# Bayesian Adaptive Designs for Clinical Trials

Jason Connor

ConfluenceStat

Jason@ConfluenceStat.com

412-860-3113

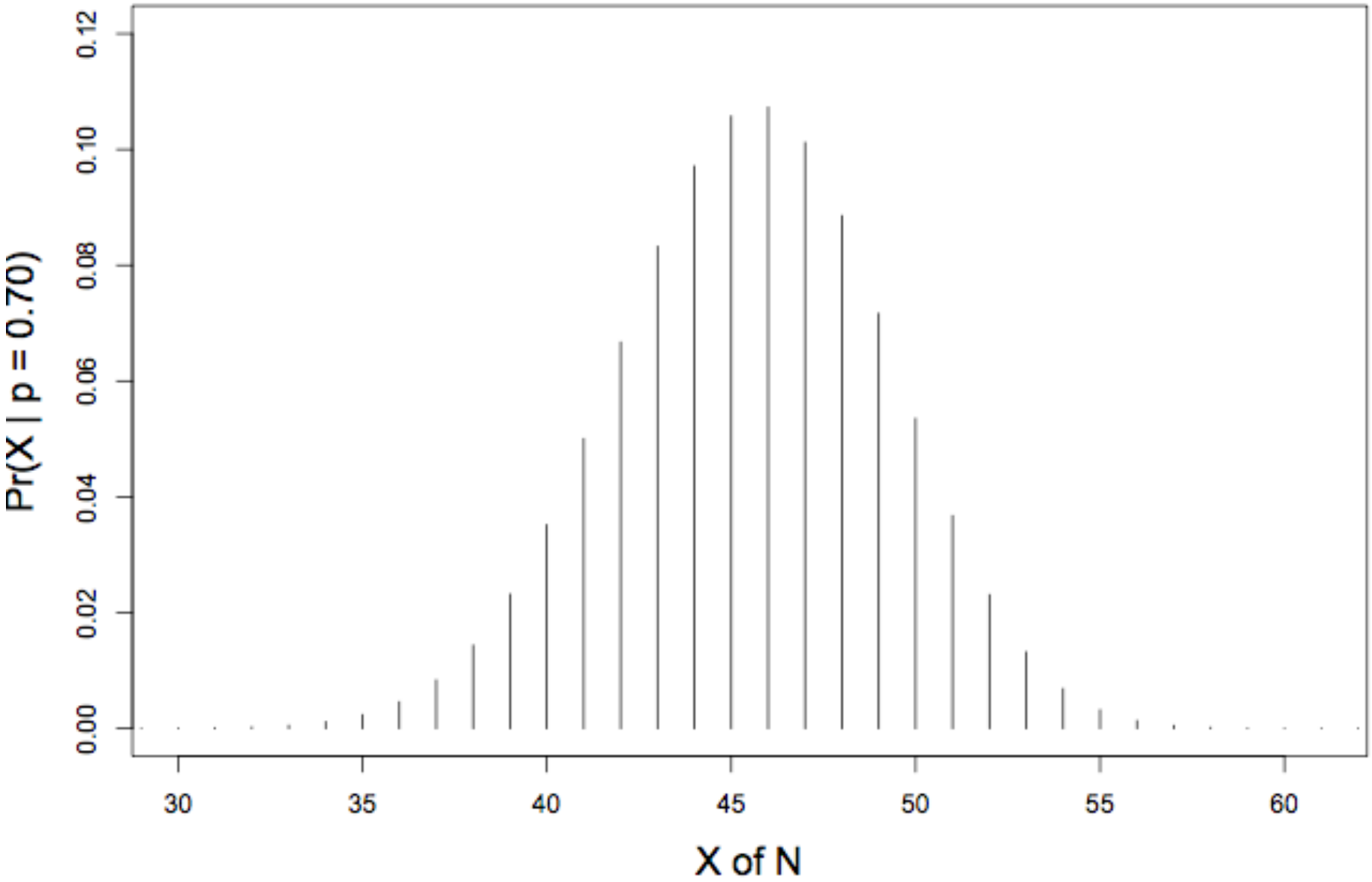
Phase 3

# Why be adaptive on sample size?

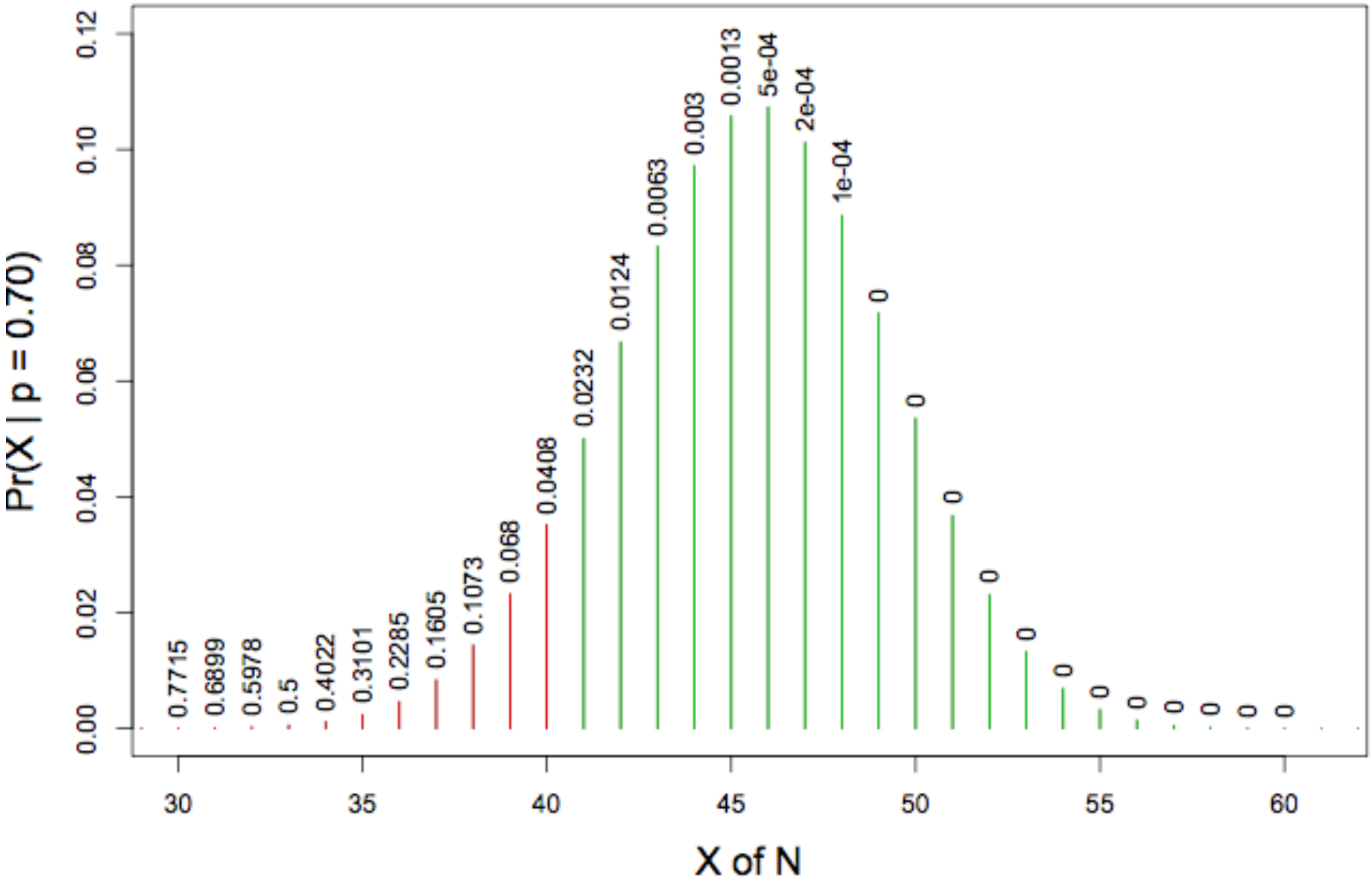
- Doctor comes to you.
- Well documented historical success rate = 50%
- Claims his therapy has 70% success
- "How many patients do I need to be statistically significant?"

N = 65 gives 90% power to reject Ho: p = 0.50 when p = 0.7

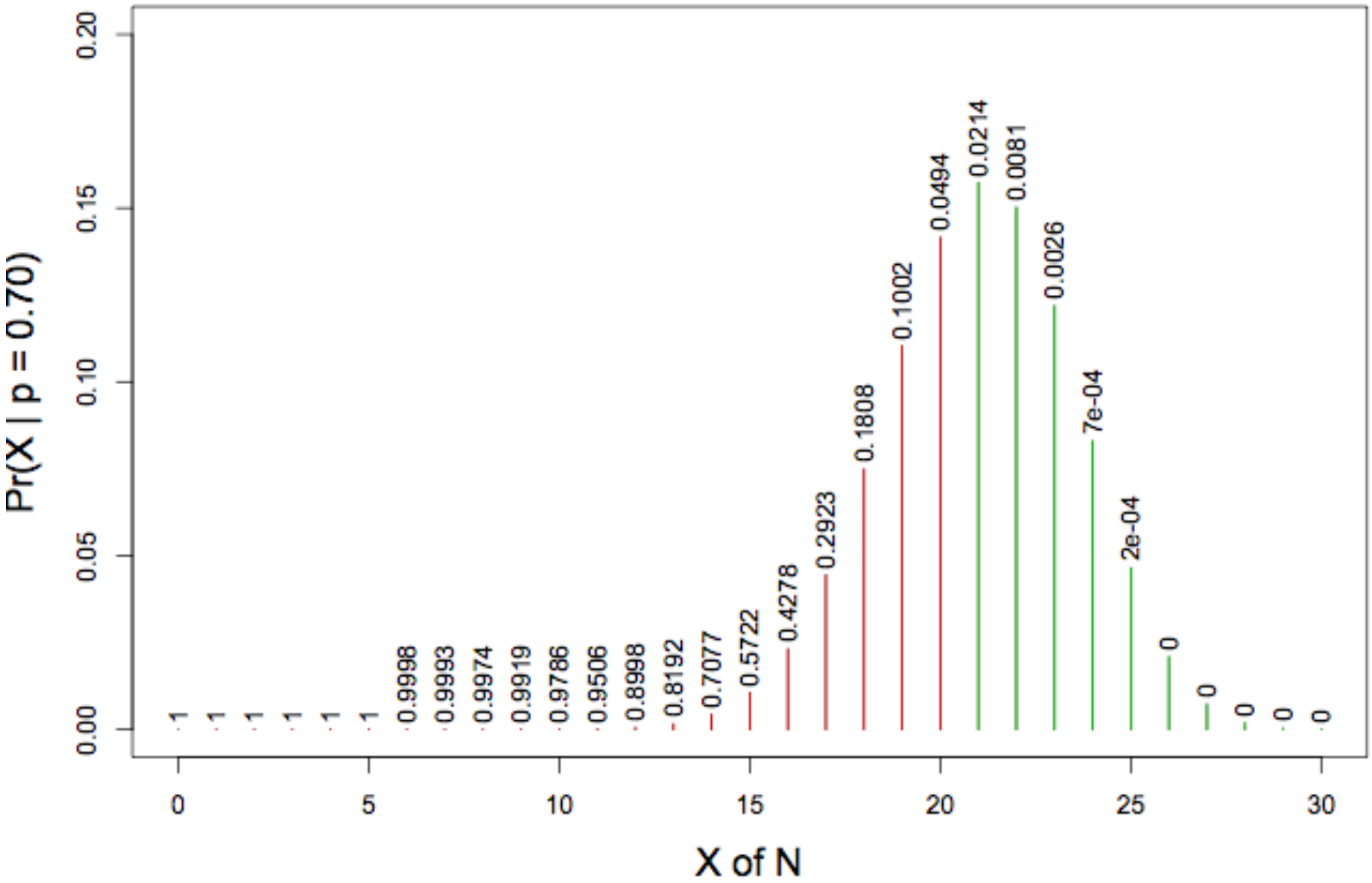
N = 65 gives 90% power to reject Ho: p = 0.50 when p = 0.7



# N = 65 gives 90% power to reject Ho: p = 0.50 when p = 0.7

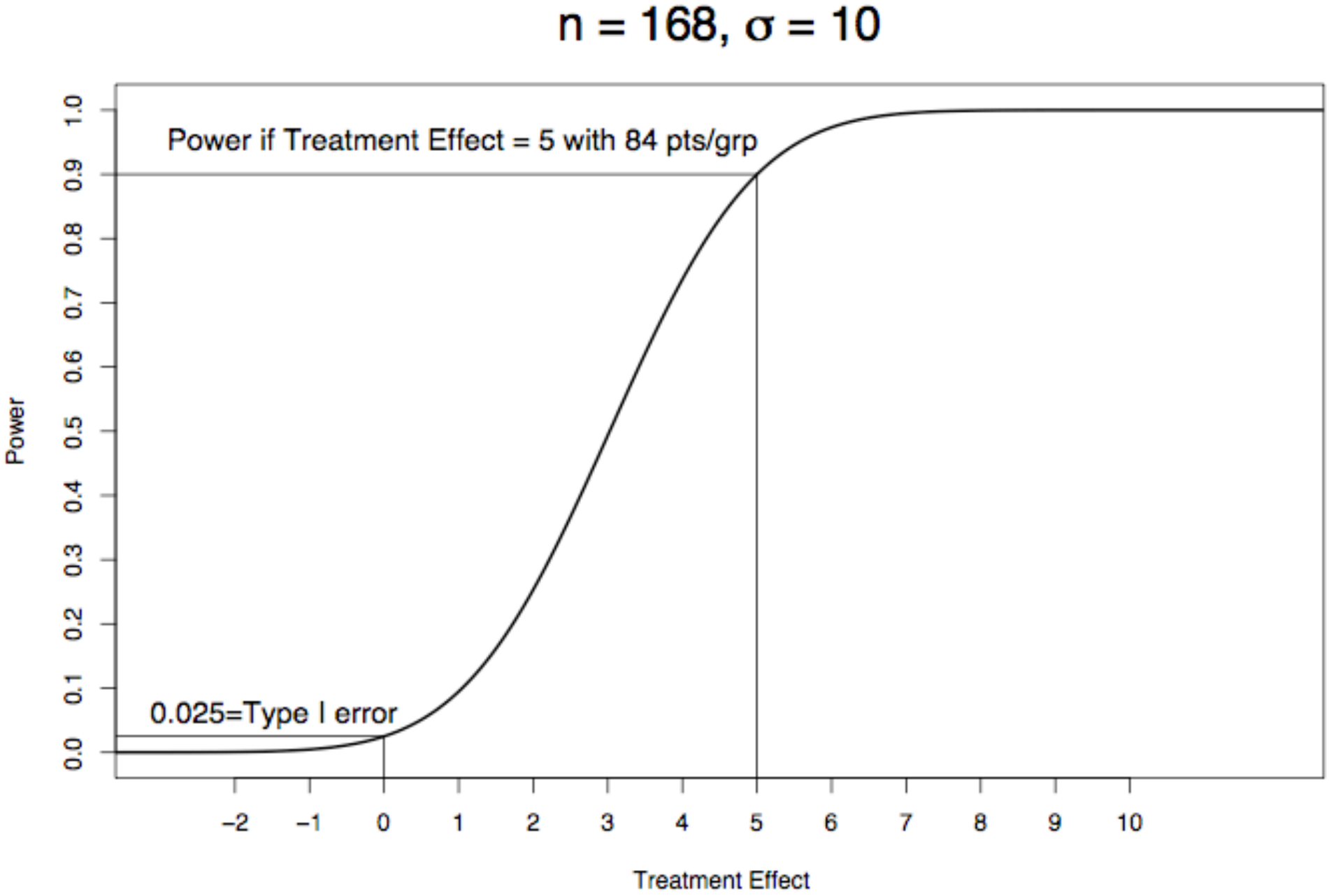


# If observed = 70% only need N = 30 not N=65!

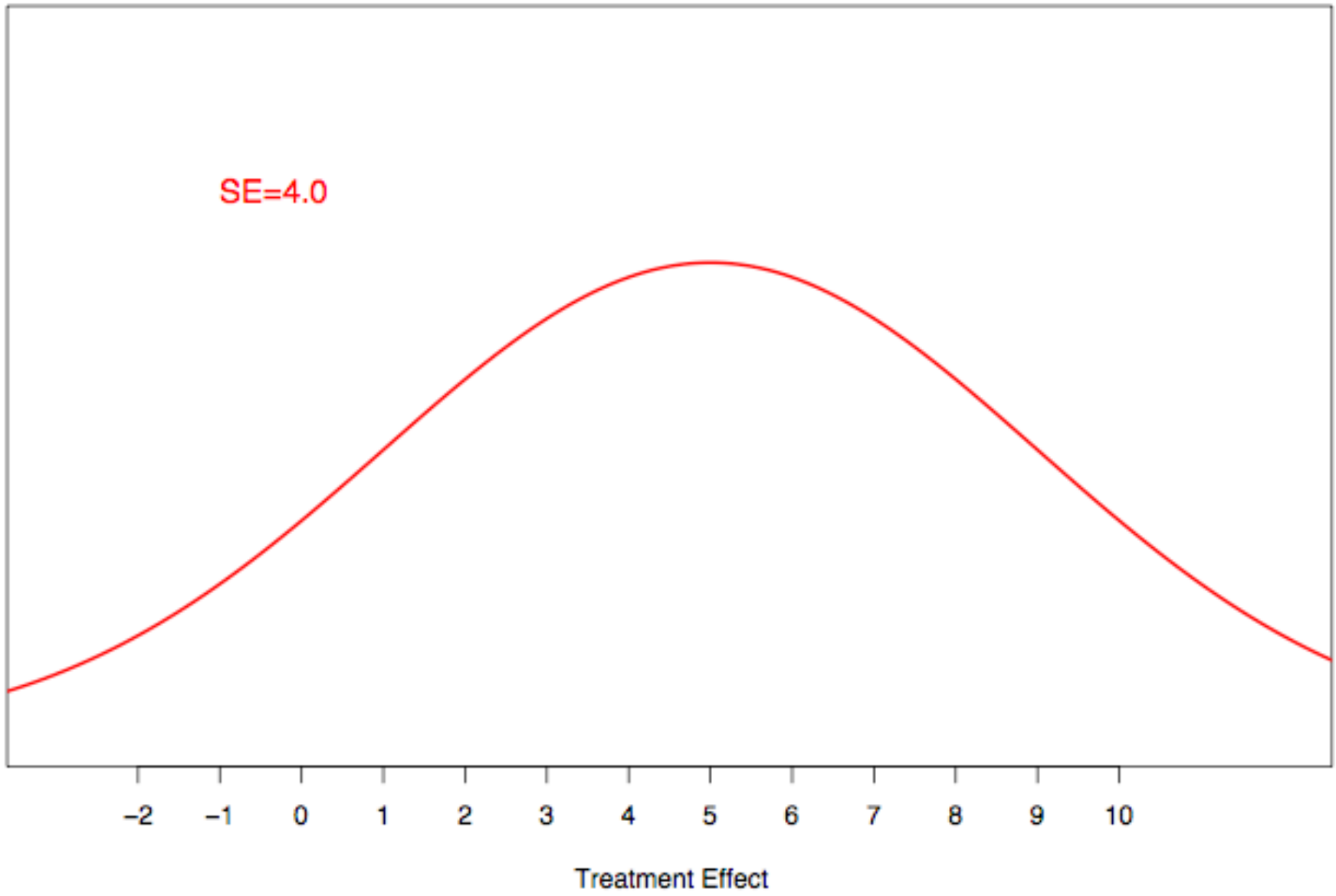


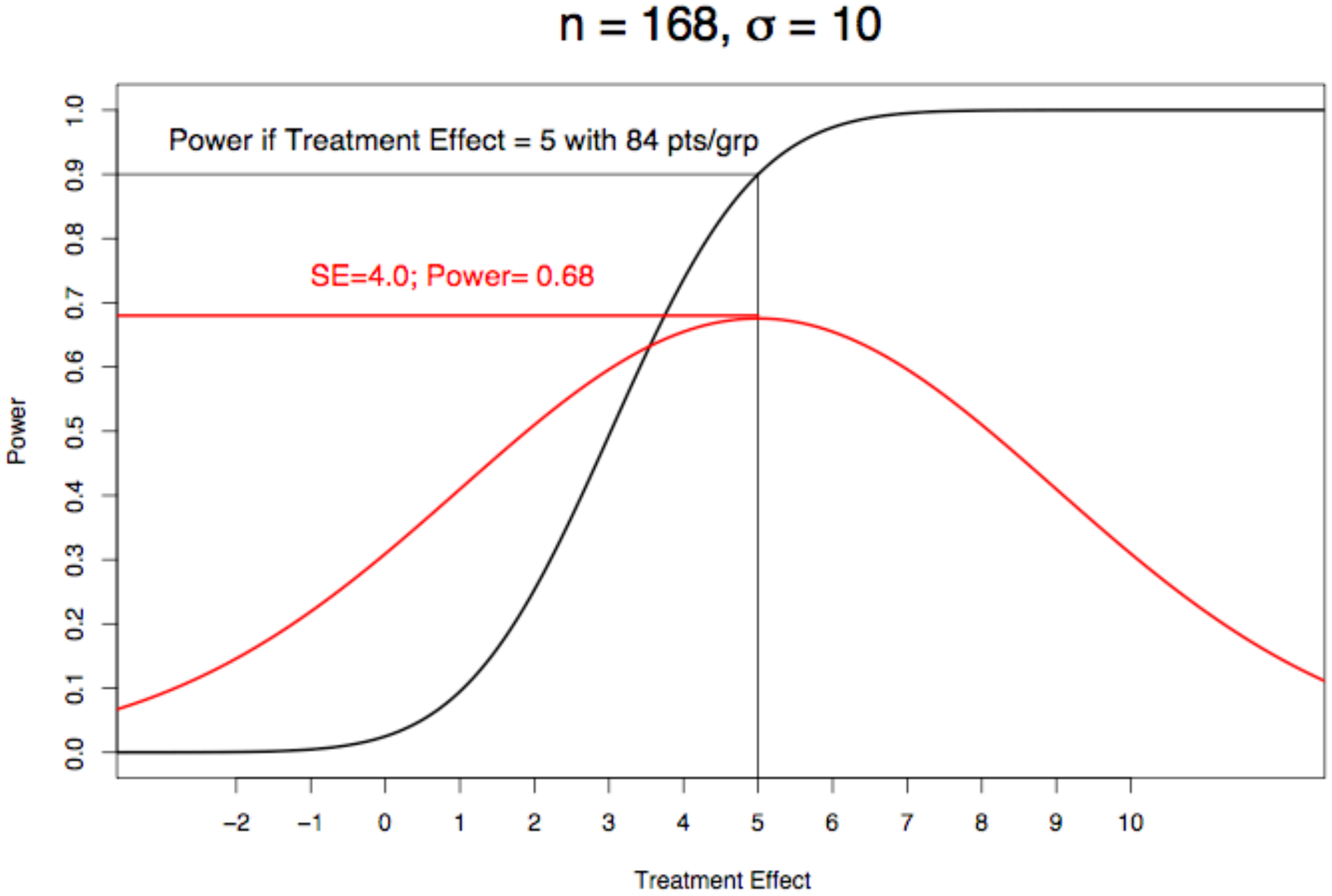
# Why be adaptive?

- Doctor comes to you.
- Claims her treatment increases IQ by 5 points
- SD = 10
- "How many patients do I need to have 90% power to demonstrate superiority?"


n = 168, σ = 10
Power if Treatment Effect = 5 with 84 pts/grp
0.025=Type I error
Power
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
−2
−1
0
1
2
3
4
5
6
7
8
9
10
Treatment Effect

# We've ignored the error in the pilot data




n = 168, σ = 10
Power if Treatment Effect = 5 with 84 pts/grp
SE=4.0; Power= 0.68
Power
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
−2
−1
0
1
2
3
4
5
6
7
8
9
10
Treatment Effect

# Phase 3 / Confirmatory Trials

- CDER/CBER Phase 3
- CDRH Confirmatory
- The final test before market
- Control of Type I error rate very important
- Tend not to adaptively randomize
  - Fear of drift
  - Usually two arm
  - No power benefit with adaptive rand. in 2-arm trial

# What is Different About Confirmatory Trials

- Type I error is a dominant factor
- Adjusting the design (goal) in order to accommodate adaptive aspects must still control type I error
- Predictive probabilities much more relevant than posterior probabilities
- Very well-defined goal.
  - A "game" you win or lose

# Confirmatory Trials & Bayes

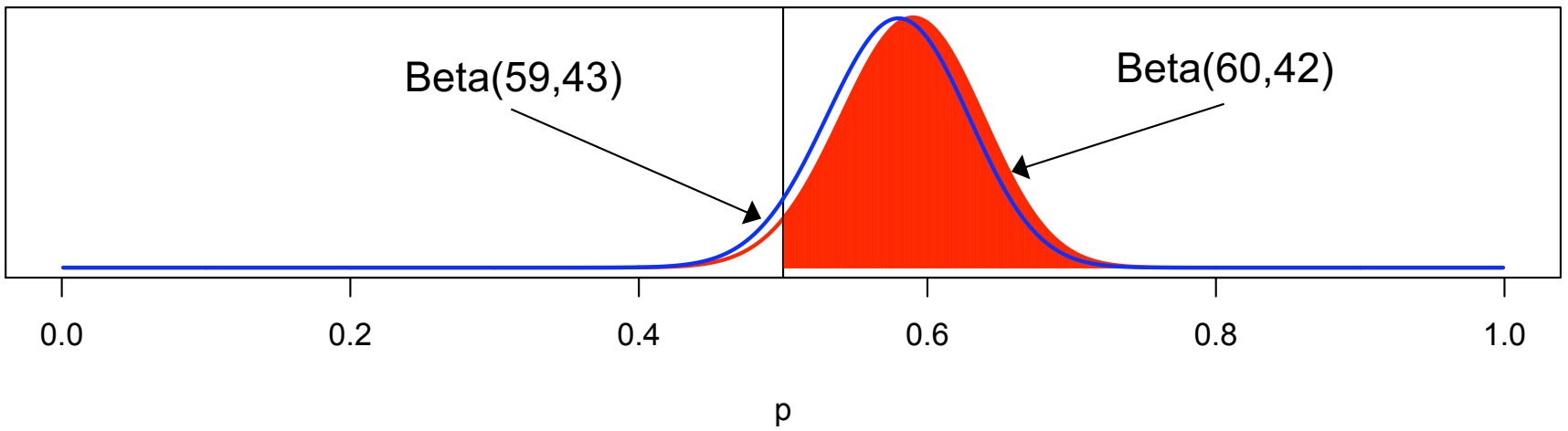
- You can't have an informative prior and control Type I error
  - Assuming the informative prior claims the treatment starts off better than the control
  - Kopp-Schneider, Calderazzo, & Wiesenfarth, Biometric Journal, 2019.

