q\}$, noting that we can probably just estimate the tail probability once for each unique combination of $(\tilde{y}_k^{a,j}, y_k^{w,j})$ and scale by the number of occurences, reducing the number of $\tilde{P}_k$ we need to compute.
- Pre-determine the values which have relatively large contribution to the posterior predictive density (e.g. say within 10^{-6} of the largest probability) and only compute $\tilde{P}_k$ for these values, noting that this will slightly under-estimate the probability by not by much more than 10^{-6}.

For small sample sizes should just enumerate over all values, but for larger predicted sample sizes use Monte Carlo.

Some notes:

- $\text{PPoS}_k \leq P_k$
- $\tilde{n} \to \infty \implies \text{PPoS}_k \to P_k$?