# Simple Trial

- Binomial data
- One-armed trial
- $n = 100$
- Need to show $p > 0.5$
- $H_o$: $p \leq 0.5$
- $H_a$: $p > 0.5$

# Phase 3 & Priors

- Simple Trial:
  - Binary data. Observe x ~ Bin(100,p)
  - Need to show Pr(p > 0.5 | x out of 100) > 0.95
  - Assume p ~ Beta(1,1) prior
  - Pr(p > 0.5 | 59 out of 100) = 0.963
  - Pr(P > 0.5 | 58 out of 100) = 0.944



# Phase 3 & Priors

- Simple Trial:
  - Binary data. Observe $x \sim Bin(100,p)$
  - Need to show $Pr(p > 0.5 \mid x \text{ out of } 100) > 0.95$
  - Assume $p \sim Beta(1,1)$ prior
  - $Pr(p > 0.5 \mid 59 \text{ out of } 100) = 0.963$
  - $Pr(P > 0.5 \mid 58 \text{ out of } 100) = 0.944$
- $Pr(X \geq 59 \mid p = 0.50) = 0.044$
  - Simple binomial calculation
  - This is Type I error and is < 5%
  - Bayesian trial
  - Good frequentist properties

# Phase 3 & Priors

- Now we have "pure" information

  9 of 10 successes in European trial

- Instead of Beta(1,1) prior use Beta(1+9, 1+1)

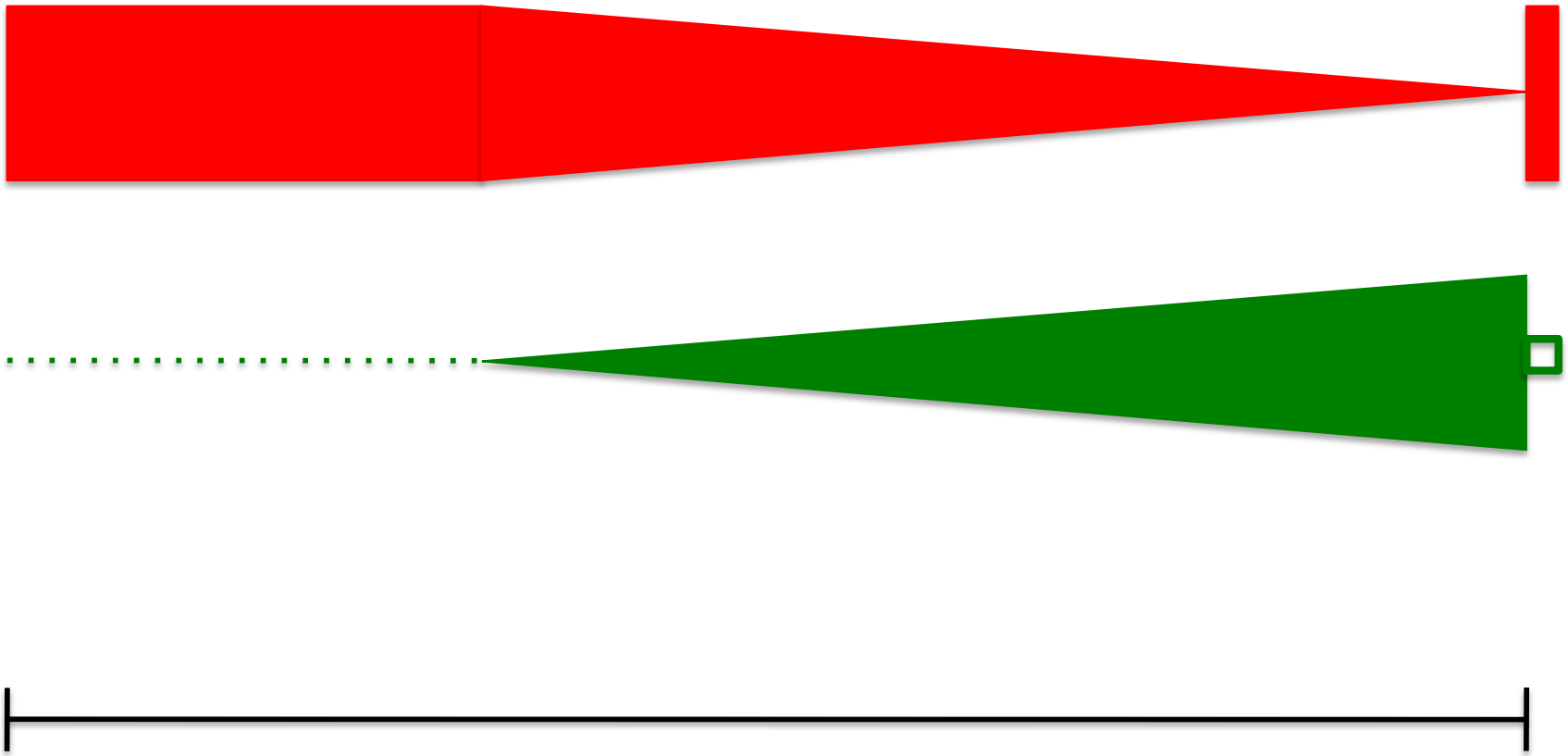
  = Beta(10,2) prior

- Regulatory agrees it is reasonable to use this as the prior
- Fixed design: for $Pr[p > 0.5 \mid \text{data}] \geq 0.95$
  - $Pr(p > 0.5 \mid 55 \text{ out of } 100, \alpha=10,\beta=2) = 0.956$
  - $Pr(P > 0.5 \mid 54 \text{ out of } 100, \alpha=10,\beta=2) = 0.936$
- $Pr(X \geq 55 \mid p = 0.50) = 0.184$
  - Type I error is inflated

# Phase 3 & Priors

- Solution to control Type I error
  - Raise the post probability threshold from 0.95 bar to 0.99 (like decreasing critical level)
  - $Pr(p > 0.5 \mid 59 \text{ out of } 100, \alpha=10,\beta=2) = 0.993$
  - $Pr(P > 0.5 \mid 58 \text{ out of } 100, \alpha=10,\beta=2) = 0.989$
  - $Pr(X \geq 59 \mid p = 0.50) = 0.044$
- Need a Beta(59+10,41+2) for a win…59 is back!!!
- The type I error "restriction" forces 59/100 regardless of prior…
- Can't allow beneficial priors AND force Type I of "new" experiment!

# Posterior/Predictive

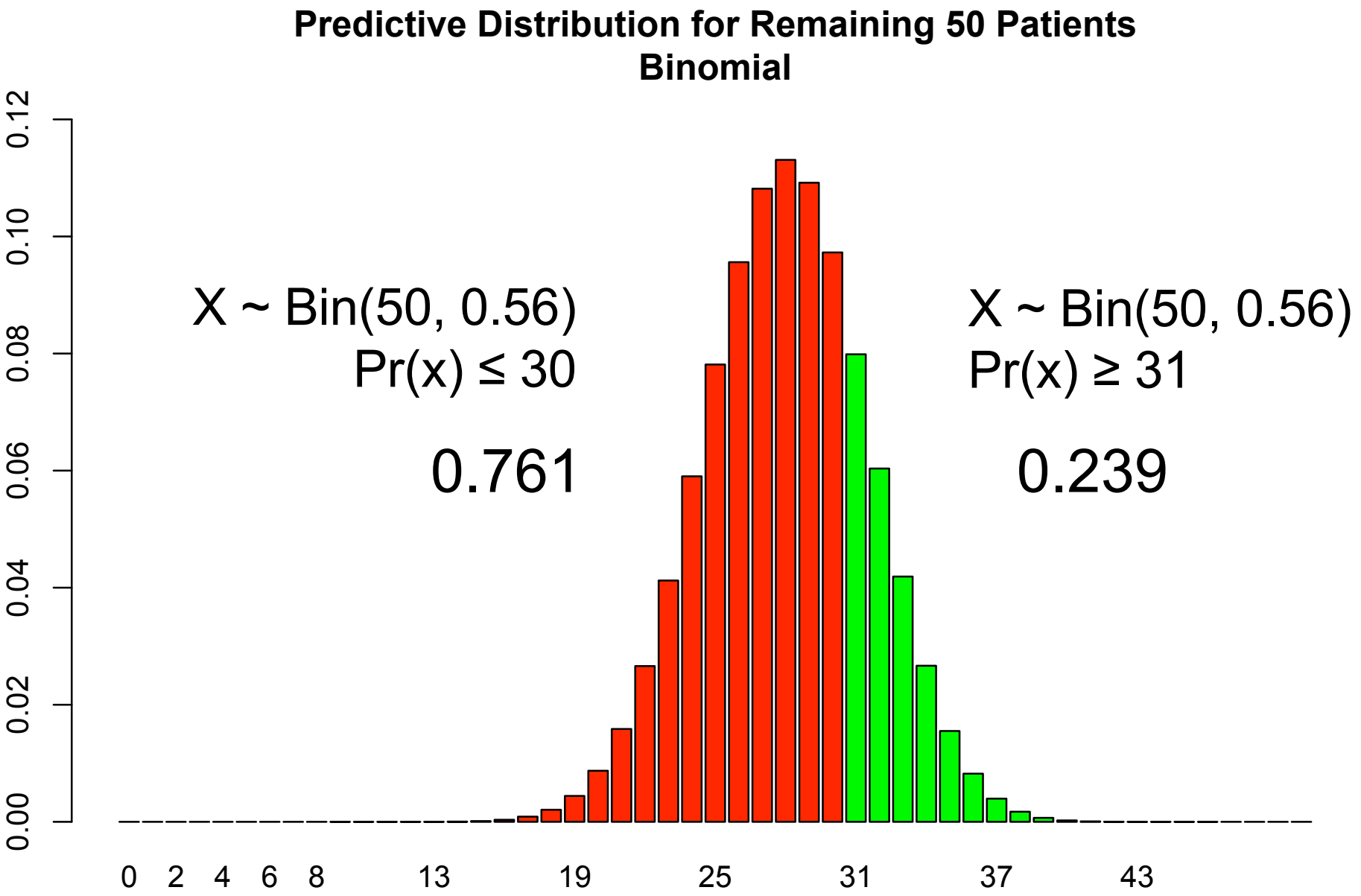


# Predictive Probabilities

- 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
  - $\Pr(p > 0.5 \mid 59 \text{ out of } 100) = 0.963$
  - $\Pr(p > 0.5 \mid 58 \text{ out of } 100) = 0.944$

# Predictive Probabilities

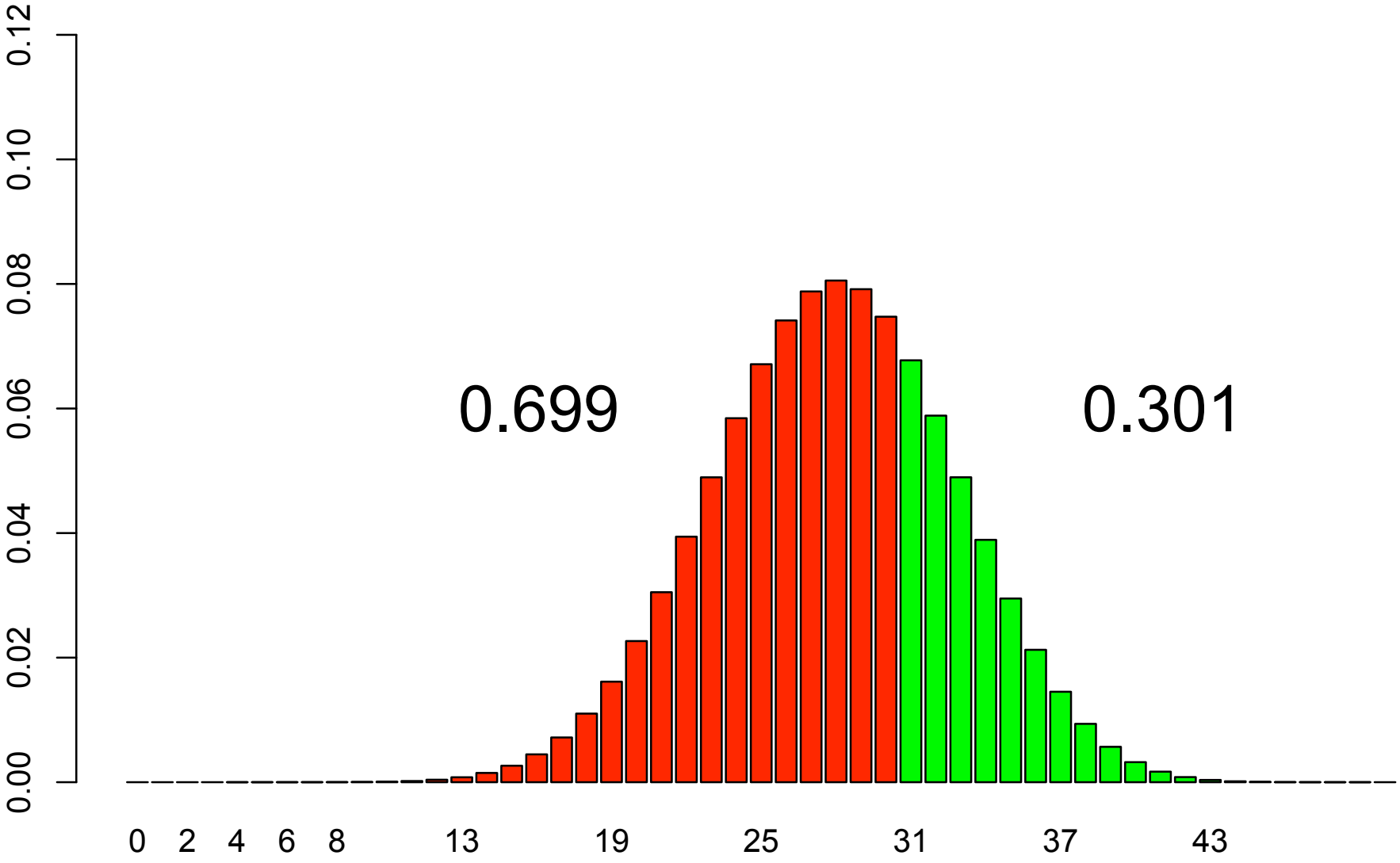
- Simple Trial:
  - Binary data. Observe $x \sim \text{Bin}(100, p)$
  - Need to show $\Pr(p > 0.5 \mid x \text{ out of } 100) > 0.95$
  - Assume $p \sim \text{Beta}(1,1)$ prior
  - $\Pr(p > 0.5 \mid 59 \text{ out of } 100) = 0.963$
  - $\Pr(p > 0.5 \mid 58 \text{ out of } 100) = 0.944$
- Observe data half way through
  - See 28/50 successes
  - Need to see 31/50 to meet threshold
  - What is predictive probability of trial success?

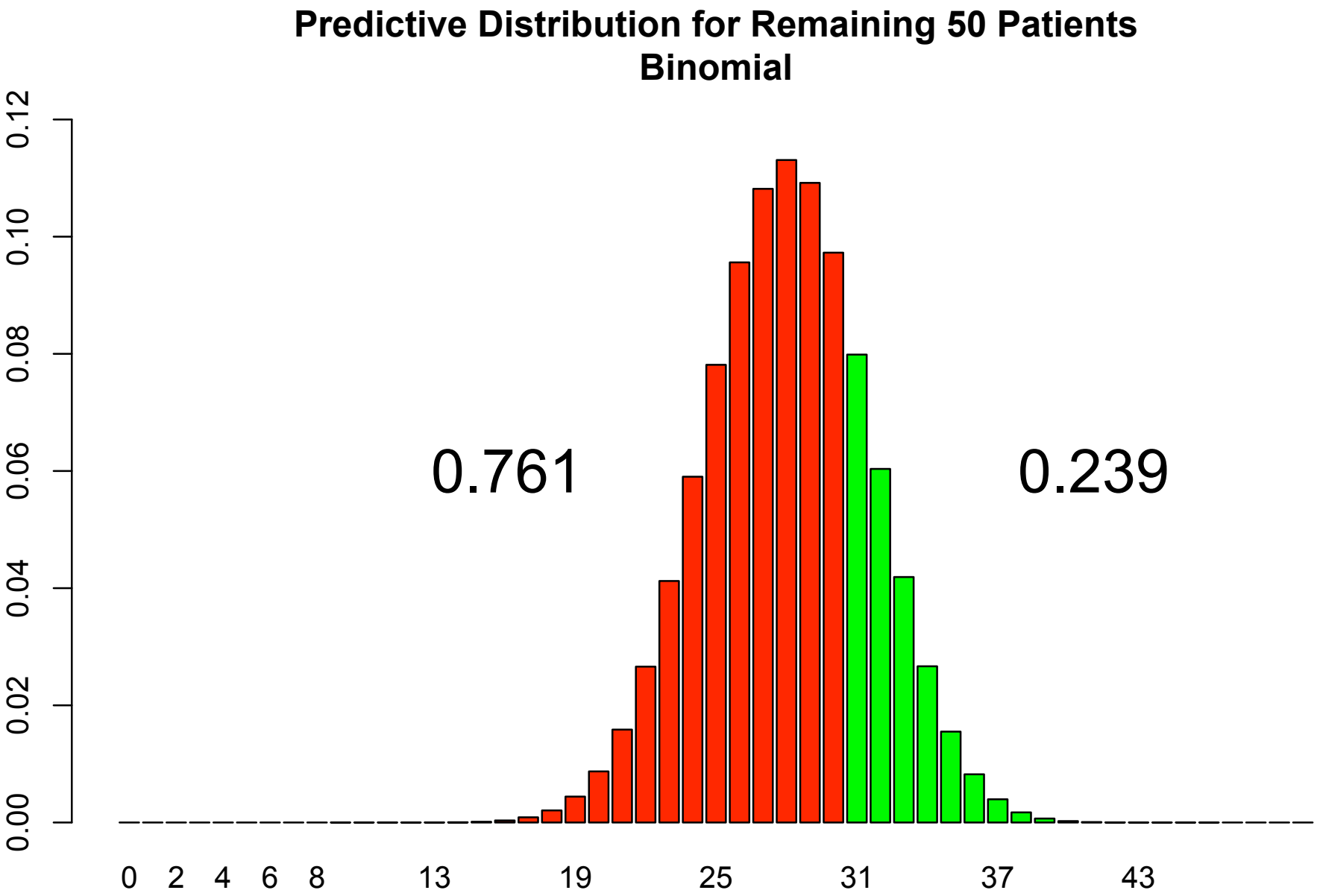

Predictive Distribution for Remaining 50 Patients
Binomial
X ~ Bin(50, 0.56)
Pr(x) ≤ 30
0.761
X ~ Bin(50, 0.56)
Pr(x) ≥ 31
0.239
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0
2
4
6
8
13
19
25
31
37
43

# Predictive Probabilities

- Know we need $x \geq 59$ at trial's end
- Have $x_1 = 28$
- Need $x_2 \geq 31$
- $p \sim$ Beta(1+28, 1+22)
- $x_2 \sim$ Binomial(50, $p$)
- $x_2 \sim$ Beta-binomial(50, $\alpha$=29, $\beta$=23)

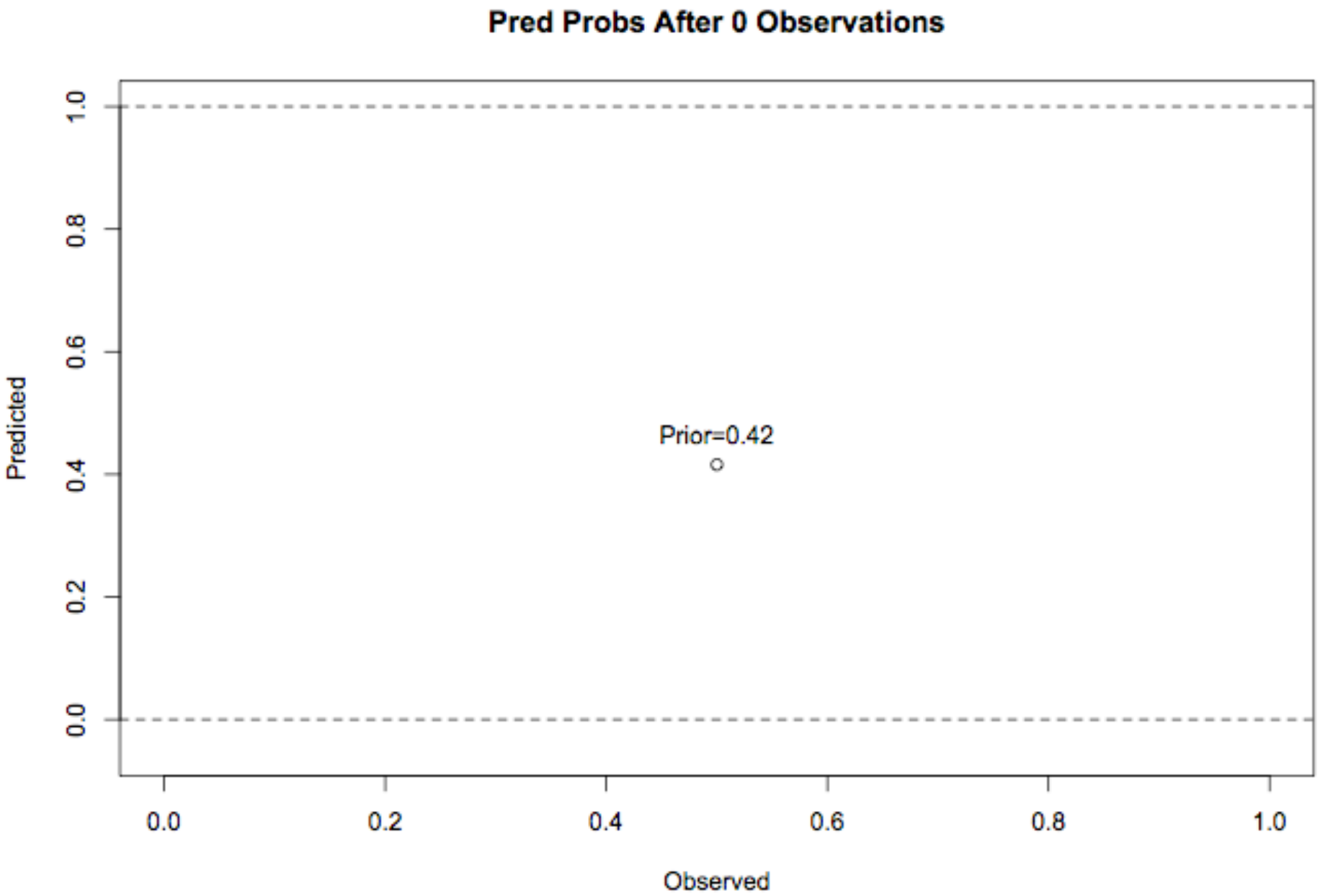
$$\Pr(\text{Win Trial}) = \sum_{x_2=31}^{50} \left\{ \binom{50}{x_2} \frac{B(x_2 + 29, 50 - x_2 + 23)}{B(29,22)} \right\} = 0.301$$

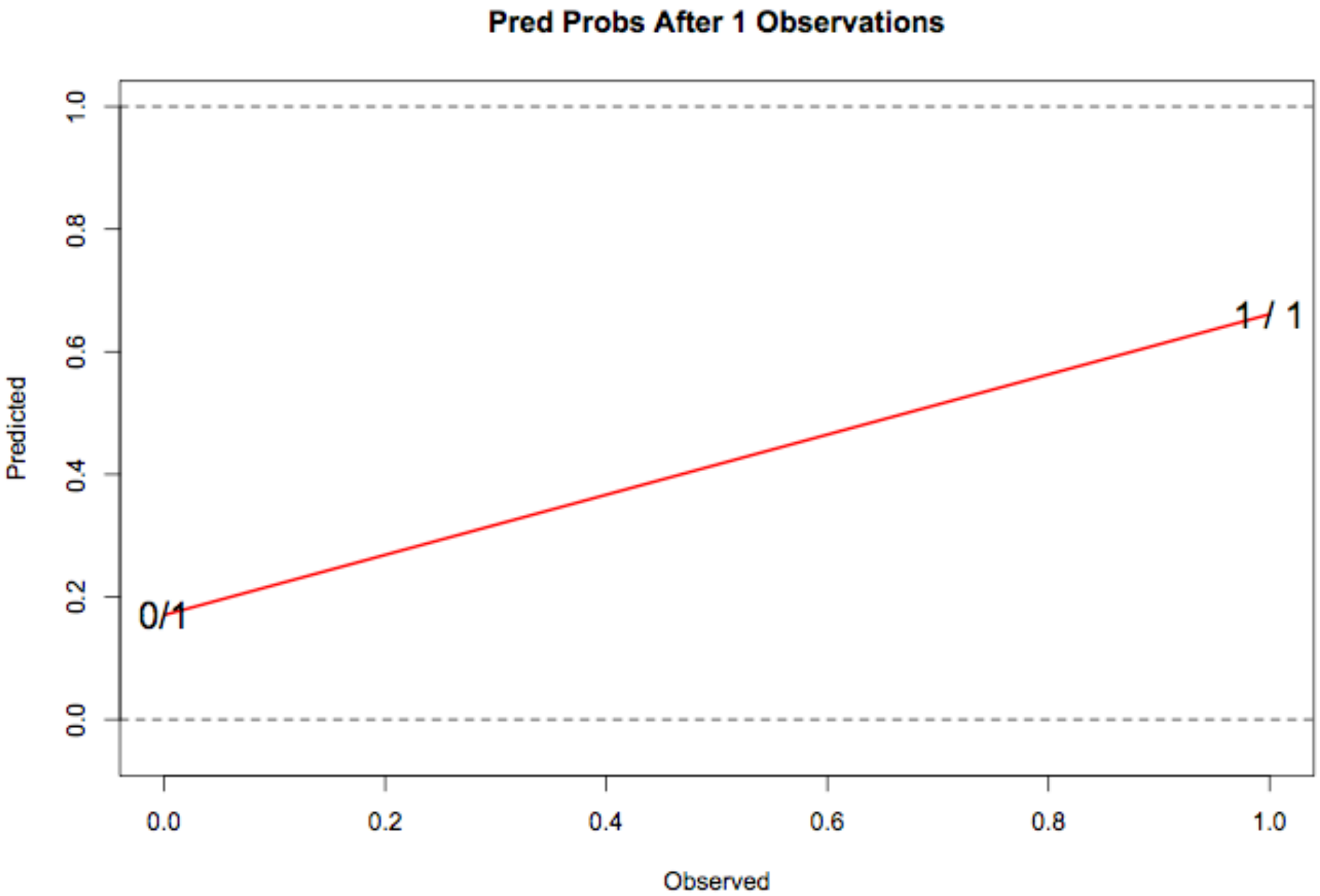

Predictive Distribution for Remaining 50 Patients
Beta-Binomial
0.699
0.301
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0
2
4
6
8
13
19
25
31
37
43

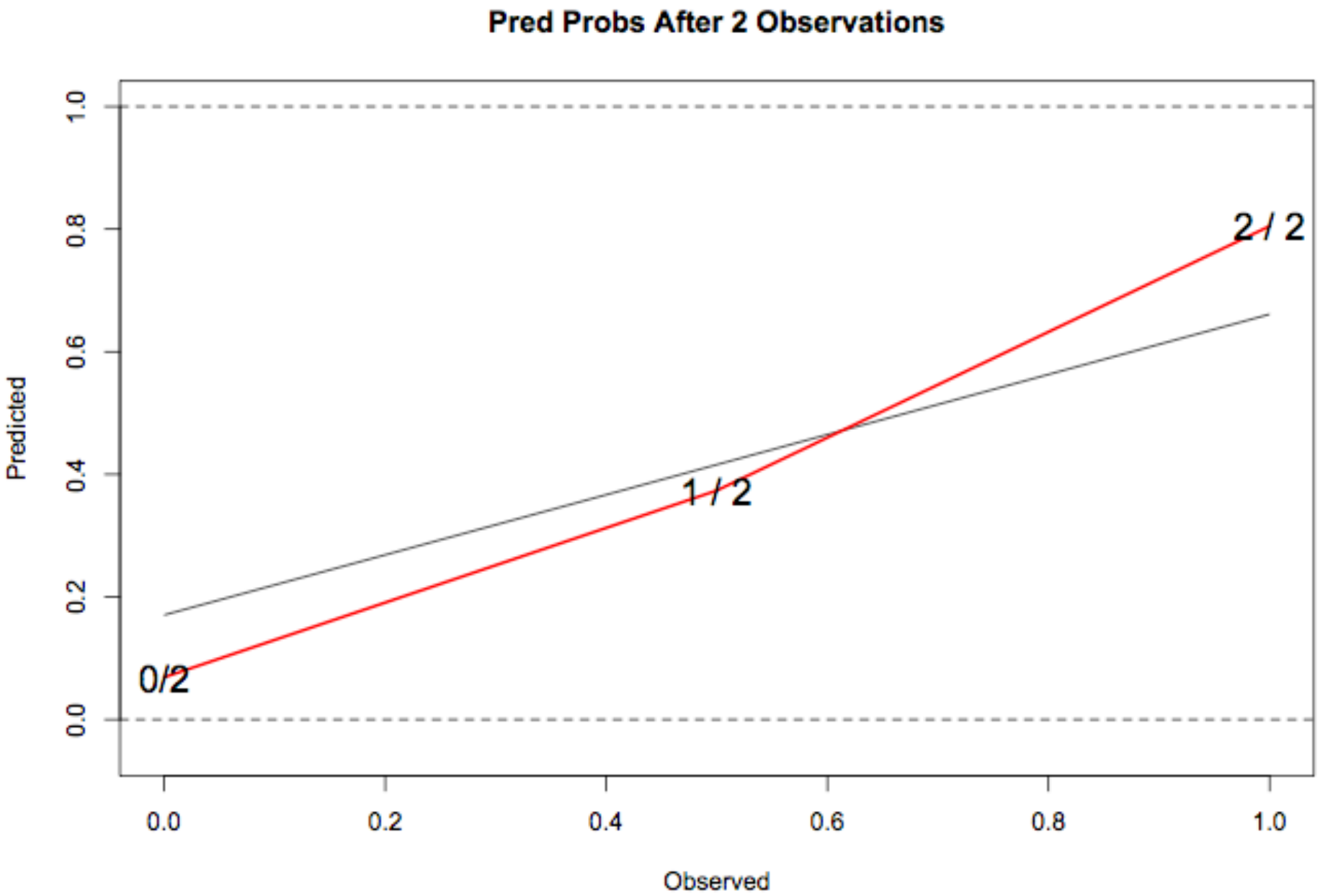

Predictive Distribution for Remaining 50 Patients
Binomial
0.761
0.239
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0
2
4
6
8
13
19
25
31
37
43

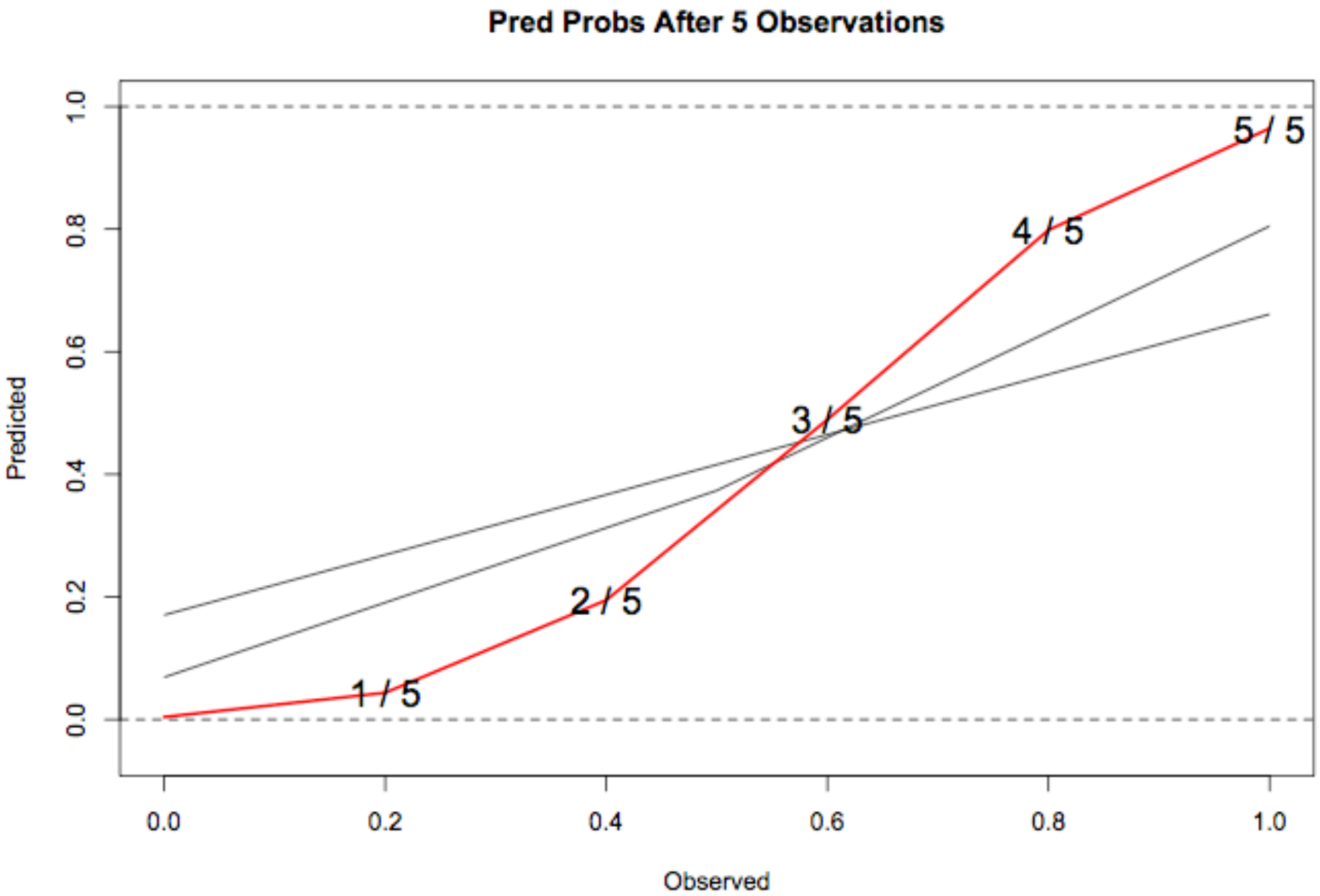
# R code for predictive probability

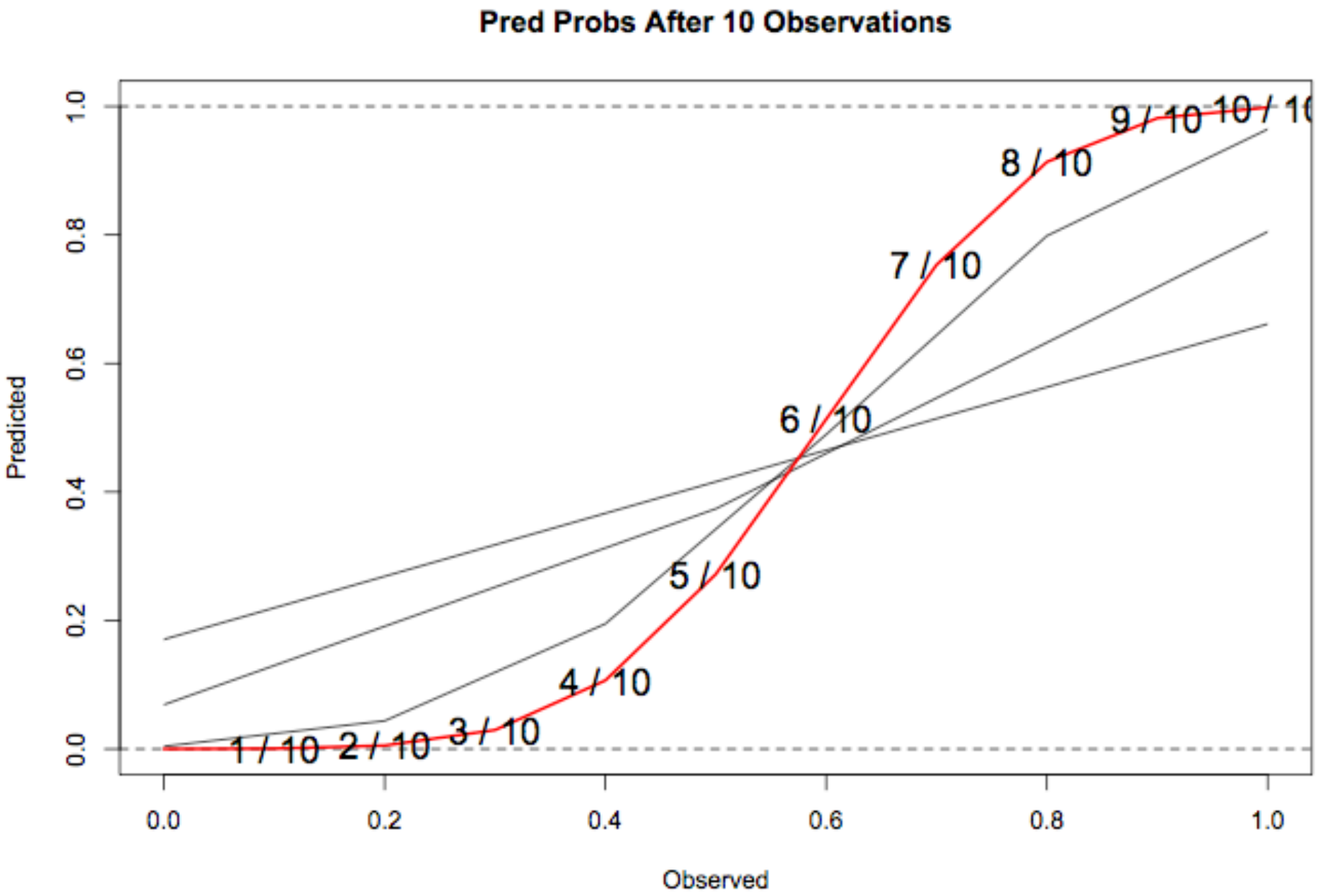
```
> ### VIA SIMULATION
> alpha <- 1; beta <- 1
> x <- 28; N <- 50
>
> p <- rbeta(1000000, alpha+x, beta+N-x)
> x.new <- rbinom(1000000, 50, p)
>
> mean(x.new >= 31)
[1] 0.301132
>
>
>
> ### VIA DIRECT CALCULATION
> N.new <- 50
> x.new <- 0:50
> prob <- choose(N.new,x.new) *
+       beta(alpha+x+x.new,(beta+N-x)+(N-x.new)) /
+          beta(alpha+x,       (beta+N-x))
> sum(prob)
[1] 1
> sum(prob[x.new >= 31])
[1] 0.3010906
> barplot(prob, names.arg=0:50, col=c(rep(2,31), rep(3,20)),
+        main="Predictive Distribution for Remaining 50 patients")
```

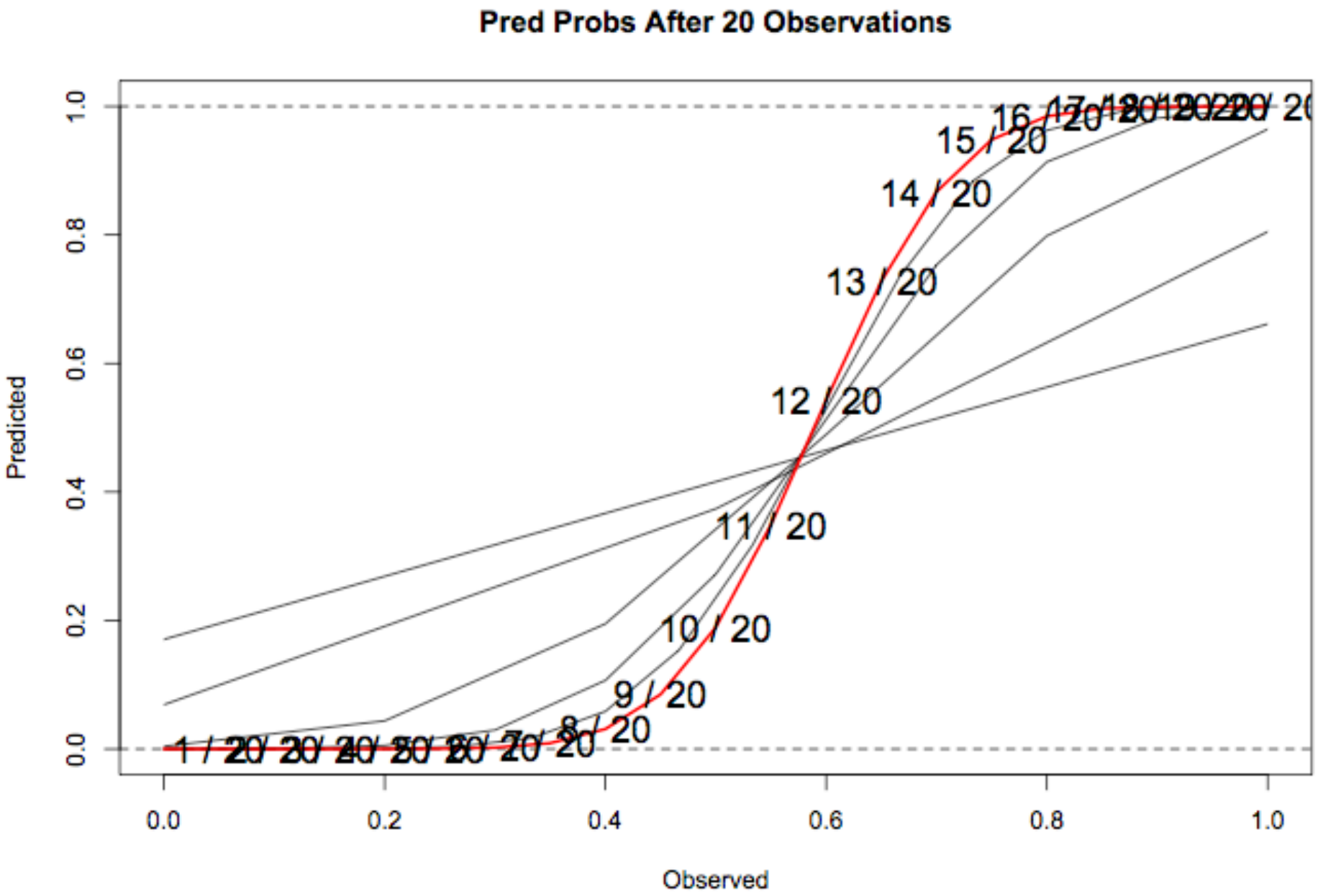

Pred Probs After 0 Observations
Predicted
1.0
0.8
0.6
0.4
0.2
0.0
Prior=0.42
0.0
0.2
0.4
0.6
0.8
1.0
Observed

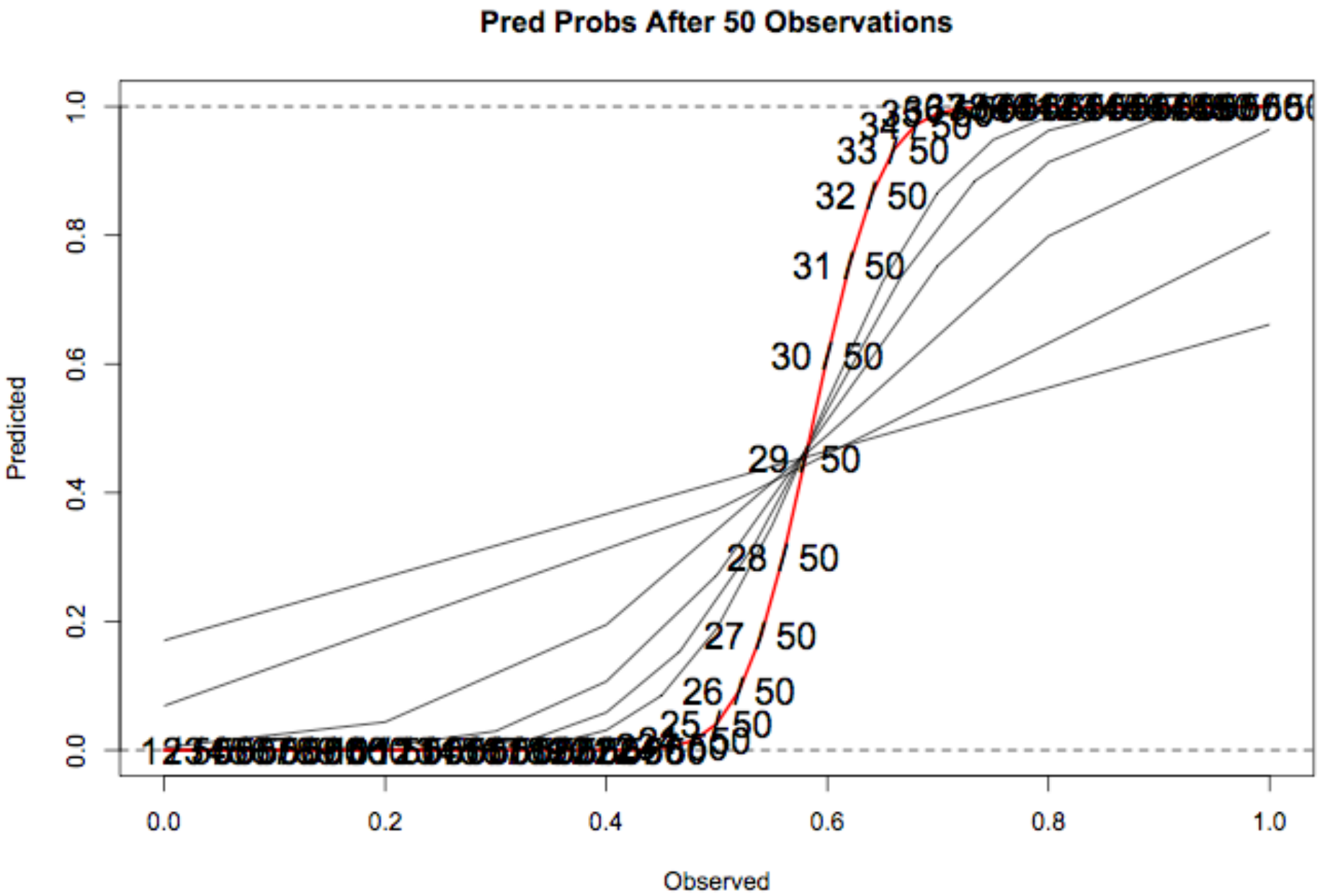

Pred Probs After 1 Observations
Predicted
Observed
0.0
0.2
0.4
0.6
0.8
1.0
0/1
1 / 1

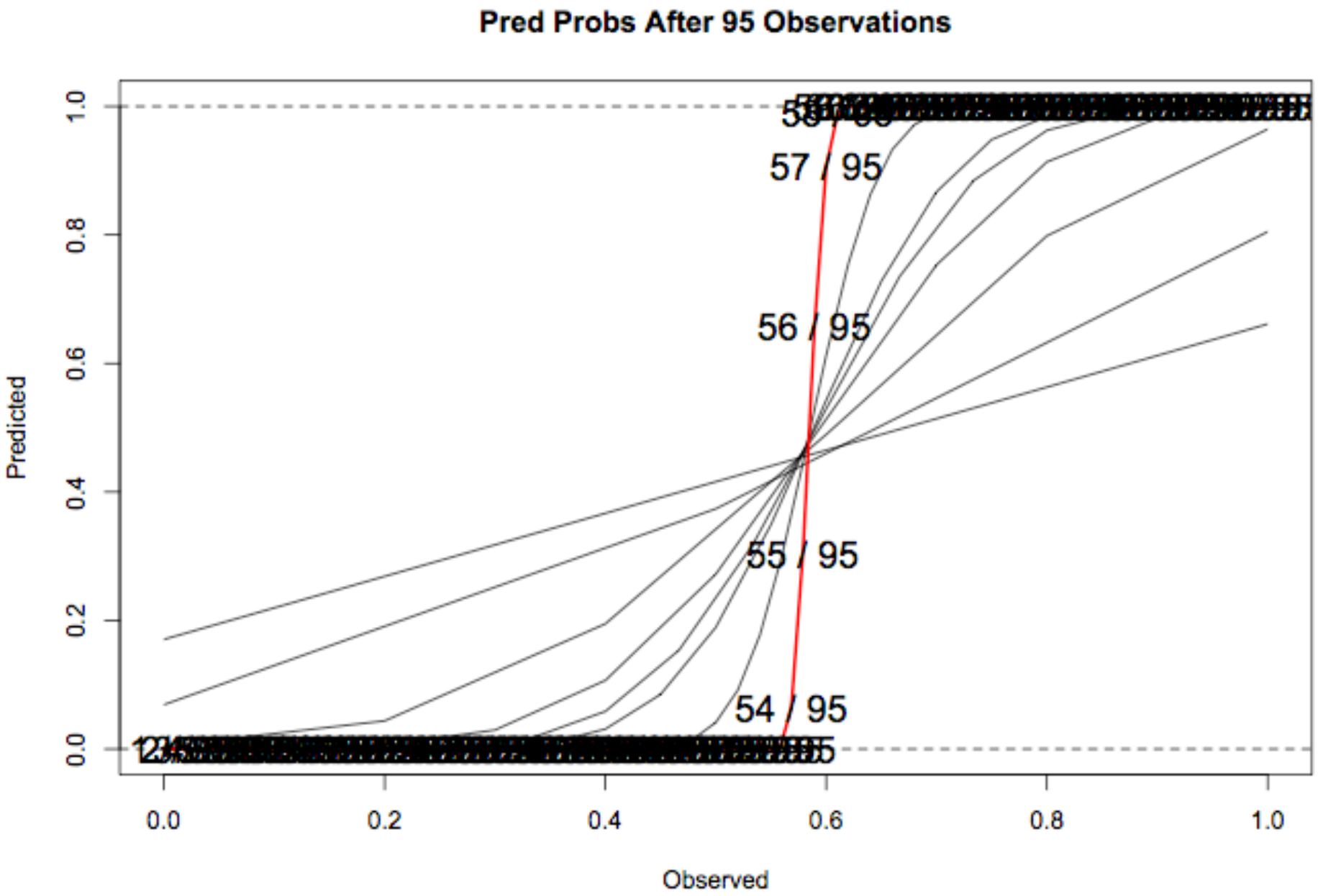

Pred Probs After 2 Observations
Predicted
Observed
0.0
0.2
0.4
0.6
0.8
1.0
0/2
1 / 2
2 / 2

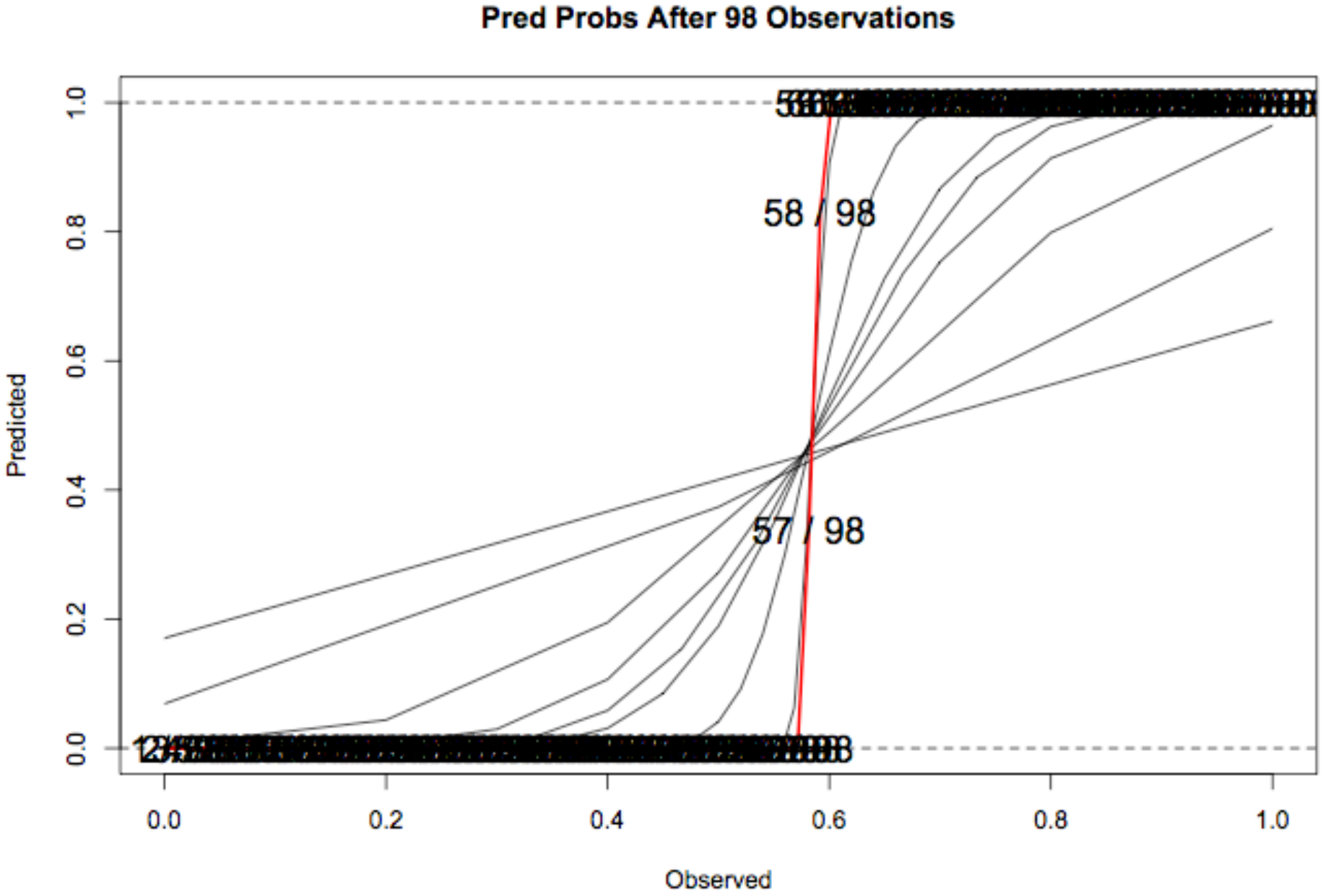

Pred Probs After 5 Observations
Predicted
Observed
0.0
0.2
0.4
0.6
0.8
1.0
1 / 5
2 / 5
3 / 5
4 / 5
5 / 5

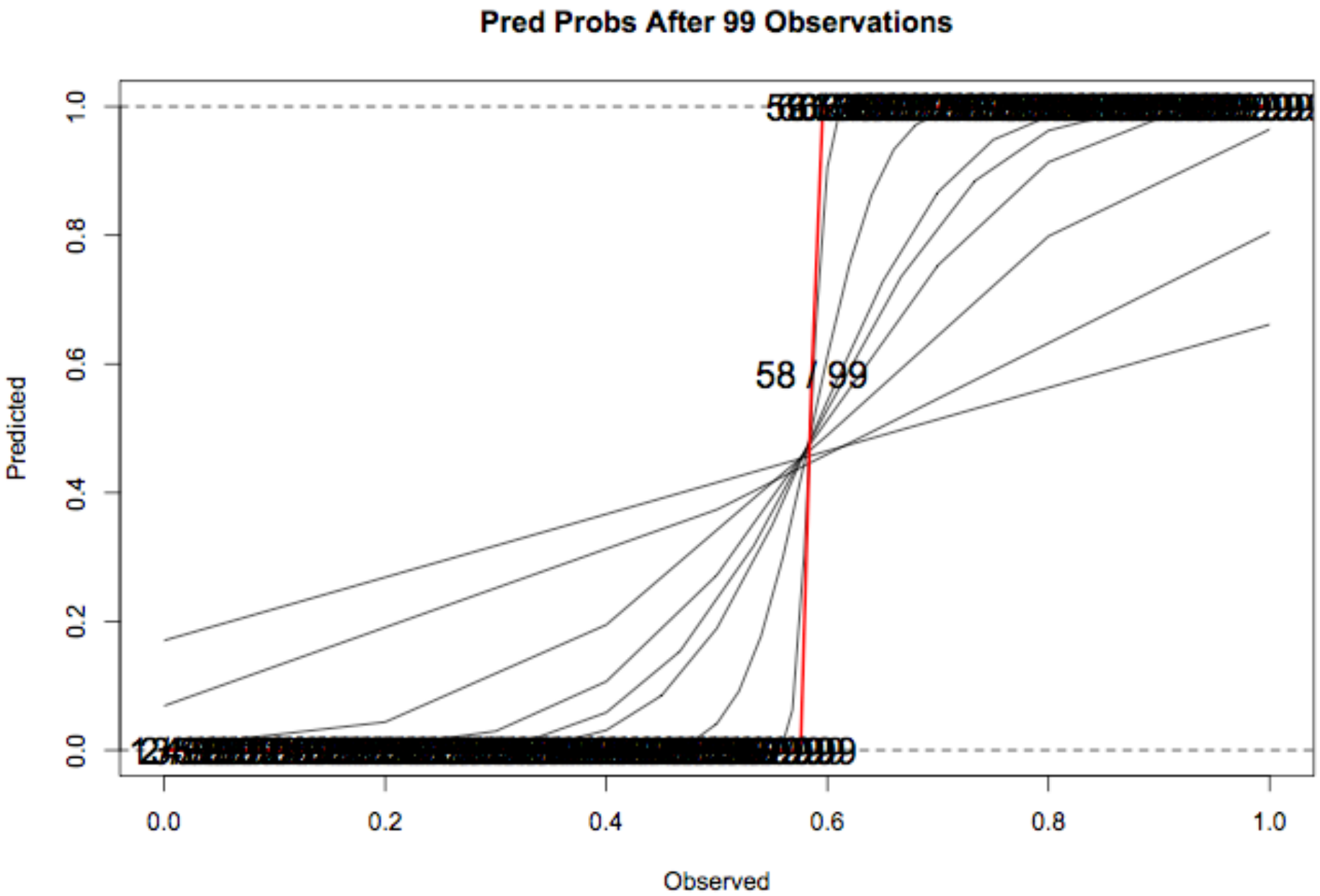

Pred Probs After 10 Observations
Predicted
Observed
0.0
0.2
0.4
0.6
0.8
1.0
1 / 10
2 / 10
3 / 10
4 / 10
5 / 10
6 / 10
7 / 10
8 / 10
9 / 10
10 / 10


Pred Probs After 20 Observations
Predicted
Observed
0.0
0.2
0.4
0.6
0.8
1.0
16 / 20
15 / 20
14 / 20
13 / 20
12 / 20
11 / 20
10 / 20
9 / 20
8 / 20


Pred Probs After 50 Observations
Predicted
Observed
0.0
0.2
0.4
0.6
0.8
1.0
34 / 50
33 / 50
32 / 50
31 / 50
30 / 50
29 / 50
28 / 50
27 / 50
26 / 50
25 / 50


Pred Probs After 95 Observations
57 / 95
56 / 95
55 / 95
54 / 95
Predicted
Observed
0.0
0.2
0.4
0.6
0.8
1.0


Pred Probs After 98 Observations
58 / 98
57 / 98
Predicted
Observed
0.0
0.2
0.4
0.6
0.8
1.0


Pred Probs After 99 Observations
58 / 99
Predicted
Observed
0.0
0.2
0.4
0.6
0.8
1.0

# Predictive Probabilities

- Observe 12 / 20 (60%)
  - Need 47 / 80 successes; 59% or better rest of way
  - *p-value* = 0.25, Pr($p$>0.5) = 0.81
  - Predictive probability of success @ 100 = 0.54

# Predictive Probabilities

- Observe 12 / 20 (60%)
  - Need 47 / 80 successes; 59% or better rest of way
  - *p-value* = 0.25, Pr($p$>0.5) = 0.81
  - Predictive probability of success @ 100 = 0.54
- Observe 28 / 50 (56%)
  - Need 31/50 successes; 62% or better rest of way
  - *p-value* = 0.24, Pr($p$>0.5) = 0.80
  - Predictive probability of success @ 100 = 0.30

# Predictive Probabilities

- Observe 12 / 20 (60%)
  - Need 47 / 80 successes; 59% or better rest of way
  - *p-value* = 0.25, Pr($p$>0.5) = 0.81
  - Predictive probability of success @ 100 = 0.54
- Observe 28 / 50 (56%)
  - Need 31/50 successes; 62% or better rest of way
  - *p-value* = 0.24, Pr($p$>0.5) = 0.80
  - Predictive probability of success @ 100 = 0.30
- Observe 41 / 75 (54.7%)
  - Need 18/25 successes; 72% or better rest of way
  - *p-value* = 0.24, Pr($p$>0.5) = 0.79
  - Predictive probability of success @ 100 = 0.086

# Predictive Probabilities

- Observe 12 / 20 (60%)
  - Need 47 / 80 successes; 59% or better rest of way
  - *p-value* = 0.25, Pr($p$>0.5) = 0.81
  - Predictive probability of success @ 100 = 0.54
- Observe 28 / 50 (56%)
  - Need 31/50 successes; 62% or better rest of way
  - *p-value* = 0.24, Pr($p$>0.5) = 0.80
  - Predictive probability of success @ 100 = 0.30
- Observe 41 / 75 (54.7%)
  - Need 18/25 successes; 72% or better rest of way
  - *p-value* = 0.24, Pr($p$>0.5) = 0.79
  - Predictive probability of success @ 100 = 0.086

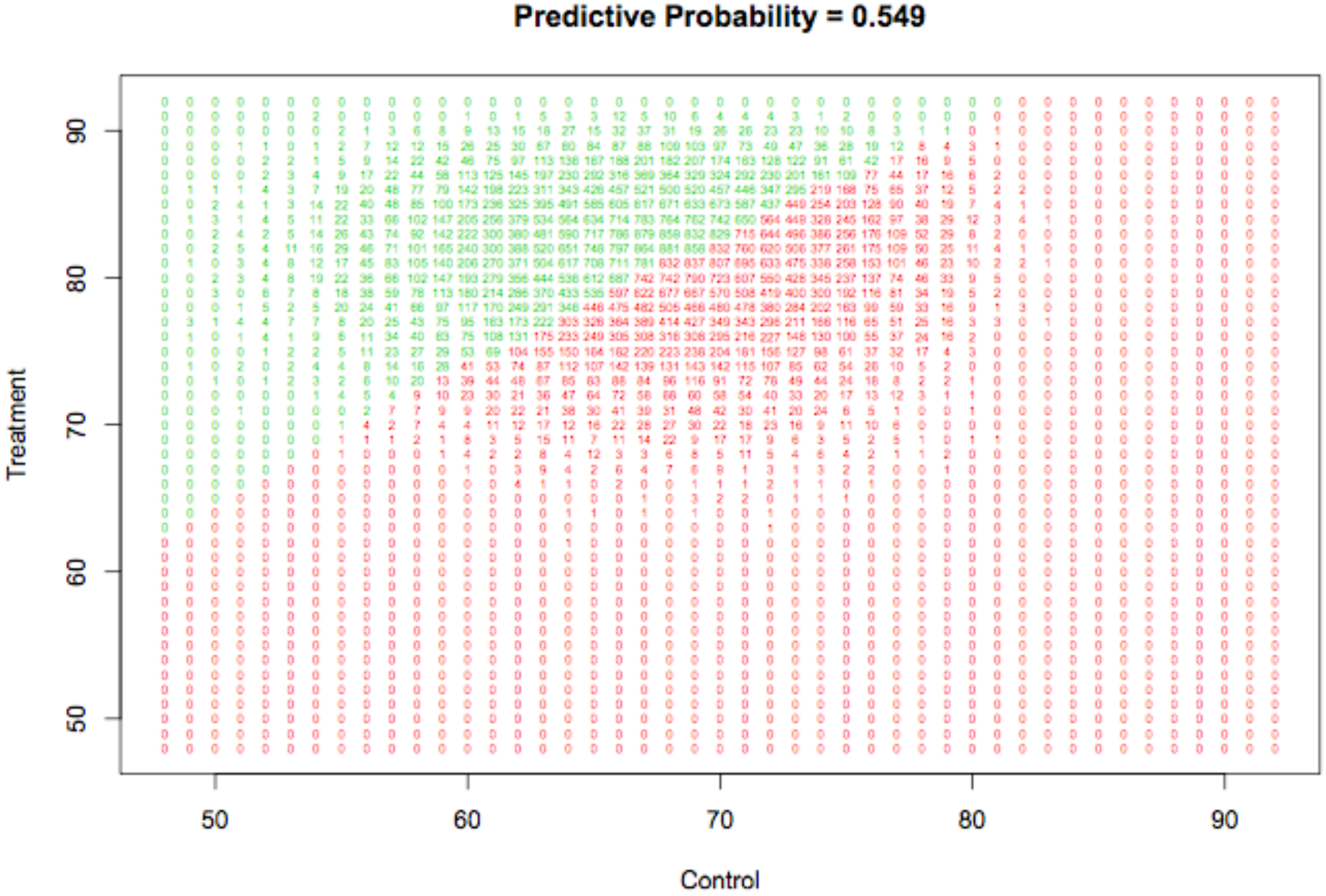
# Another trial

- $N_G = 100$ in Treatment & Control Group
- Testing $p_t > p_c$ using Fisher's Exact Test @ 0.025
- Observe
  - 34/50 in Control Group
  - 41/50 in Treatment Group
- What is predictive probability of success?

# Predictive Probability

```
alpha <- 1; beta <- 1
xc <- 34; nc <- 50
xt <- 41; nt <- 50

pc <- rbeta(100000, alpha+xc, beta+nc-xc)
pt <- rbeta(100000, alpha+xt, beta+nt-xt)

xc.total <- xc + rbinom(100000, 50, pc)
xt.total <- xt + rbinom(100000, 50, pt)

p.values <- rep(NA,100000)
for(i in 1:100000){
  p.values[i] <- fisher.test(
      matrix(c(xc.total[i], 100-xc.total[i],
               xt.total[i], 100-xt.total[i]),nrow=2),
               alternative="less")$p.value
}

> mean(p.value<0.025)
[1] 0.549
```

GREEN numbers are when it's statistically superior
RED are cases not significant



# But what if we have historical data

- Well known historical data, $p_c = 60\% \pm 5\%$

- Expected from pilot studies, $p_t = 80\% \pm 15\%$

# But what if we have historical data

- Well known historical data, $p_c = 60\% \pm 5\%$
- Expected from pilot studies, $p_t = 80\% \pm 15\%$
- Beta distribution defined by p~Beta(α,β) has mean & variance

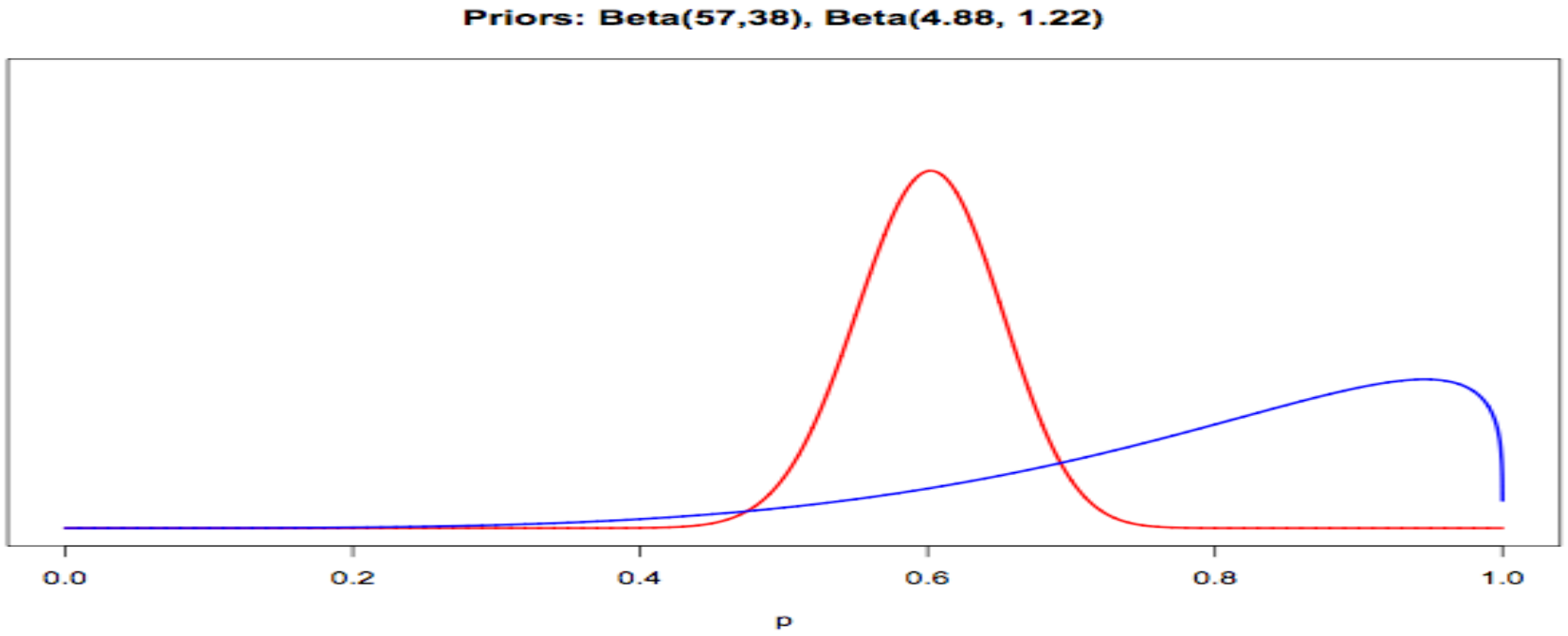
$$E(p) = \frac{\alpha}{\alpha + \beta} \qquad V(p) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

- Solve for α & β

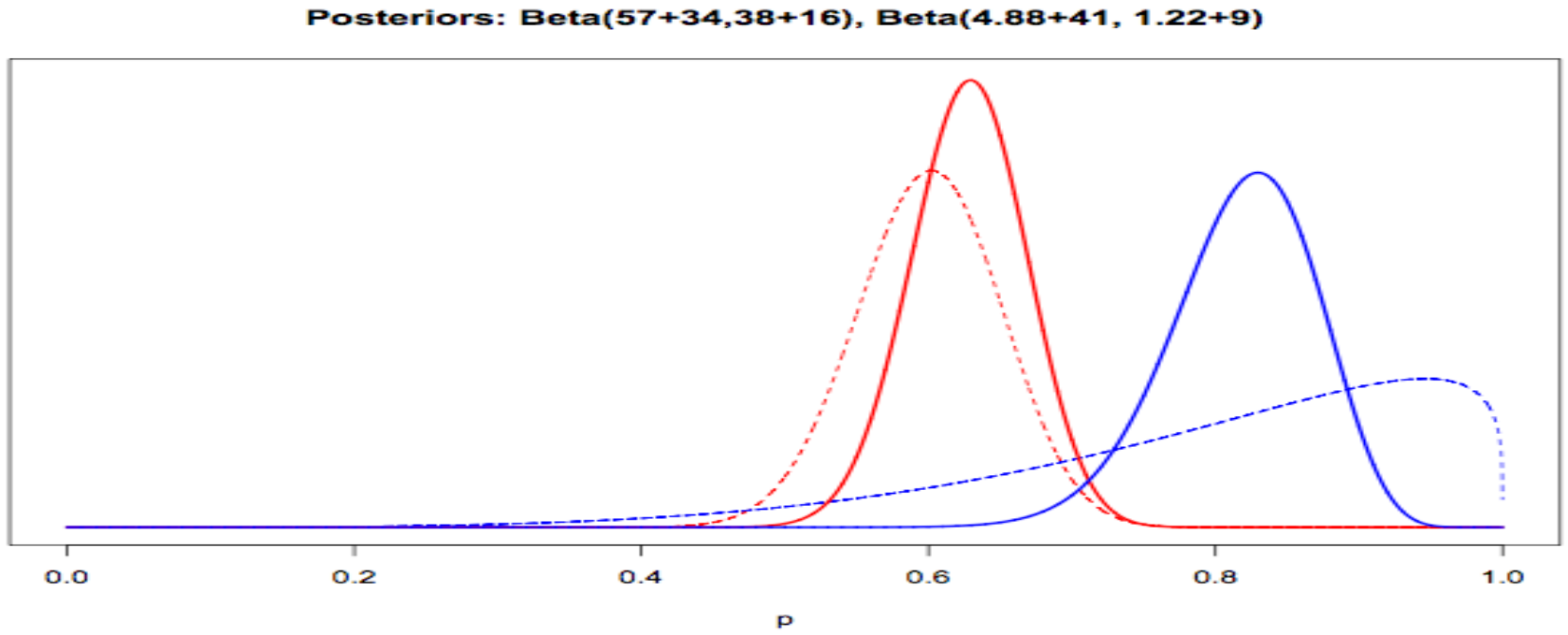
$$\frac{\alpha}{\alpha + \beta} = 0.6 \qquad \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} = 0.05^2$$

# But what if we have historical data

- Well known historical data, $p_c = 60\% \pm 5\%$
  - $\alpha_c = 57$, $\beta_c = 38$
- Expected from pilot studies, $p_t = 80\% \pm 15\%$
  - $\alpha_t = 4.8888$, $\beta_t = 1.2222$

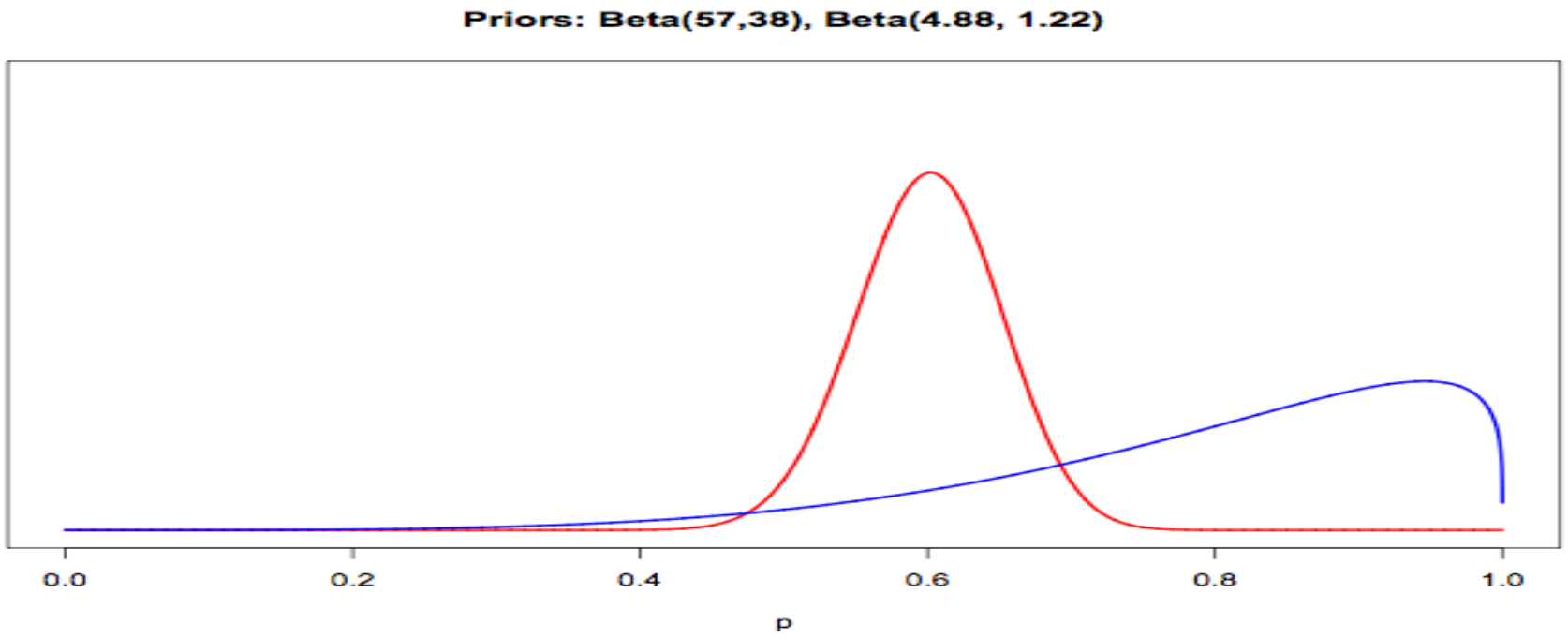


```
alpha.c <- 57; beta.c <- 38; alpha.t <- 4.888888; beta.t <- 1.222222
xc <- 34; nc <- 50; xt <- 41; nt <- 50
pc <- rbeta(100000, alpha.c+xc, beta.c+nc-xc)
pt <- rbeta(100000, alpha.t+xt, beta.t+nt-xt)
xc.total <- xc + rbinom(100000, 50, pc)
xt.total <- xt + rbinom(100000, 50, pt)
p.values <- rep(NA,100000)
for(i in 1:100000){
  p.values[i] <- fisher.test(matrix(c(xc.total[i], 100-xc.total[i],
                                      xt.total[i], 100-xt.total[i]),nrow=2),
                                      alternative="less")$p.value
                    }
> mean(p.values<0.025)
[1] 0.73422
```



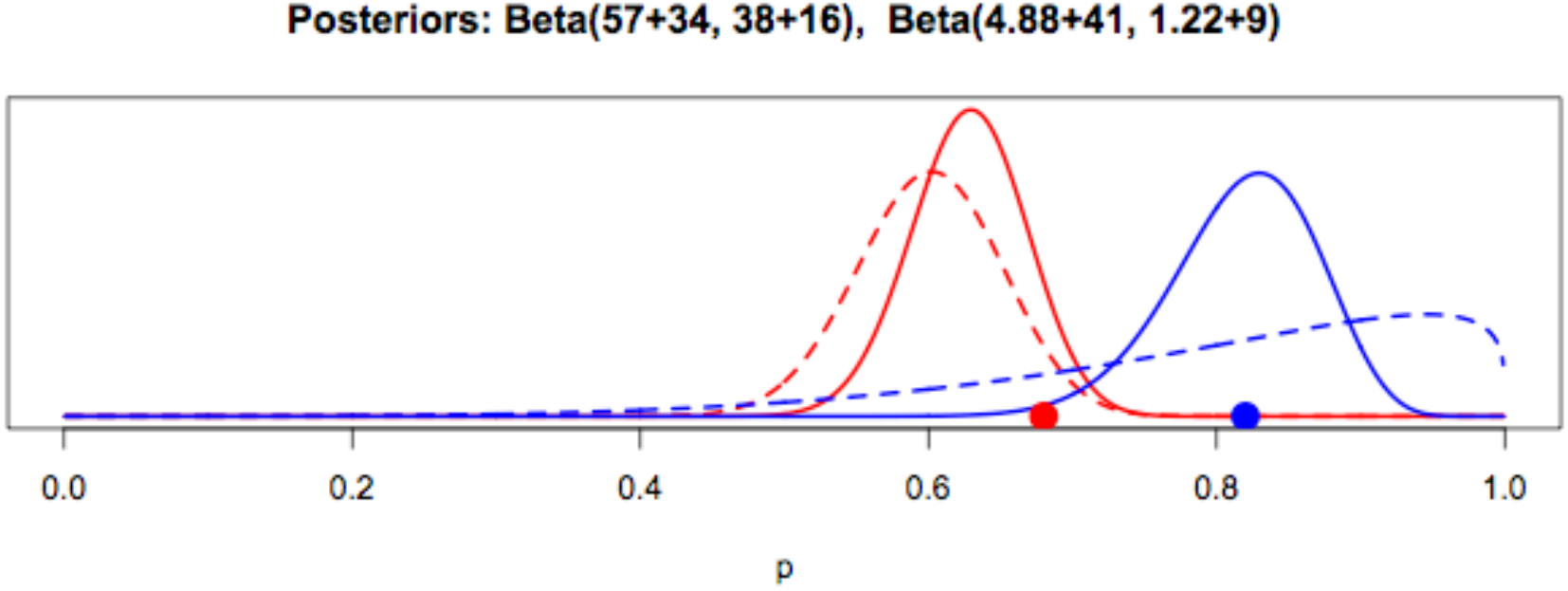
# Downweight Historical Information

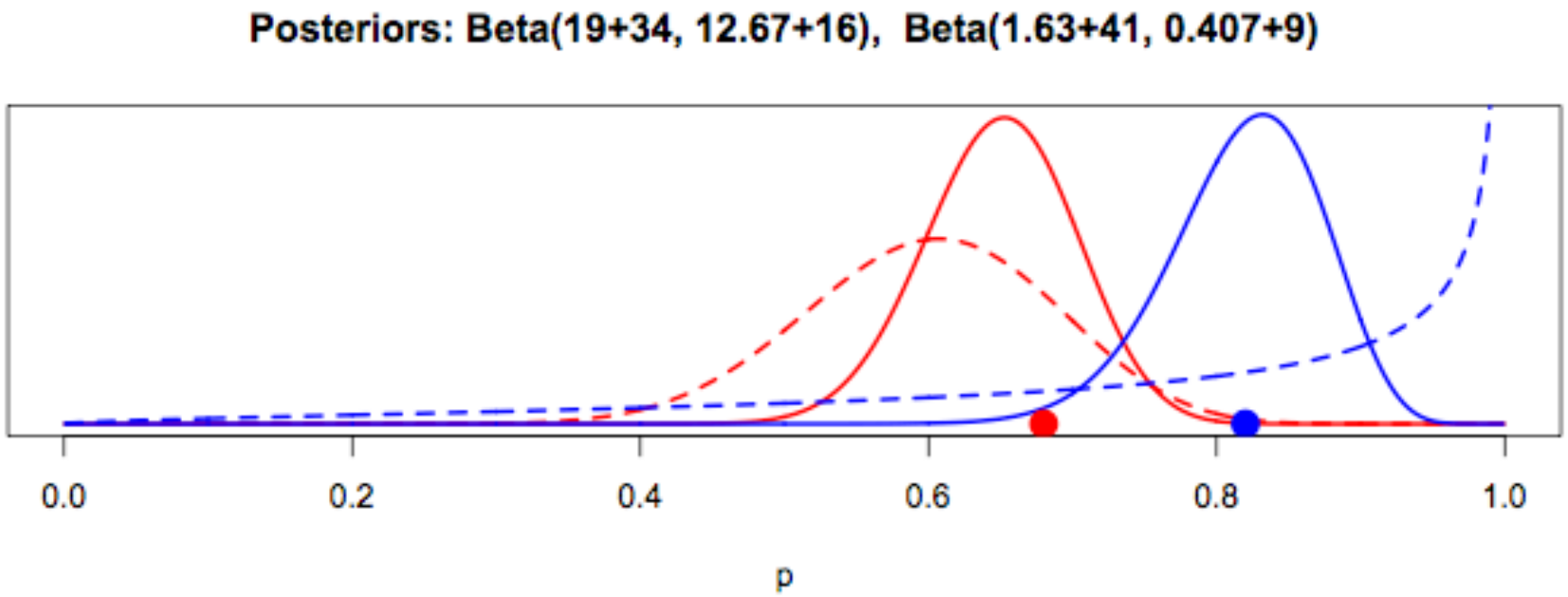
- Well known historical data, $p_c = 60\% \pm 5\%$
  - $\alpha_c = 57$, $\beta_c = 38$, 95 patients' worth of info
- Expected from pilot studies, $p_t = 80\% \pm 15\%$
  - $\alpha_t = 4.8888$, $\beta_t = 1.2222$ 6.1 pts' worth of info



# Downweight Historical Information

- Well known historical data, $p_c = 60\% \pm 5\%$
  - $\alpha_c = 57$, $\beta_c = 38$, 95 patients' worth of info
  - New data is 50/(50+95) = 34% of information
- Expected from pilot studies, $p_t = 80\% \pm 15\%$
  - $\alpha_t = 4.8888$, $\beta_t = 1.2222$, 6.1 patients' worth of info
  - New data is 50/(50+6) = 89% of information
- Downweight each prior so it includes 1/3 as much information
  - $\alpha_c = 19$, $\beta_c = 12.6667$, 31.67 patients' worth of info
  - $\alpha_t = 1.63$, $\beta_t = 0.407$, 2 patients' worth of info


Posteriors: Beta(57+34, 38+16), Beta(4.88+41, 1.22+9)
0.0
0.2
0.4
0.6
0.8
1.0
p


Posteriors: Beta(19+34, 12.67+16), Beta(1.63+41, 0.407+9)
0.0
0.2
0.4
0.6
0.8
1.0
p

# Phase 3 Cancer Design

- Binary endpoint, complete response observed at 45 days post treatment
  - Consider CR vs. PFS vs. OS?
  - Primary analysis chi-square test
- Expect 20% improvement vs. control
- Use Bayesian prediction to determine sample size necessary for success in frequentist trial
  - Bayesian 'behind the curtain'

# Statistical Model

- Final analysis: Chi-square test
- Interim analyses with
  - $N = N_c + N_t$ patients enrolled; $n = n_c + n_t$ complete
  - $x_c \sim \text{Binomial}(n_c, p_c)$; $p_c \sim \text{Beta}(1,1)$
  - $x_t \sim \text{Binomial}(n_t, p_t)$; $p_t \sim \text{Beta}(1,1)$
  - $N = N_c + N_t$ $N_c = n_c + n_c^*$ $N_t = n_t + n_t^*$
  - $x_c^* \sim \text{Beta-binomial}(n_c^*, 1+x_c, 1+n_c-x_c)$
  - $x_t^* \sim \text{Beta-binomial}(n_t^*, 1+x_t, 1+n_t-x_t)$

$$PP_N = \sum_{x_c^*=0}^{n_c^*} \sum_{x_t^*=0}^{n_t^*} pr\left(x_c^*\right) pr\left(x_t^*\right) I\left\{\chi^2_{p-value}\left(x_c + x_c^*, N_c, x_t + x_t^*, N_t\right) < 0.05\right\}$$

# Design Questions

- What should sample size range be?
  - Most sponsor can do is 300 patients
    - Step 1, calculate power of fixed 300 patient trial

```
> bpower(n1=150, n2=150, p1=0.6, p2=0.8)
  Power
  0.969
```

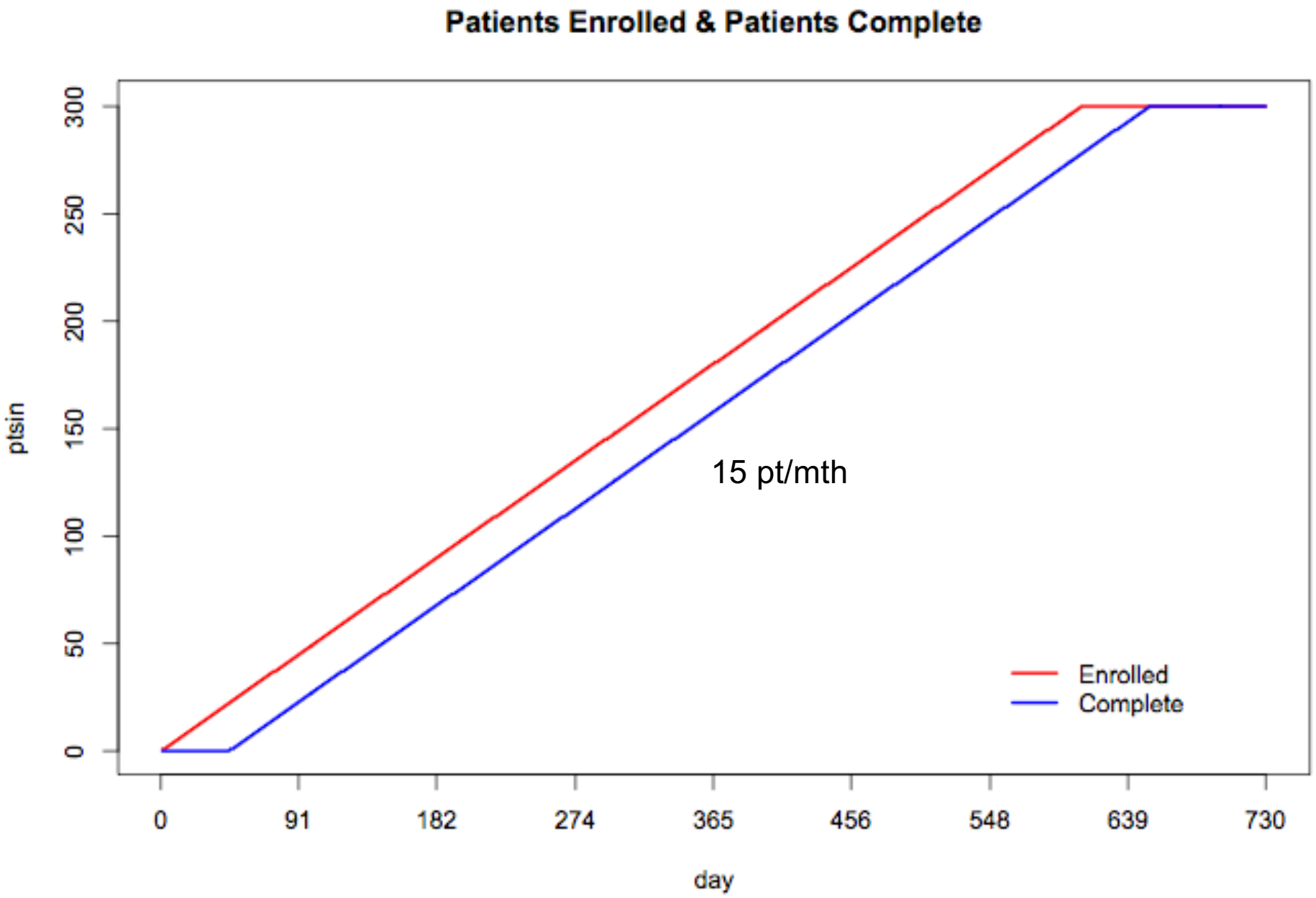
  - Best case want to go to FDA with ≥150 patients
  - We'll see if 300 is enough, if not we'll go back to the company with evidence they need to up the cap

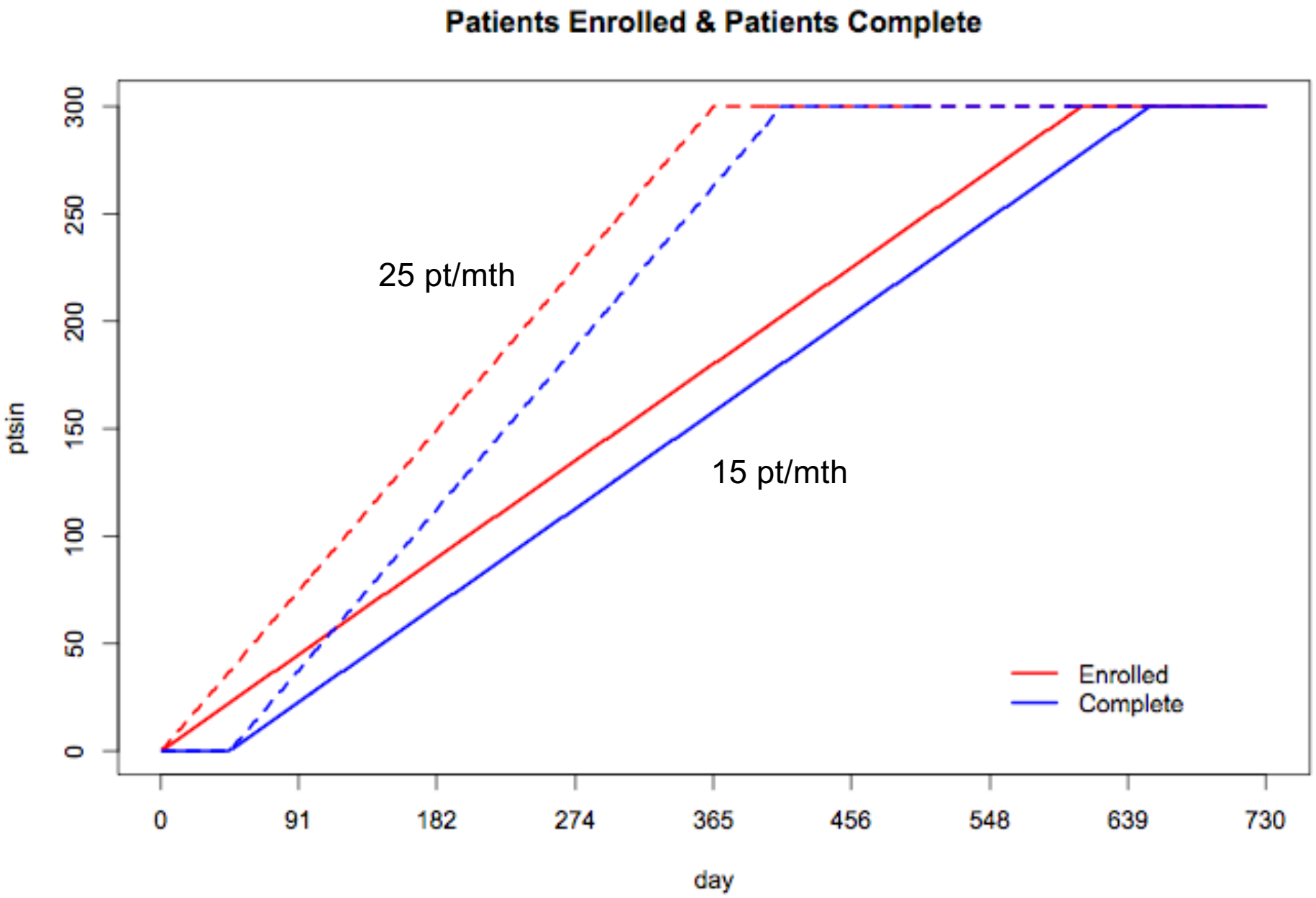
```
> bpower(n1=150, n2=150, p1=0.6, p2=0.75)
  Power
  0.795
```

Smallest win: 60% (80/150) vs. 72% (108/150) → p=0.03

# Design Questions

- Can we use an adaptive design?
  - Expect 15-25 patients per month
  - "Fast" outcome at 45 days
  - 22-37 outstanding patients at any analysis
  - If we do first look @ 150 patients enrolled
    - 128 with complete data with 15 pt/month accrual
    - 113 with complete data with 25 pt/month accrual
  - Usually accrual ramps up, assume constant here
  - Don't want to interfere with accrual
    - Don't pause accrual at each interim analysis
    - Decide whether to stop accrual while accruing


Patients Enrolled & Patients Complete
ptsin
0
50
100
150
200
250
300
0
91
182
274
365
456
548
639
730
day
15 pt/mth
Enrolled
Complete


Patients Enrolled & Patients Complete
ptsin
0
50
100
150
200
250
300
25 pt/mth
15 pt/mth
Enrolled
Complete
0
91
182
274
365
456
548
639
730
day

# Design Questions

- How often to do interim looks?
  - Every 25 patients is every 1-1⅔ months
  - Manageable, may be CRO fee for every look

# Design Questions

- How to decide when to stop accrual for predicted success?

# Design Questions

- How to decide when to stop accrual for predicted success?
  - Use predictive probabilities
  - At each interim analysis ask

    "If we stop enrolling & wait for all outstanding patients to reach their 45-day outcomes, what is the probability we 'win'?"
  - If high, stop, wait, & analyze

    How high?

    I never want to stop then lose! (and so far haven't)

# Design Questions

- How to decide when to stop accrual for futility (if at all)?

# Design Questions

- How to decide when to stop accrual for futility (if at all)?
  - Use predictive probabilities
  - At each analysis ask

    "If we enrolling to the 300-patient maximum then wait for all patients to reach their 45-day outcomes, what is the probability we 'win'?"
  - If low, stop for futility?

    How low?

    More aggressive, more likely to stop a good trial

# Design Questions

- What priors to use for predictive probabilities Beta dists?
  - Pretty new, let's be conservative with Beta(1,1) for treatment & control
  - Could use historical (or downweighted historical) priors here Incentive to have an 'honest' prior
  - Don't use prior in final analysis, frequentist test
- Stop for predicted success if $PP_N > S_N = 0.90$
- Stop for futility if $PP_{Nmax} < F_N = 0.10$

# Sketch of my simulation code

- Define when to analyze, priors, cap, accrual rate, alpha level, efficacy
  - Factors I'll change a lot during discussions with sponsor
- Subroutine for patient accrual & randomization
- Subroutine to generate patient response & dropout
- Subroutine for interim analysis
  - Factors in time of analysis, which patients enrolled, which pts have outcomes
  - Outputs predictive probability of success with current $n$ and at maximum $N$
- Subroutine for decision
  - Stop for predicted success, stop for cap, stop for futility, keep going
- Final analysis at $n$ where trial stopped
- Track trial size, win or lose, reason for stopping, number of looks, trial duration

Control Rate= 0.6000
Exper Rate = 0.8000

Accrual Rate (pts/month): 15.00
Number of Sims 1000
Minimum Sample Size 150
Maximum Sample Size 300
CV 0.0250
Cuts 0.9000 0.1000

| | Mean | SD |
|---|---|---|
| Sample Size | 179.60 | 45.10 |

| | Lose | Win |
|---|---|---|
| Success | 0.008 | 0.897 |
| Cap | 0.012 | 0.048 |
| Futility | 0.035 | 0.000 |
| Total | 0.055 | 0.945 |

| Look | Lose | Win | Total |
|---|---|---|---|
| 150 | 0.020 | 0.565 | 0.585 |
| 175 | 0.005 | 0.118 | 0.123 |
| 200 | 0.002 | 0.091 | 0.093 |
| 225 | 0.004 | 0.069 | 0.073 |
| 250 | 0.006 | 0.028 | 0.034 |
| 275 | 0.006 | 0.026 | 0.032 |
| 300 | 0.012 | 0.048 | 0.060 |
| Tot | 0.055 | 0.945 | 1.000 |

```
Control Rate=     0.6000
 Exper Rate =     0.8000

Accrual Rate (pts/month):   15.00
        Number of Sims        1000
    Minimum Sample Size        150
    Maximum Sample Size        300
                     CV     0.0250
          Cuts    0.9000    0.1000

                    Mean        SD
    Sample Size   179.60     45.10

                  Lose       Win
   Success       0.008     0.897
       Cap       0.012     0.048
  Futility       0.035     0.000
     Total       0.055     0.945

Look      Lose       Win     Total
 150     0.020     0.565     0.585
 175     0.005     0.118     0.123
 200     0.002     0.091     0.093
 225     0.004     0.069     0.073
 250     0.006     0.028     0.034
 275     0.006     0.026     0.032
 300     0.012     0.048     0.060
 Tot     0.055     0.945     1.000
```

Fixed trial of 300 provided 96.9% power

This design provides 94.5% power with average sample size just 180 patients

Control Rate= 0.6000
Exper Rate = 0.8000

Accrual Rate (pts/month): 15.00
Number of Sims 1000
Minimum Sample Size 150
Maximum Sample Size 300
CV 0.0250
Cuts 0.9000 0.1000

| | Mean | SD |
|---|---|---|
| Sample Size | 179.60 | 45.10 |

| | Lose | Win |
|---|---|---|
| Success | 0.008 | 0.897 |
| Cap | 0.012 | 0.048 |
| Futility | 0.035 | 0.000 |
| Total | 0.055 | 0.945 |

| Look | Lose | Win | Total |
|---|---|---|---|
| 150 | 0.020 | 0.565 | 0.585 |
| 175 | 0.005 | 0.118 | 0.123 |
| 200 | 0.002 | 0.091 | 0.093 |
| 225 | 0.004 | 0.069 | 0.073 |
| 250 | 0.006 | 0.028 | 0.034 |
| 275 | 0.006 | 0.026 | 0.032 |
| 300 | 0.012 | 0.048 | 0.060 |
| Tot | 0.055 | 0.945 | 1.000 |

Control Rate= 0.6000
Exper Rate = 0.8000

Accrual Rate (pts/month): 15.00
Number of Sims 1000
Minimum Sample Size 150
Maximum Sample Size 300
CV 0.0250
Cuts 0.9000 0.0000

| | Mean | SD |
|---|---|---|
| Sample Size | 182.65 | 49.86 |

| | Lose | Win |
|---|---|---|
| Success | 0.013 | 0.894 |
| Cap | 0.026 | 0.067 |
| Futility | 0.000 | 0.000 |
| Total | 0.039 | 0.961 |

| Look | Lose | Win | Total |
|---|---|---|---|
| 150 | 0.011 | 0.586 | 0.597 |
| 175 | 0.000 | 0.097 | 0.097 |
| 200 | 0.001 | 0.082 | 0.083 |
| 225 | 0.000 | 0.071 | 0.071 |
| 250 | 0.001 | 0.022 | 0.023 |
| 275 | 0.000 | 0.036 | 0.036 |
| 300 | 0.026 | 0.067 | 0.093 |
| Tot | 0.039 | 0.961 | 1.000 |

# Stopping Boundaries, $S_n$, $F_n$

- Need not be constant
- We stopped for predicted success but lost at the first interim analysis in 1.1% of trials
  - I never want this to happen if I can avoid it!
- Let $S_n$ be the success stopping bound
- Let $F_n$ be the futility stopping bound
- Current $S_n = 0.9$ & $F_n = 0.1$ for all $n$
- Could choose $S_n = 0.99$ for small $n$

  & $S_n = 0.9$ for higher $n$

Control Rate= 0.6000
Exper Rate = 0.8000

Accrual Rate (pts/month): 15.00
Number of Sims 1000
Minimum Sample Size 150
Maximum Sample Size 300
CV 0.0250
Cuts 0.9000 0.0000

| | Mean | SD |
|---|---|---|
| Sample Size | 182.65 | 49.86 |

| | Lose | Win |
|---|---|---|
| Success | 0.013 | 0.894 |
| Cap | 0.026 | 0.067 |
| Futility | 0.000 | 0.000 |
| Total | 0.039 | 0.961 |

| Look | Lose | Win | Total |
|---|---|---|---|
| 150 | 0.011 | 0.586 | 0.597 |
| 175 | 0.000 | 0.097 | 0.097 |
| 200 | 0.001 | 0.082 | 0.083 |
| 225 | 0.000 | 0.071 | 0.071 |
| 250 | 0.001 | 0.022 | 0.023 |
| 275 | 0.000 | 0.036 | 0.036 |
| 300 | 0.026 | 0.067 | 0.093 |
| Tot | 0.039 | 0.961 | 1.000 |

Control Rate= 0.6000
Exper Rate = 0.8000

Accrual Rate (pts/month): 15.00
Number of Sims 1000
Minimum Sample Size 150
Maximum Sample Size 300
CV 0.0250
Cuts 0.9500 0.0000

| | Mean | SD |
|---|---|---|
| Sample Size | 186.47 | 53.61 |

| | Lose | Win |
|---|---|---|
| Success | 0.001 | 0.905 |
| Cap | 0.032 | 0.062 |
| Futility | 0.000 | 0.000 |
| Total | 0.033 | 0.967 |

| Look | Lose | Win | Total |
|---|---|---|---|
| 150 | 0.000 | 0.520 | 0.520 |
| 175 | 0.001 | 0.135 | 0.136 |
| 200 | 0.000 | 0.110 | 0.110 |
| 225 | 0.000 | 0.054 | 0.054 |
| 250 | 0.000 | 0.053 | 0.053 |
| 275 | 0.000 | 0.033 | 0.033 |
| 300 | 0.032 | 0.062 | 0.094 |
| Tot | 0.033 | 0.967 | 1.000 |

Control Rate= 0.6000
Exper Rate = 0.8000

Accrual Rate (pts/month): 15.00
Number of Sims 1000
Minimum Sample Size 150
Maximum Sample Size 300
CV 0.0250
Cuts 0.9500 0.0000

| | Mean | SD |
|---|---|---|
| Sample Size | 186.47 | 53.61 |

| | Lose | Win |
|---|---|---|
| Success | 0.001 | 0.905 |
| Cap | 0.032 | 0.062 |
| Futility | 0.000 | 0.000 |
| Total | 0.033 | 0.967 |

| Look | Lose | Win | Total |
|---|---|---|---|
| 150 | 0.000 | 0.520 | 0.520 |
| 175 | 0.001 | 0.135 | 0.136 |
| 200 | 0.000 | 0.110 | 0.110 |
| 225 | 0.000 | 0.054 | 0.054 |
| 250 | 0.000 | 0.053 | 0.053 |
| 275 | 0.000 | 0.033 | 0.033 |
| 300 | 0.032 | 0.062 | 0.094 |
| Tot | 0.033 | 0.967 | 1.000 |

Control Rate= 0.6000
Exper Rate = 0.8000

Accrual Rate (pts/month): 15.00
Number of Sims 1000
Minimum Sample Size 150
Maximum Sample Size 300
CV 0.0250
Cuts 0.9500 0.0500

| | Mean | SD |
|---|---|---|
| Sample Size | 183.82 | 46.57 |

| | Lose | Win |
|---|---|---|
| Success | 0.001 | 0.915 |
| Cap | 0.014 | 0.048 |
| Futility | 0.022 | 0.000 |
| Total | 0.037 | 0.963 |

| Look | Lose | Win | Total |
|---|---|---|---|
| 150 | 0.012 | 0.513 | 0.525 |
| 175 | 0.003 | 0.139 | 0.142 |
| 200 | 0.004 | 0.108 | 0.112 |
| 225 | 0.001 | 0.061 | 0.062 |
| 250 | 0.000 | 0.056 | 0.056 |
| 275 | 0.003 | 0.038 | 0.042 |
| 300 | 0.014 | 0.048 | 0.063 |
| Tot | 0.037 | 0.963 | 1.000 |

Control Rate= 0.6000
Exper Rate = 0.8000

Accrual Rate (pts/month): 15.00
Number of Sims 1000
Minimum Sample Size 150
Maximum Sample Size 300
CV 0.0250
Cuts 0.9500 0.0500

| | Mean | SD |
|---|---|---|
| Sample Size | 183.82 | 46.57 |

| | Lose | Win |
|---|---|---|
| Success | 0.001 | 0.915 |
| Cap | 0.014 | 0.048 |
| Futility | 0.022 | 0.000 |
| Total | 0.037 | 0.963 |

| Look | Lose | Win | Total |
|---|---|---|---|
| 150 | 0.012 | 0.513 | 0.525 |
| 175 | 0.003 | 0.139 | 0.142 |
| 200 | 0.004 | 0.108 | 0.112 |
| 225 | 0.001 | 0.061 | 0.062 |
| 250 | 0.000 | 0.056 | 0.056 |
| 275 | 0.003 | 0.038 | 0.042 |
| 300 | 0.014 | 0.048 | 0.063 |
| Tot | 0.037 | 0.963 | 1.000 |

Control Rate= 0.6000
Exper Rate = 0.8000

Accrual Rate (pts/month): 15.00
Number of Sims 1000
Minimum Sample Size 150
Maximum Sample Size 300
CV 0.0250
Cuts 0.9500 0.1000

| | Mean | SD |
|---|---|---|
| Sample Size | 183.20 | 48.53 |

| | Lose | Win |
|---|---|---|
| Success | 0.001 | 0.892 |
| Cap | 0.015 | 0.065 |
| Futility | 0.027 | 0.000 |
| Total | 0.043 | 0.957 |

| Look | Lose | Win | Total |
|---|---|---|---|
| 150 | 0.017 | 0.546 | 0.564 |
| 175 | 0.006 | 0.118 | 0.124 |
| 200 | 0.001 | 0.093 | 0.094 |
| 225 | 0.000 | 0.054 | 0.054 |
| 250 | 0.002 | 0.049 | 0.051 |
| 275 | 0.002 | 0.032 | 0.034 |
| 300 | 0.015 | 0.065 | 0.080 |
| Tot | 0.043 | 0.957 | 1.000 |

Control Rate= 0.6000
Exper Rate = 0.7500

Accrual Rate (pts/month): 15.00
Number of Sims 5000
Minimum Sample Size 150
Maximum Sample Size 300
CV 0.0250
Cuts 0.9500 0.0500

| | Mean | SD |
|---|---|---|
| Sample Size | 217.45 | 59.78 |

| | Lose | Win |
|---|---|---|
| Success | 0.009 | 0.639 |
| Cap | 0.083 | 0.152 |
| Futility | 0.116 | 0.000 |
| Total | 0.209 | 0.791 |

| Look | Lose | Win | Total |
|---|---|---|---|
| 150 | 0.044 | 0.260 | 0.304 |
| 175 | 0.017 | 0.100 | 0.117 |
| 200 | 0.012 | 0.086 | 0.098 |
| 225 | 0.016 | 0.068 | 0.084 |
| 250 | 0.018 | 0.067 | 0.085 |
| 275 | 0.019 | 0.057 | 0.076 |
| 300 | 0.083 | 0.152 | 0.235 |
| Tot | 0.209 | 0.791 | 1.000 |

Control Rate= 0.6000
Exper Rate = 0.7500

Accrual Rate (pts/month): 15.00
Number of Sims 5000
Minimum Sample Size 150
Maximum Sample Size 300
CV 0.0250
Cuts 0.9500 0.1000

| | Mean | SD |
|---|---|---|
| Sample Size | 211.28 | 57.80 |

| | Lose | Win |
|---|---|---|
| Success | 0.008 | 0.654 |
| Cap | 0.063 | 0.128 |
| Futility | 0.148 | 0.000 |
| Total | 0.219 | 0.781 |

| Look | Lose | Win | Total |
|---|---|---|---|
| 150 | 0.064 | 0.263 | 0.327 |
| 175 | 0.024 | 0.105 | 0.129 |
| 200 | 0.020 | 0.088 | 0.108 |
| 225 | 0.016 | 0.072 | 0.088 |
| 250 | 0.017 | 0.073 | 0.090 |
| 275 | 0.015 | 0.053 | 0.068 |
| 300 | 0.063 | 0.128 | 0.191 |
| Tot | 0.219 | 0.781 | 1.000 |

Control Rate= 0.6000
Exper Rate = 0.6000

Accrual Rate (pts/month): 15.00
Number of Sims 5000
Minimum Sample Size 150
Maximum Sample Size 300
CV 0.0250
Cuts 0.9500 0.0500

| | Mean | SD |
|---|---|---|
| Sample Size | 187.32 | 49.97 |

| | Lose | Win |
|---|---|---|
| Success | 0.002 | 0.020 |
| Cap | 0.066 | 0.012 |
| Futility | 0.900 | 0.000 |
| Total | 0.968 | 0.032 |

| Look | Lose | Win | Total |
|---|---|---|---|
| 150 | 0.519 | 0.008 | 0.527 |
| 175 | 0.117 | 0.002 | 0.119 |
| 200 | 0.079 | 0.002 | 0.081 |
| 225 | 0.079 | 0.003 | 0.082 |
| 250 | 0.062 | 0.002 | 0.064 |
| 275 | 0.046 | 0.002 | 0.048 |
| 300 | 0.066 | 0.012 | 0.078 |
| Tot | 0.968 | 0.032 | 1.000 |

Control Rate= 0.6000
Exper Rate = 0.6000

Accrual Rate (pts/month): 15.00
Number of Sims 1000
Minimum Sample Size 150
Maximum Sample Size 300
CV 0.0250
Cuts 0.9500 0.1000

| | Mean | SD |
|---|---|---|
| Sample Size | 176.31 | 44.02 |

| | Lose | Win |
|---|---|---|
| Success | 0.002 | 0.019 |
| Cap | 0.041 | 0.009 |
| Futility | 0.929 | 0.000 |
| Total | 0.972 | 0.028 |

| Look | Lose | Win | Total |
|---|---|---|---|
| 150 | 0.634 | 0.006 | 0.640 |
| 175 | 0.103 | 0.004 | 0.107 |
| 200 | 0.073 | 0.003 | 0.076 |
| 225 | 0.047 | 0.003 | 0.050 |
| 250 | 0.042 | 0.002 | 0.044 |
| 275 | 0.033 | 0.001 | 0.034 |
| 300 | 0.041 | 0.009 | 0.050 |
| Tot | 0.972 | 0.028 | 1.000 |

# Enough!

- Settle on
  - Success Bound = 0.95
  - Futility Bound = 0.10
- Type I error was 0.028 -- too high
  - Pivotal trial, we need this to be ≤ 0.025
  - Hard to calculate analytically
  - Need to simulate over many scenarios
  - Then convince ourselves & FDA we've explored the whole null space

# Intuition Check

- Use critical value = 0.025
- Simulate with 4 accrual rates, 10k sims/scenario
- Will the Type I error rates change with accrual rate? If so how?
- How will sample sizes change?

| Accrual (pts/mth) | Mean N | Type I error |
|---|---|---|
| 5 | | |
| 15* | 177 | 0.030 |
| 25 | | |
| 50 | | |

*Slightly different than previous slide because 10,000 sims each

# Intuition Check

- Use critical value = 0.025
- Simulate with 4 accrual rates, 10k sims/scenario
- Will the Type I error rates change with accrual rate? If so how?
- How will sample sizes change?

| Accrual (pts/mth) | Mean N | Type I error |
|---|---|---|
| 5 | 172 | 0.039 |
| 15 | 177 | 0.030 |
| 25 | 182 | 0.028 |
| 50 | 195 | 0.027 |

# Find Critical Value for $\alpha = 0.025$

- Assume accrual won't be slower than 15/month
- Explore range of true $p_c$ & $p_t$
- Find right critical value by trial & error
  - 10,000 sims each using 0.6 vs. 0.6
  - Sqrt(0.025*0.975/10000) = 0.0016

| Critv | 0.40 | 0.50 | 0.60 | 0.70 | 0.80 |
|---|---|---|---|---|---|
| 0.025 | | | 0.030 | | |

# Find Critical Value for $\alpha = 0.025$

- Assume accrual won't be slower than 15/month
- Explore range of true $p_c$ & $p_t$
- Find right critical value by trial & error
  - 10,000 sims each using 0.4 vs. 0.4 to 0.8 vs. 0.8
  - Sqrt(0.025*0.975/10000) = 0.0016

| Critv | 0.40 | 0.50 | 0.60 | 0.70 | 0.80 |
|---|---|---|---|---|---|
| 0.025 | | | 0.030 | | |
| 0.020 | 0.024 | 0.026 | 0.026 | 0.024 | 0.025 |

# Find Critical Value for $\alpha = 0.025$

- Assume accrual won't be slower than 15/month
- Explore range of true $p_c$ & $p_t$
- Find right critical value by trial & error
  - 10,000 sims each using 0.4 vs. 0.4 to 0.8 vs. 0.8
  - Sqrt(0.025*0.975/10000) = 0.0016

| Critv | 0.40 | 0.50 | 0.60 | 0.70 | 0.80 |
|---|---|---|---|---|---|
| 0.025 | | | 0.030 | | |
| 0.020 | 0.024 | 0.026 | 0.026 | 0.024 | 0.025 |
| 0.018 | 0.024 | 0.021 | 0.023 | 0.023 | 0.020 |

# Find Critical Value for $\alpha = 0.025$

- Assume accrual won't be slower than 15/month
- Explore range of true $p_c$ & $p_t$
- Find right critical value by trial & error
  - 10,000 sims each using 0.4 vs. 0.4 to 0.8 vs. 0.8
  - Sqrt(0.025*0.975/10000) = 0.0016

| Critv | 0.40 | 0.50 | 0.60 | 0.70 | 0.80 |
|---|---|---|---|---|---|
| 0.025 | | | 0.030 | | |
| 0.020 | 0.024 | 0.026 | 0.026 | 0.024 | 0.025 |
| 0.018 | 0.024 | 0.021 | 0.024 | 0.023 | 0.020 |
| 0.019 | 0.022 | 0.026 | 0.024 | 0.024 | 0.024 |

Let's go with 0.018

If a real trial I'd run 100,000 or 1M sims and try to get as much power as possible

# Example Trial #1

```
Simulation #  14        Analysis #  150
 Group           N    Obs    Suc
 Control        75     68     35      51%
 Treatment      75     68     49      72%
 P_N    =  0.9360 >  0.950 ? No,    P_Nmax =  0.9180 <  0.100 ? No
 Continue to enroll
```

# Example Trial #1

```
Simulation #  14        Analysis #  150
 Group           N    Obs    Suc
 Control        75     68     35     51%
 Treatment      75     68     49     72%
 P_N    =  0.9360 >  0.950 ? No,   P_Nmax =  0.9180 <  0.100 ? No
 Continue to enroll

Simulation #  14        Analysis #  175
 Group           N    Obs    Suc
 Control        88     73     39     53%
 Treatment      87     72     53     74%
 P_N    =  0.9370 >  0.950 ? No,   P_Nmax =  0.9360 <  0.100 ? No
 Continue to enroll
```

# Example Trial #1

```
Simulation #  14        Analysis #  150
 Group           N    Obs    Suc
 Control        75     68     35     51%
 Treatment      75     68     49     72%
 P_N    =  0.9360 >  0.950 ? No,   P_Nmax =  0.9180 <  0.100 ? No
 Continue to enroll

Simulation #  14        Analysis #  175
 Group           N    Obs    Suc
 Control        88     73     39     53%
 Treatment      87     72     53     74%
 P_N    =  0.9370 >  0.950 ? No,   P_Nmax =  0.9360 <  0.100 ? No
 Continue to enroll

Simulation #  14        Analysis #  200
 Group           N    Obs    Suc
 Control       100     91     48     53%
 Treatment     100     90     68     76%
 P_N    =  >.9999 >  0.950 ? YES,  P_Nmax =  0.9900 <  0.100 ? No
 Stop for predicted success
```

# Example Trial #1

```
Simulation #  14        Analysis #  150
 Group           N    Obs    Suc
 Control        75     68     35     51%
 Treatment      75     68     49     72%
 P_N    =  0.9360 >  0.950 ? No,   P_Nmax =  0.9180 <  0.100 ? No
 Continue to enroll

Simulation #  14        Analysis #  175
 Group           N    Obs    Suc
 Control        88     73     39     53%
 Treatment      87     72     53     74%
 P_N    =  0.9370 >  0.950 ? No,   P_Nmax =  0.9360 <  0.100 ? No
 Continue to enroll

Simulation #  14        Analysis #  200
 Group           N    Obs    Suc
 Control       100     91     48     53%
 Treatment     100     90     68     76%
 P_N    =  >.9999 >  0.950 ? YES,  P_Nmax =  0.9900 <  0.100 ? No
 Stop for predicted success

Simulation #  14    Final Analysis  200
     Group       N    Obs    Suc
 Control       100    100     52     52%
 Treatment     100    100     76     76%
 Successful trial,   p-value = 0.001 < 0.0180
```

# Example Trial #2

```
Simulation #  10        Analysis #  150
 Group           N    Obs    Suc
 Control        75     66     40      61%  (need to see +20
 Treatment      75     65     44      68%   successes for win @
   150)
 P_n    =  0.0000 >  0.950 ? No,    P_Nmax =  0.2590 <  0.100 ? No
 Continue to enroll

Simulation #  10        Analysis #  175
 Group           N    Obs    Suc
 Control        88     80     47      59%
 Treatment      87     79     51      65%
 P_n    =  0.0000 >  0.950 ? No,    P_Nmax =  0.1020 <  0.100 ? No
 Continue to enroll

Simulation #  10        Analysis #  200
 Group           N    Obs    Suc
 Control       100     90     55      61%  (need to see +18
 Treatment     100     89     57      64%   successes for win @
   300)
 P_n    =  0.0000 >  0.950 ? No,    P_Nmax =  0.0360 <  0.100 ? YES
 Stop for futility
 Unsuccessful trial
```

# Final Operating Characteristics

## $S_n = 0.95, F_n = 0.10$

| $p_c$ | $p_t$ | Mean N | Futility | Max & Win | PredSuc & Win | Power |
|---|---|---|---|---|---|---|
| 0.60 | 0.60 | 175 | 0.937 | 0.046<br>0.009 | 0.016<br>0.015 | 0.024 |
| 0.60 | 0.65 | 199 | 0.775 | 0.145<br>0.041 | 0.081<br>0.075 | 0.117 |
| 0.60 | 0.70 | 220 | 0.478 | 0.247<br>0.114 | 0.275<br>0.267 | 0.381 |
| 0.60 | 0.75 | 216 | 0.195 | 0.216<br>0.143 | 0.590<br>0.580 | 0.723 |
| 0.60 | 0.80 | 189 | 0.039 | 0.088<br>0.073 | 0.873<br>0.868 | 0.942 |

# Final Operating Characteristics

## $S_n = 0.95, F_n = 0.05$

| $p_c$ | $p_t$ | Mean N | Futility | Max & Win | PredSuc & Win | Power |
|---|---|---|---|---|---|---|
| 0.60 | 0.60 | 185 | 0.913 | 0.071<br>0.009 | 0.017<br>0.015 | 0.025 |
| 0.60 | 0.65 | 212 | 0.716 | 0.200<br>0.053 | 0.084<br>0.079 | 0.132 |
| 0.60 | 0.70 | 231 | 0.407 | 0.314<br>0.131 | 0.280<br>0.271 | 0.401 |
| 0.60 | 0.75 | 221 | 0.143 | 0.256<br>0.155 | 0.601<br>0.591 | 0.746 |
| 0.60 | 0.80 | 190 | 0.025 | 0.095<br>0.074 | 0.880<br>0.876 | 0.950 |

# Final Operating Characteristics vs. Fixed Frequentist Trials

| $p_c$ | $p_t$ | B-A Mean N | B-A Power | F-Power 300 | F-Power BA Mean |
|---|---|---|---|---|---|
| 0.60 | 0.60 | 175<br>185 | 0.024<br>0.025 | 0.025 | 0.025 |
| 0.60 | 0.65 | 199<br>212 | 0.12<br>0.13 | 0.14 | 0.11 |
| 0.60 | 0.70 | 220<br>231 | 0.38<br>0.40 | 0.44 | 0.34 |
| 0.60 | 0.75 | 216<br>221 | 0.72<br>0.75 | 0.79 | 0.66 |
| 0.60 | 0.80 | 189<br>190 | 0.94<br>0.95 | 0.969 | 0.86 |

# Summary / Thoughts?

# Imagine

- Imagine investigators do a case-control study
- Identify cases
  - Patients with hypertension
- Identify controls
  - People without hypertension with the same demographics (age, gender, marital status)
- See statistically significant increase in blood pressure between cases & controls
- Would JAMA publish this paper?

- They did: Bassler et al, March 23/31, 2010, V303, No12, 1180-1187.

# Stopping Randomized Trials Early for Benefit and Estimation of Treatment Effects

## Systematic Review and Meta-regression Analysis

Dirk Bassler, MD, MSc
Matthias Briel, MD, MSc
Victor M. Montori, MD, MSc
Melanie Lane, BA
Paul Glasziou, MBBS, PhD
Qi Zhou, PhD
Diane Heels-Ansdell, MSc
Stephen D. Walter, PhD
Gordon H. Guyatt, MD, MSc
and the STOPIT-2 Study Group

**Context** Theory and simulation suggest that randomized controlled trials (RCTs) stopped early for benefit (truncated RCTs) systematically overestimate treatment effects for the outcome that precipitated early stopping.

**Objective** To compare the treatment effect from truncated RCTs with that from meta-analyses of RCTs addressing the same question but not stopped early (nontruncated RCTs) and to explore factors associated with overestimates of effect.

**Data Sources** Search of MEDLINE, EMBASE, Current Contents, and full-text journal content databases to identify truncated RCTs up to January 2007; search of MEDLINE, Cochrane Database of Systematic Reviews, and Database of Abstracts of Reviews of Effects to identify systematic reviews from which individual RCTs were extracted up to January 2008.

**Study Selection** Selected studies were RCTs reported as having stopped early for benefit and matching nontruncated RCTs from systematic reviews. Independent reviewers with medical content expertise, working blinded to trial results, judged the eligibility of the nontruncated RCTs based on their similarity to the truncated RCTs.

**Data Extraction** Reviewers with methodological expertise conducted data extraction independently.

**Results** The analysis included 91 truncated RCTs asking 63 different questions and 424 matching nontruncated RCTs. The pooled ratio of relative risks in truncated RCTs vs matching nontruncated RCTs was 0.71 (95% confidence interval, 0.65-0.77). This difference was independent of the presence of a statistical stopping rule and the methodological quality of the studies as assessed by allocation concealment and blinding. Large differences in treatment effect size between truncated and nontruncated RCTs (ratio of relative risks <0.75) occurred with truncated RCTs having fewer than 500 events. In 39 of the 63 questions (62%), the pooled effects of the nontruncated RCTs failed to demonstrate significant benefit.

**Conclusions** Truncated RCTs were associated with greater effect sizes than RCTs not stopped early. This difference was independent of the presence of statistical stopping rules and was greatest in smaller studies.

*JAMA. 2010;303(12):1180-1187* www.jama.com

ALTHOUGH RANDOMIZED CONtrolled trials (RCTs) generally provide credible evidence of treatment effects, multiple problems may emerge when investigators terminate a trial earlier than planned,[1] especially when the decision to terminate the trial is based on the finding of an apparently beneficial treatment effect. Bias may arise because large random fluctuations of the estimated treatment effect can occur, particularly early in the progress of a trial.[2] When investigators stop a trial based on an apparently beneficial treatment effect, their

# From Abstract Study Selection

"Selected studies were RCTs reported as having stopped early for benefit and matching nontruncated RCTs from systematic reviews. Independent re-viewers with medical content expertise, working blinded to trial results, judged the eligibility of the nontruncated RCTs based on their similarity to the truncated RCTs."

# From Abstract Results

- Large differences in treatment effect size between truncated and nontruncated RCTs occurred ….
- In 39 of the 63 questions (62%), the pooled effects of the nontruncated RCTs failed to demonstrate significant benefit.

```
Control Rate=     0.6000
 Exper Rate =     0.7500

Accrual Rate (pts/month):   15.00
        Number of Sims        5000
    Minimum Sample Size        150
    Maximum Sample Size        300
                    CV     0.0250
          Cuts     0.9500   0.0500

                    Mean       SD
    Sample Size   217.45     59.78

                  Lose       Win
   Success       0.009     0.639
       Cap       0.083     0.152
  Futility       0.116     0.000
     Total       0.209     0.791

Look      Lose       Win      Total
 150     0.044     0.260     0.304
 175     0.017     0.100     0.117
 200     0.012     0.086     0.098
 225     0.016     0.068     0.084
 250     0.018     0.067     0.085
 275     0.019     0.057     0.076
 300     0.083     0.152     0.235
 Tot     0.209     0.791     1.000
```

- Previous example
- Truth is 15% benefit
- But 8.3% of time trial goes to maximum … and fails.
- The reason it goes to max is because data is ambiguous

# S. Berry, Carlin, Connor

"To illustrate the issue, consider a clinical trial in which analysis is as follows: participants found to be performing better are retrospectively placed in the experimental group and participants found not to be performing well are retrospectively placed in the control group; a statistically significant difference in outcome is found when the groups are compared. It is clear that post-treatment selection of participants, based on their outcomes, would be responsible for any observed difference."

"This is logically equivalent to the analysis reported by Bassler et al."

# Goodman, D. Berry, Wittes

"So comparing the truncated trials to the nontruncated trials is similar to comparing completed trials with large effects with those with lower effects. The difference the authors observed was both predictable and uninformative."

"Bias is a property of study procedures; it is not logically applicable to a subset of results."

Goodman SN. Systematic reviews are not biased by results from trials stopped early for benefit. *J Clin Epidemiol.* 2008;61(1):95-96.

$p_t = 0.8$ vs. $p_c = 0.6$

n=180 → 80% Power

- What is average effect size in the statistically significant trials?

$p_t = 0.8$ vs. $p_c = 0.6$

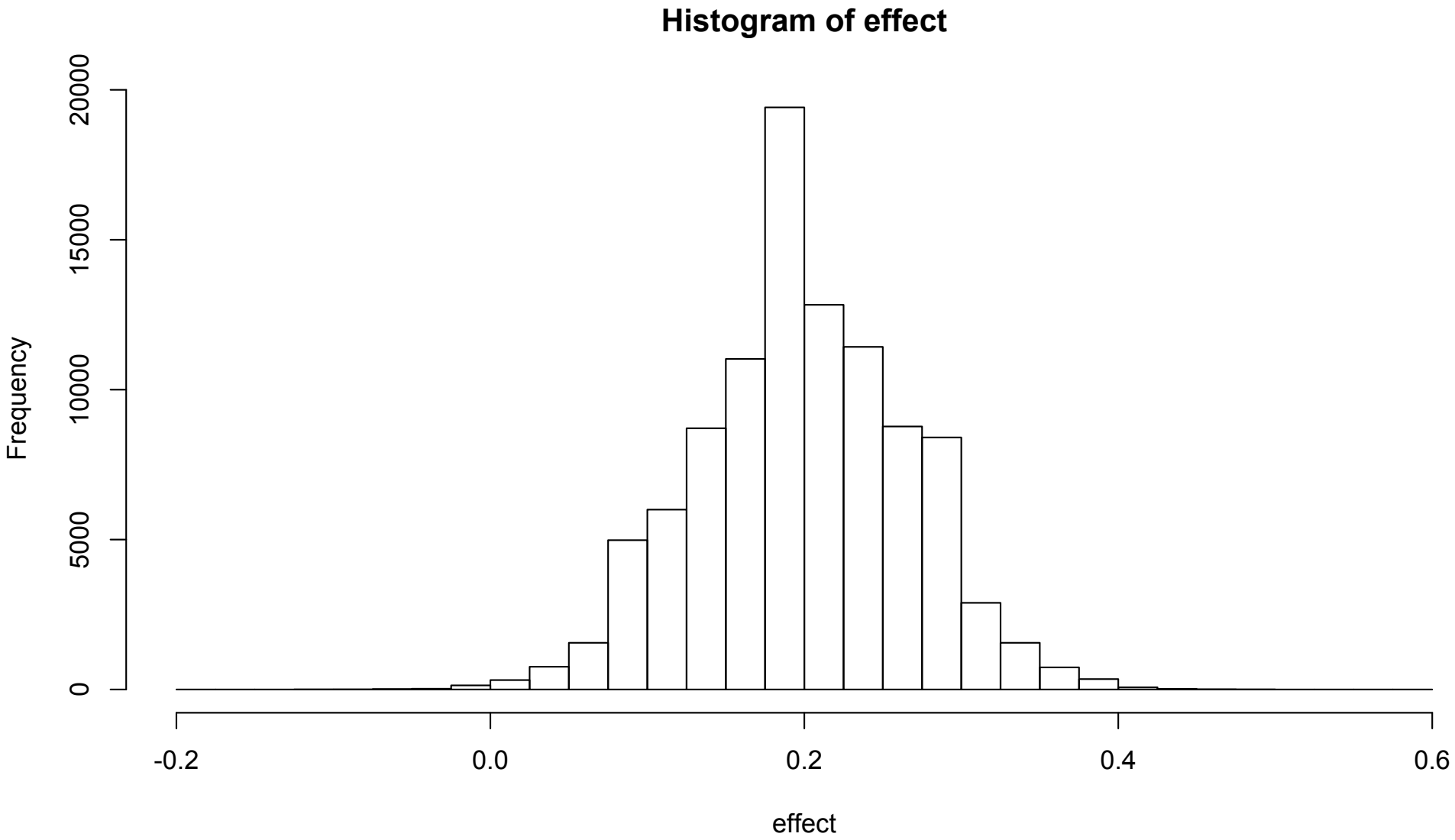
n=180 → 80% Power

- What is average effect size in the statistically significant trials?
- What is the average effect size in 100,000 simulated trials?

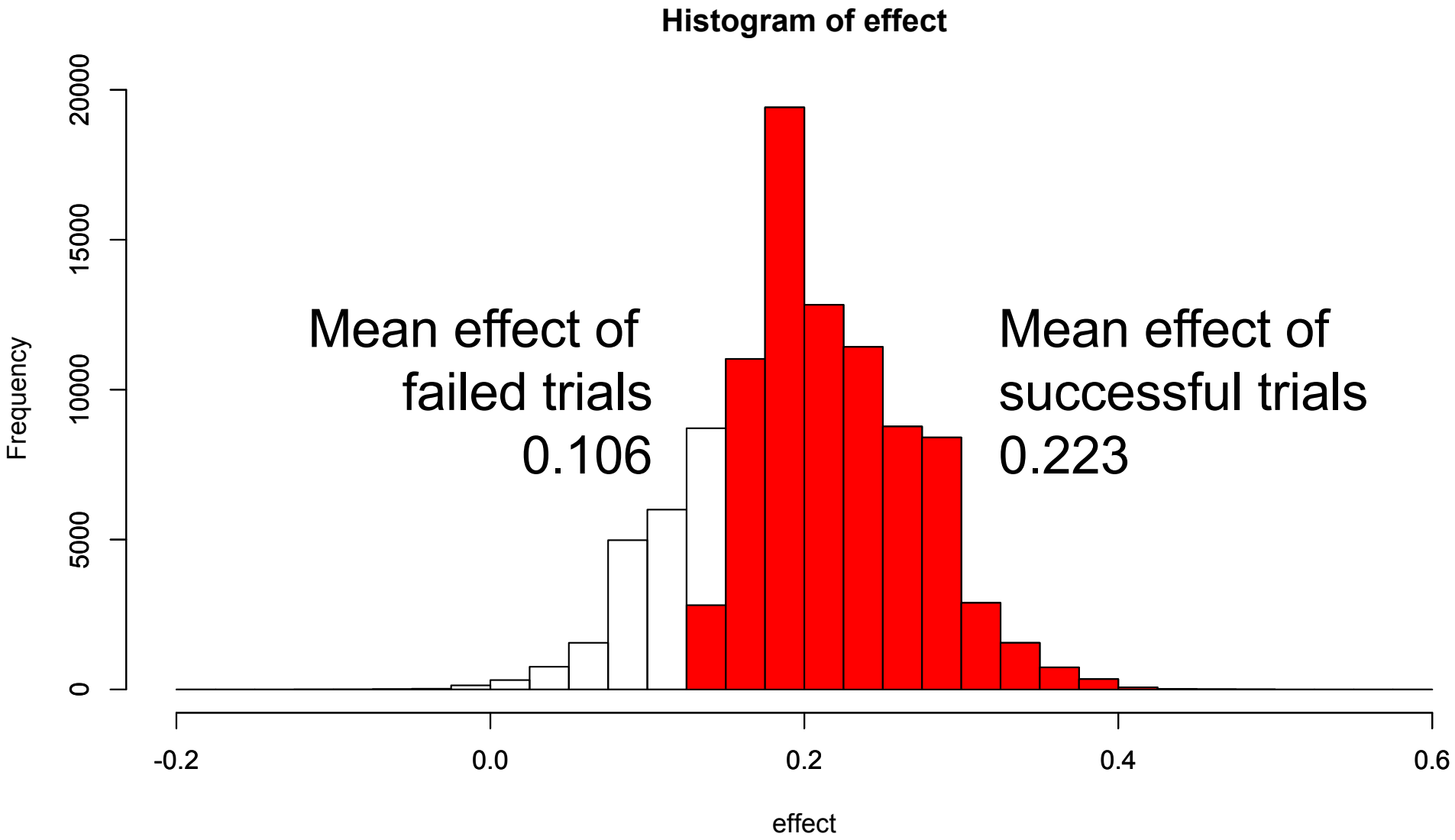
```
pvalue <- NULL; effect <- NULL

for(i in 1:100000){
x.c <- rbinom(1, 90, 0.6)
x.t <- rbinom(1, 90, 0.8)
mat <- rbind(c(x.c, 90-x.c), c(x.t, 90-x.t))
test <- chisq.test(mat)
pvalue[i] <- test$p.value
effect[i] <- x.t/90 - x.c/90
print(i)
}

hist(effect, breaks=seq(-.2, .6, by=0.025))
par(new=T)
hist(effect[pvalue<0.05], breaks=seq(-.2, .6, by=0.025), col=2, main=" ", xlab="
", ylab= " " )

> mean(pvalue < 0.05)     ### CHECK power = 80%
[1] 0.80313
> mean(effect)            ### CHECK mean effect = 0.20
[1] 0.2003593
>
> mean(effect[pvalue < 0.05])
[1] 0.2233924
> mean(effect[pvalue >= 0.05])
[1] 0.1063962
>
> 0.80 * .2233924 + 0.20 * 0.1063962
[1] 0.1999932
```

```
count <- 0
outcome <- matrix(nrow=8281, ncol=5)
for(xc in 0:90){
  for(xt in 0:90){
  count <- count + 1
  prob.of.pair <- dbinom(xc, 90, 0.6) * dbinom(xt, 90, 0.8)

  mat <- rbind(c(xc, 90-xc), c(xt, 90-xt))
  test <- chisq.test(mat)
  effect <- xt/90 - xc/90

  outcome[count, ] <- c(xc, xt, prob.of.pair, test$p.value, effect)
  print(c(xc, xt))
}}

outcome <- data.frame(outcome)
names(outcome) <- c("xc","xt","pr","pvalue","effect")

> sum(outcome$pr[outcome$pvalue < 0.05])
[1] 0.80168

> sum((outcome$effect * outcome$pr) [outcome$pvalue < 0.05]) /
sum(outcome$pr[outcome$pvalue<0.05])
[1] 0.2231661

> sum((outcome$effect * outcome$pr) [outcome$pvalue > 0.05]) /
sum(outcome$pr[outcome$pvalue>0.05])
[1] 0.1063544
```

# 100k sims 0.8 vs. 0.6, n=180



# 100k sims 0.8 vs. 0.6, n=180



# Revisit Previous Example

- Binary outcome
- Adaptive trial from 150 to 300 patients
- Expected difference 60% vs. 80%

Control Rate= 0.6000
Exper Rate = 0.8000

Accrual Rate (pts/month): 15.00
Number of Sims 1000
Minimum Sample Size 150
Maximum Sample Size 300
CV 0.0250
Cuts 0.9500 0.1000

| | Mean | SD |
|---|---|---|
| Sample Size | 183.20 | 48.53 |

| | Lose | Win |
|---|---|---|
| Success | 0.001 | 0.892 |
| Cap | 0.015 | 0.065 |
| Futility | 0.027 | 0.000 |
| Total | 0.043 | 0.957 |

| Look | Lose | Win | Total |
|---|---|---|---|
| 150 | 0.017 | 0.546 | 0.564 |
| 175 | 0.006 | 0.118 | 0.124 |
| 200 | 0.001 | 0.093 | 0.094 |
| 225 | 0.000 | 0.054 | 0.054 |
| 250 | 0.002 | 0.049 | 0.051 |
| 275 | 0.002 | 0.032 | 0.034 |
| 300 | 0.015 | 0.065 | 0.080 |
| Tot | 0.043 | 0.957 | 1.000 |

Control Rate= 0.6000
Exper Rate = 0.6000

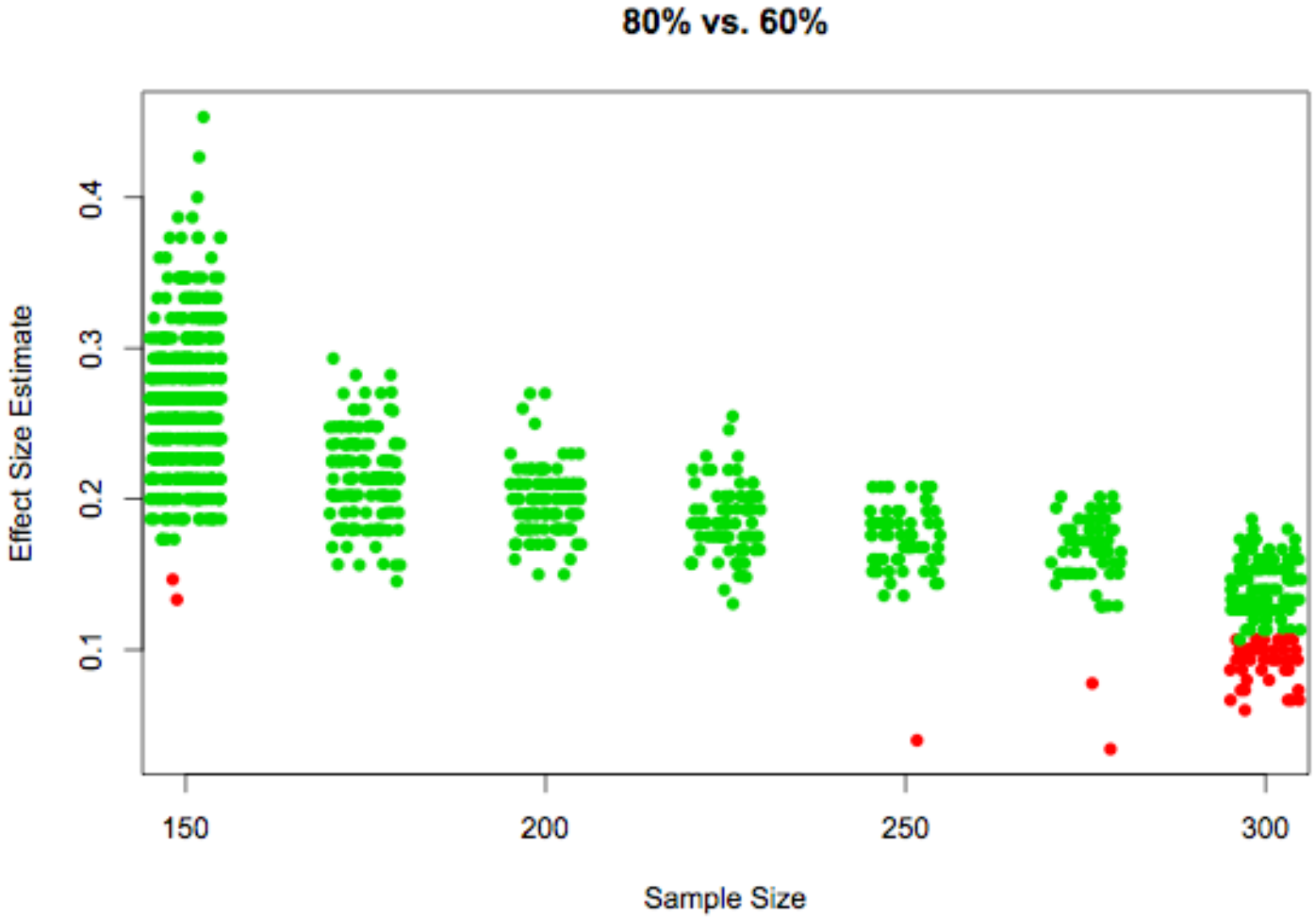
Accrual Rate (pts/month): 15.00
Number of Sims 1000
Minimum Sample Size 150
Maximum Sample Size 300
CV 0.0250
Cuts 0.9500 0.1000

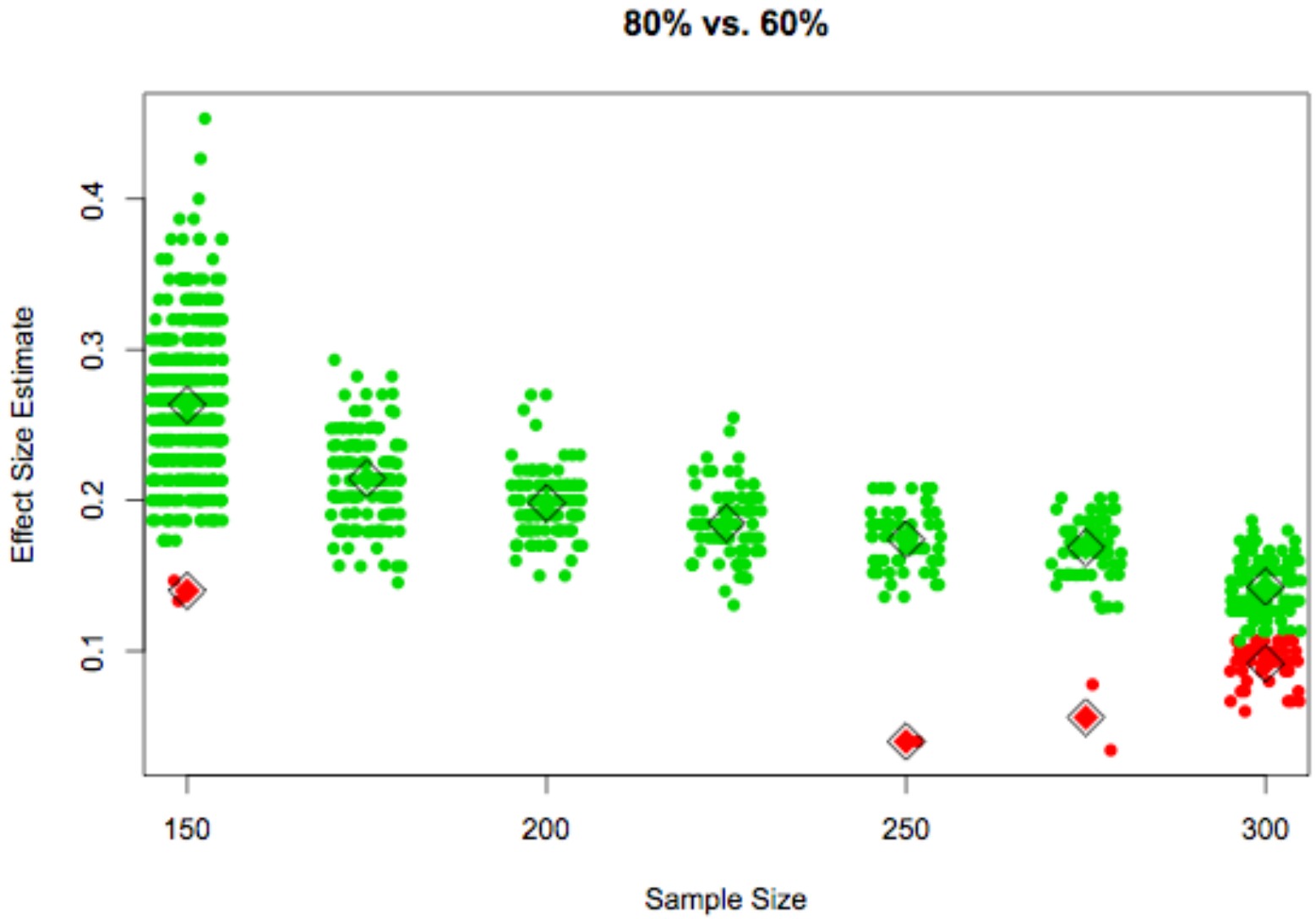
| | Mean | SD |
|---|---|---|
| Sample Size | 176.31 | 44.02 |

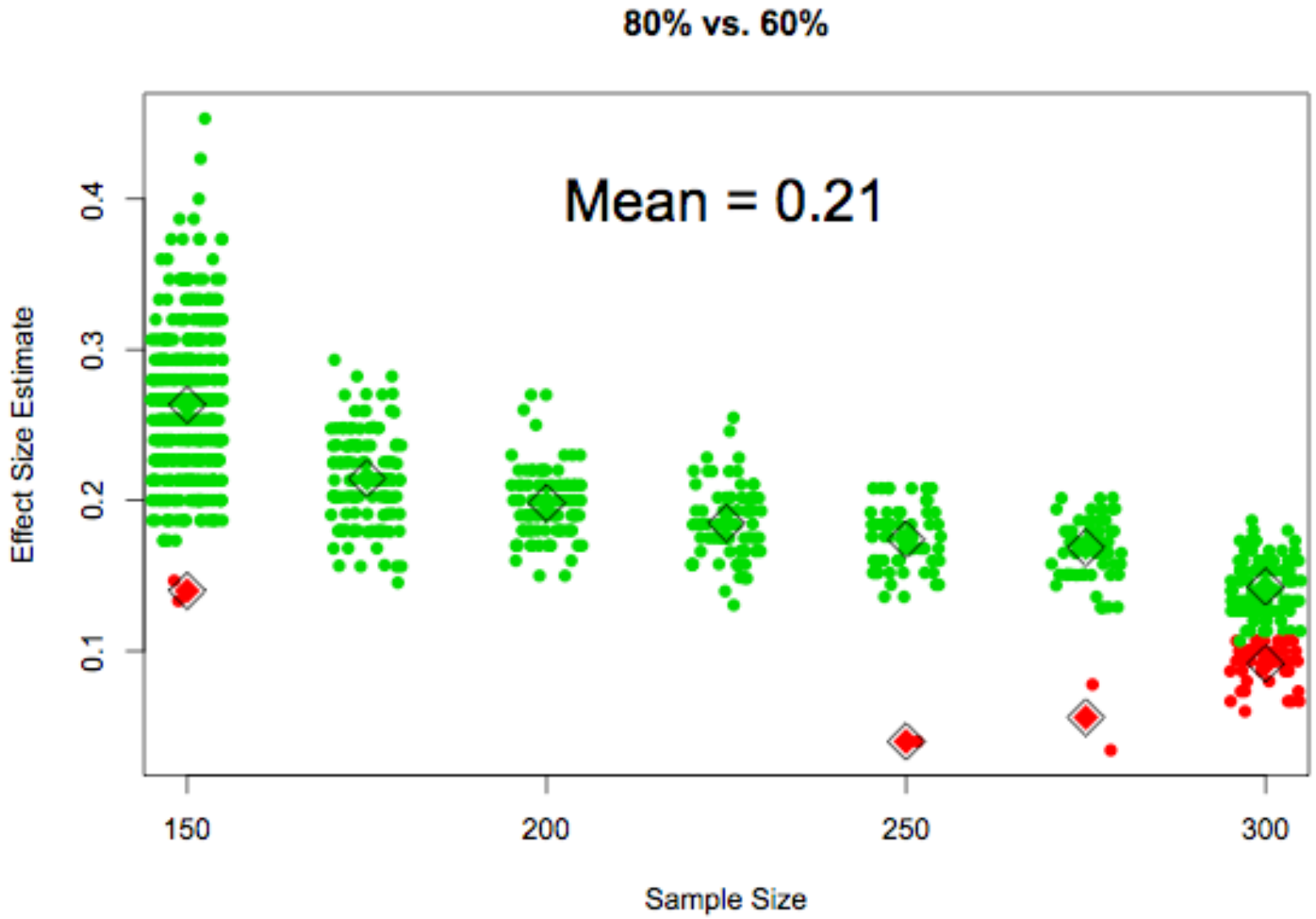
| | Lose | Win |
|---|---|---|
| Success | 0.002 | 0.019 |
| Cap | 0.041 | 0.009 |
| Futility | 0.929 | 0.000 |
| Total | 0.972 | 0.028 |

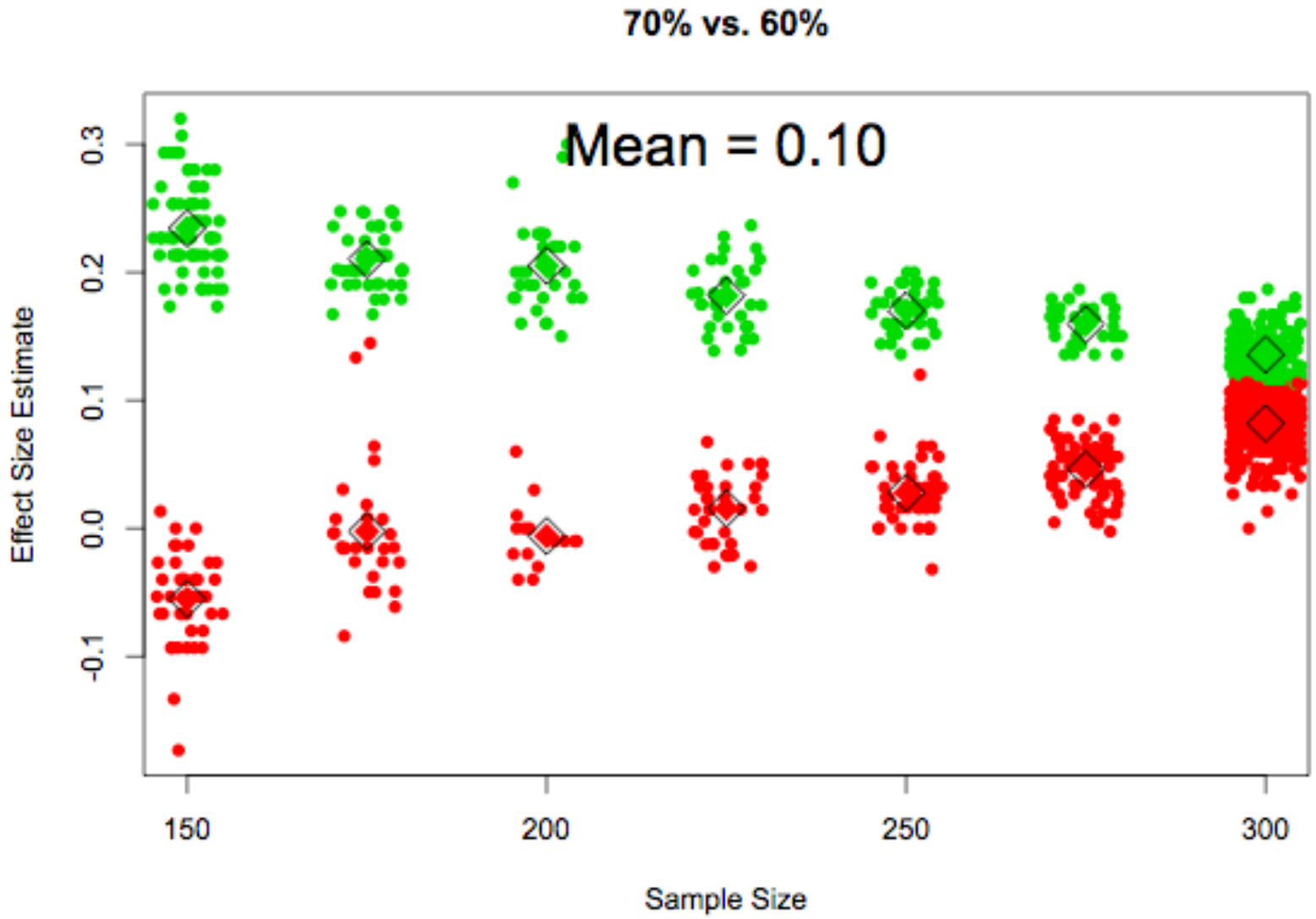
| Look | Lose | Win | Total |
|---|---|---|---|
| 150 | 0.634 | 0.006 | 0.640 |
| 175 | 0.103 | 0.004 | 0.107 |
| 200 | 0.073 | 0.003 | 0.076 |
| 225 | 0.047 | 0.003 | 0.050 |
| 250 | 0.042 | 0.002 | 0.044 |
| 275 | 0.033 | 0.001 | 0.034 |
| 300 | 0.041 | 0.009 | 0.050 |
| Tot | 0.972 | 0.028 | 1.000 |

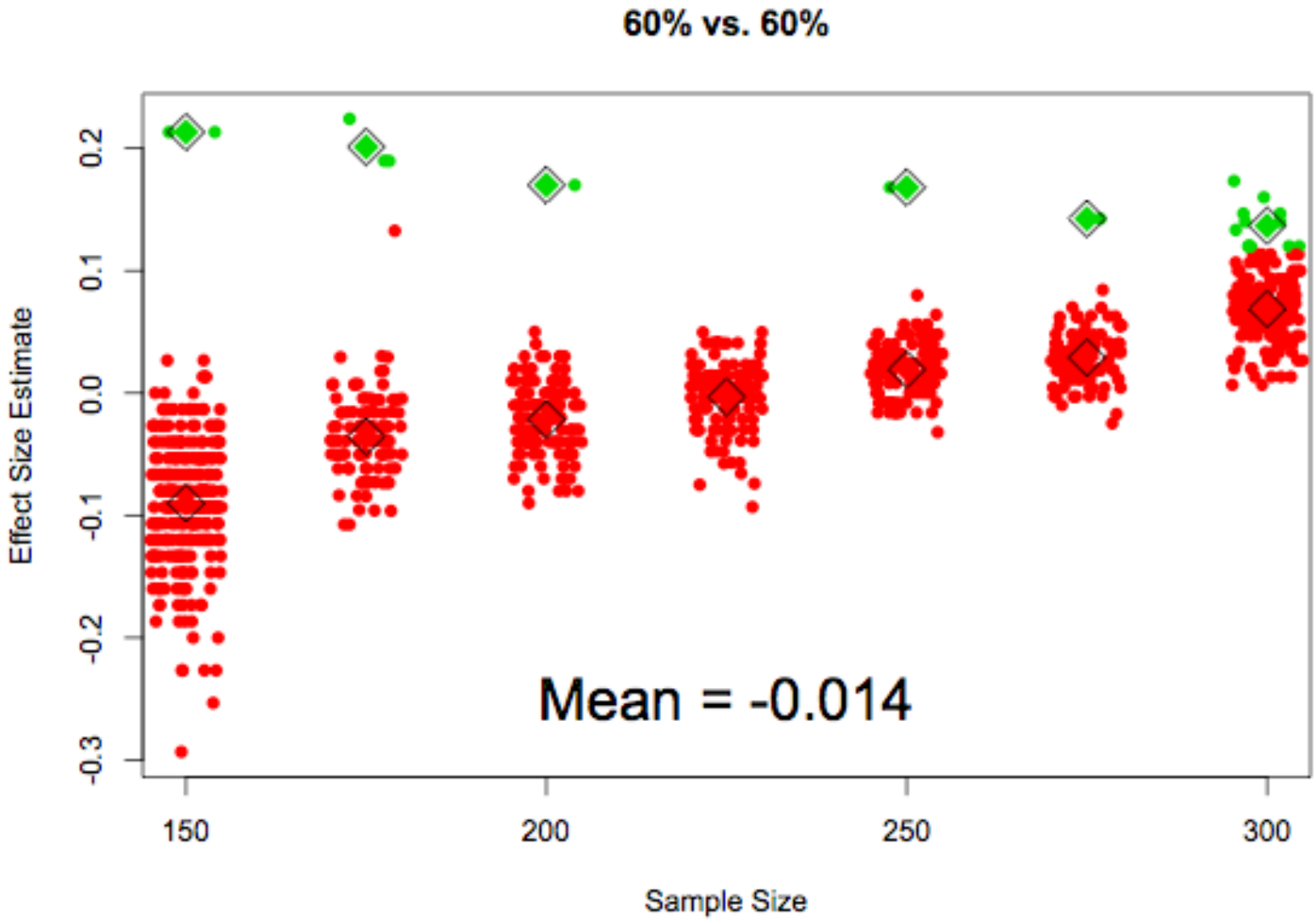
# Goldilocks Example

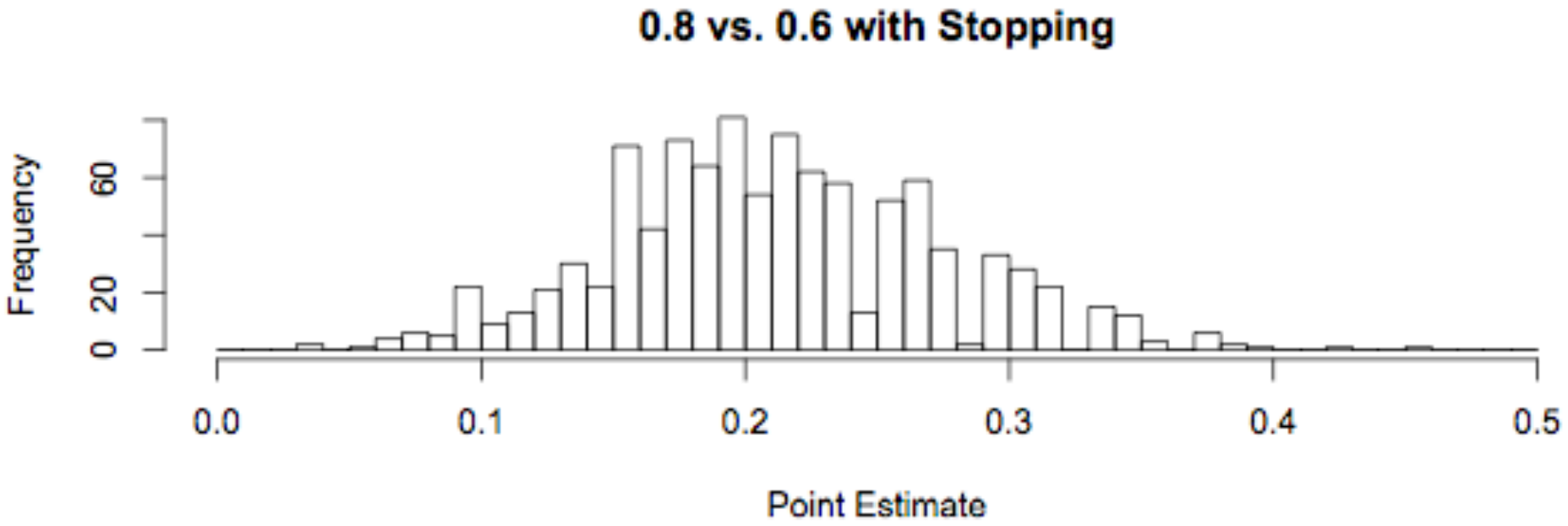


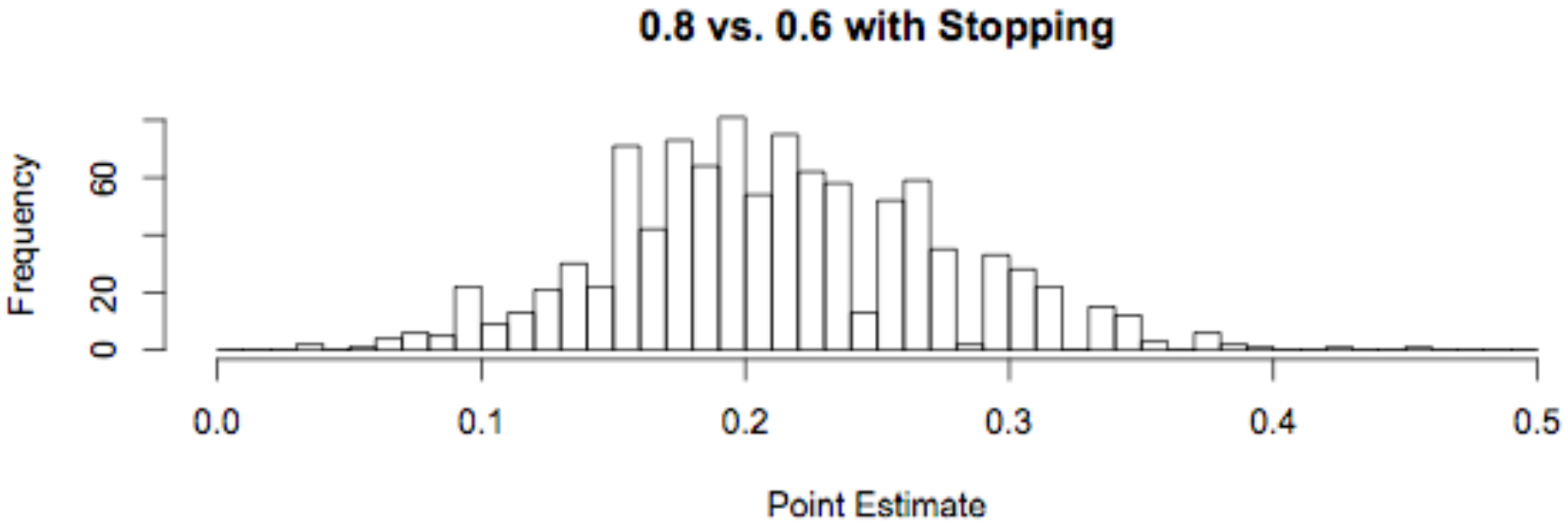

80% vs. 60%
Effect Size Estimate
0.1
0.2
0.3
0.4
150
200
250
300
Sample Size

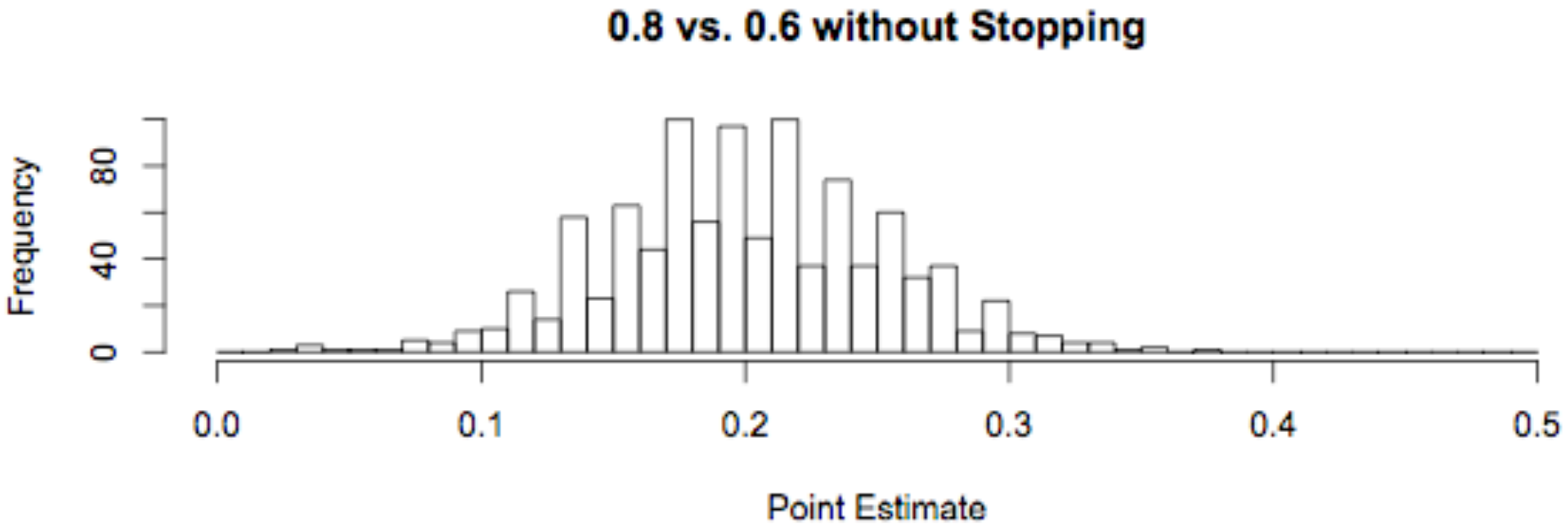

80% vs. 60%
Mean = 0.21
Effect Size Estimate
0.1
0.2
0.3
0.4
150
200
250
300
Sample Size

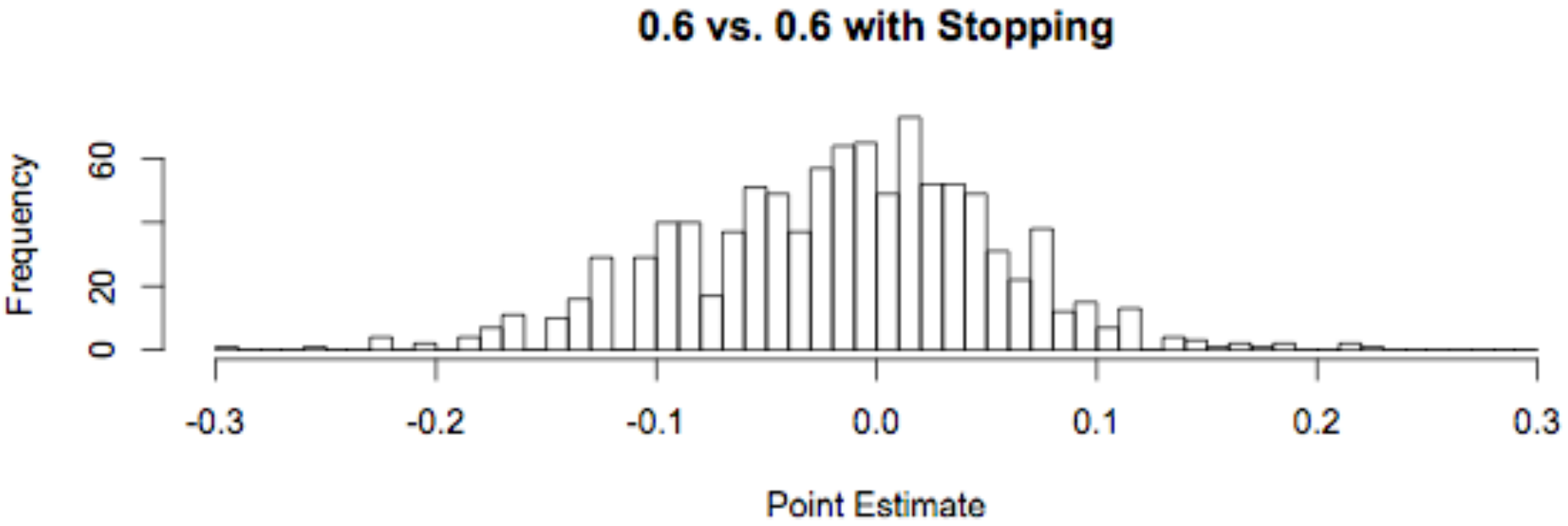

70% vs. 60%
Mean = 0.10
Effect Size Estimate
0.3
0.2
0.1
0.0
-0.1
150
200
250
300
Sample Size

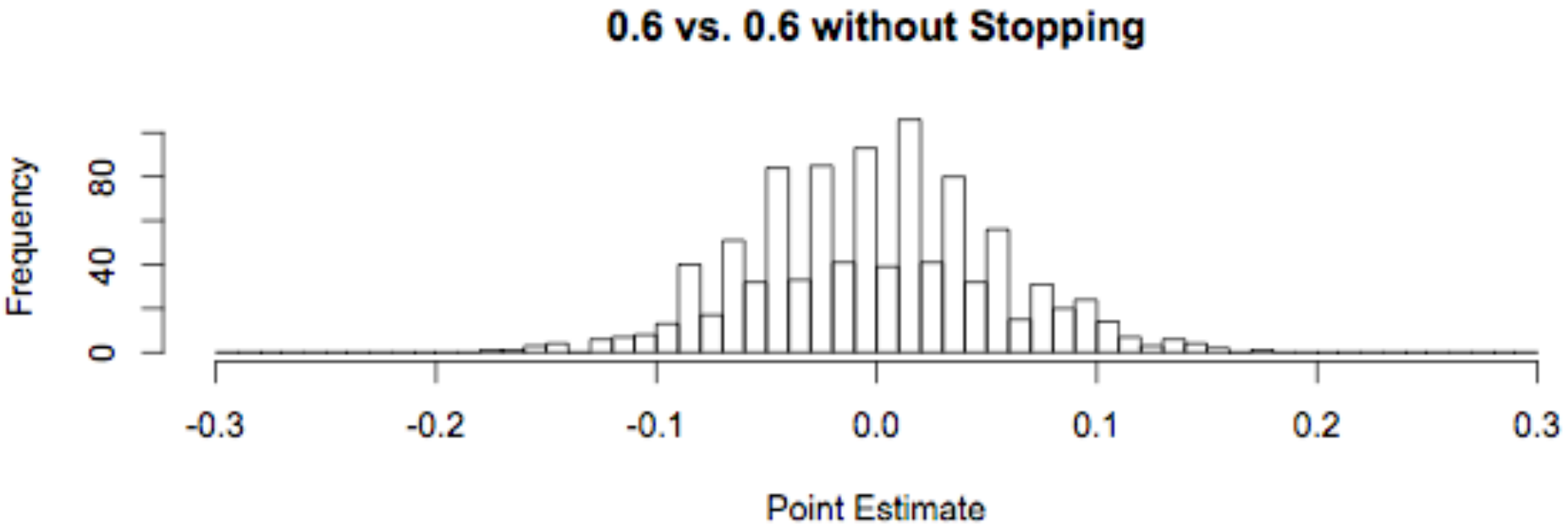

60% vs. 60%
Effect Size Estimate
0.2
0.1
0.0
-0.1
-0.2
-0.3
Mean = -0.014
150
200
250
300
Sample Size


0.8 vs. 0.6 with Stopping
Frequency
0
20
60
0.0
0.1
0.2
0.3
0.4
0.5
Point Estimate


0.8 vs. 0.6 with Stopping
Frequency
0
20
60
0.0
0.1
0.2
0.3
0.4
0.5
Point Estimate


0.8 vs. 0.6 without Stopping
Frequency
0
40
80
0.0
0.1
0.2
0.3
0.4
0.5
Point Estimate


0.6 vs. 0.6 with Stopping
Frequency
0
20
60
-0.3
-0.2
-0.1
0.0
0.1
0.2
0.3
Point Estimate

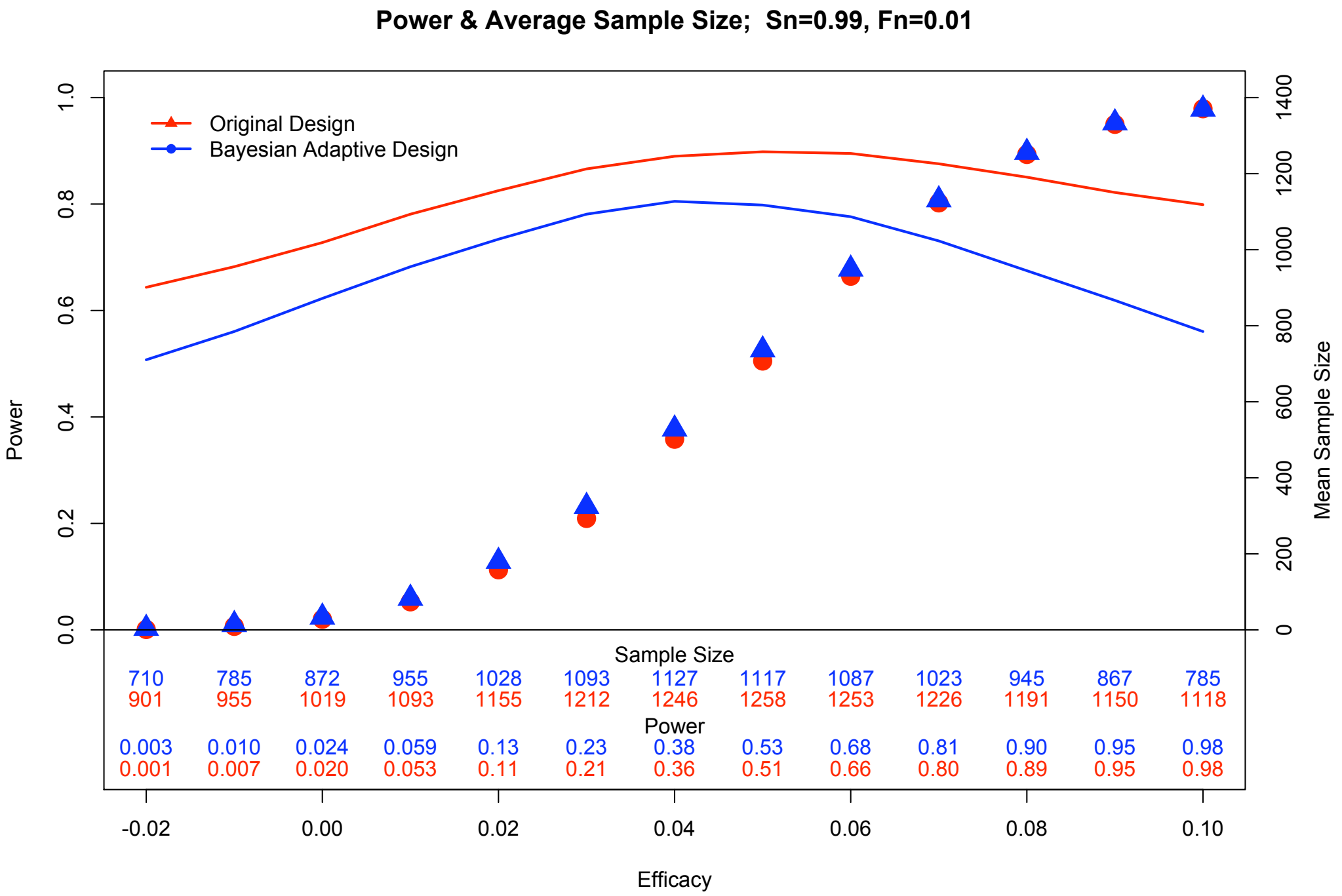

0.6 vs. 0.6 without Stopping
Frequency
0
40
80
-0.3
-0.2
-0.1
0.0
0.1
0.2
0.3
Point Estimate

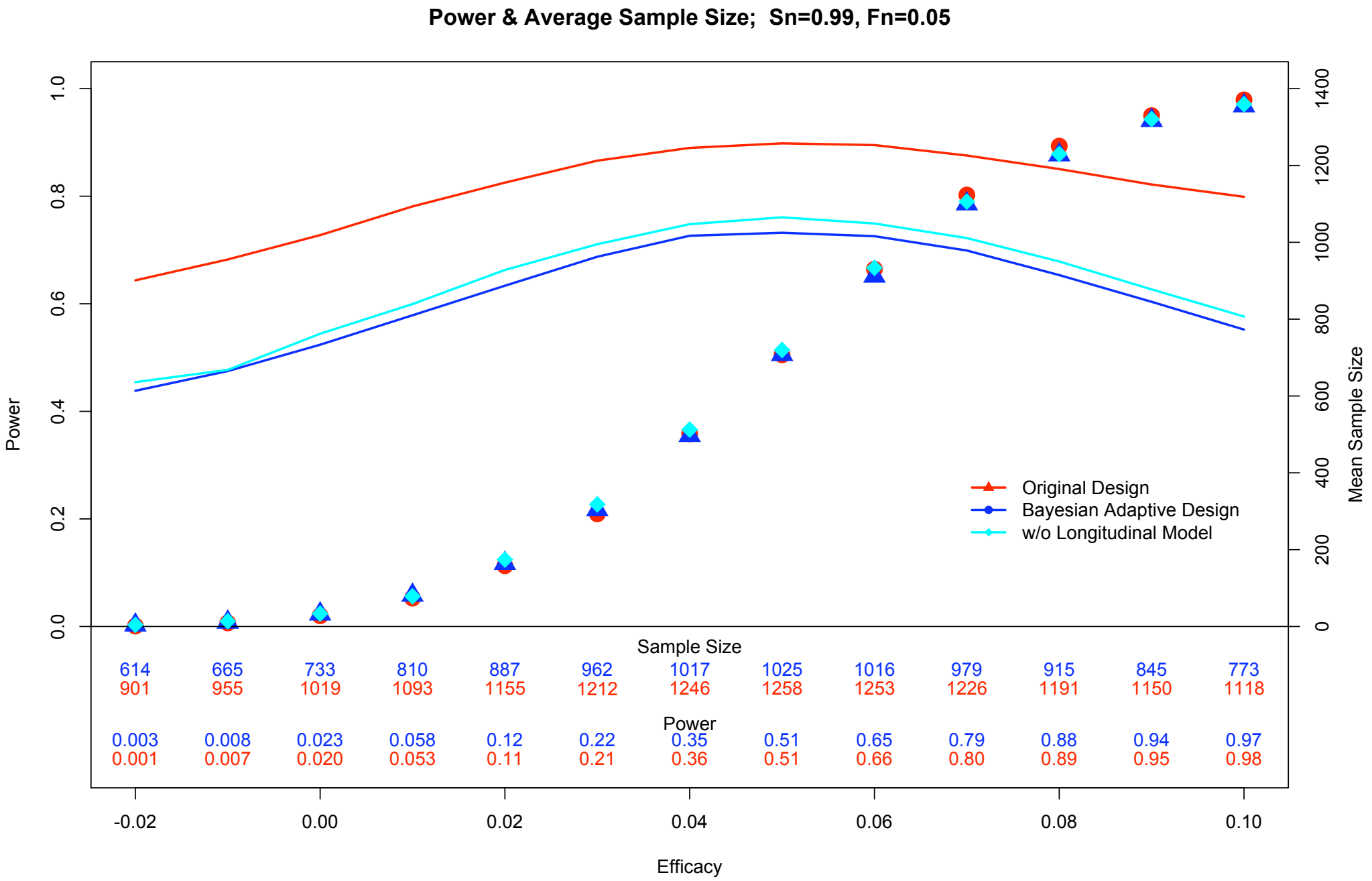
# Compare Distributions

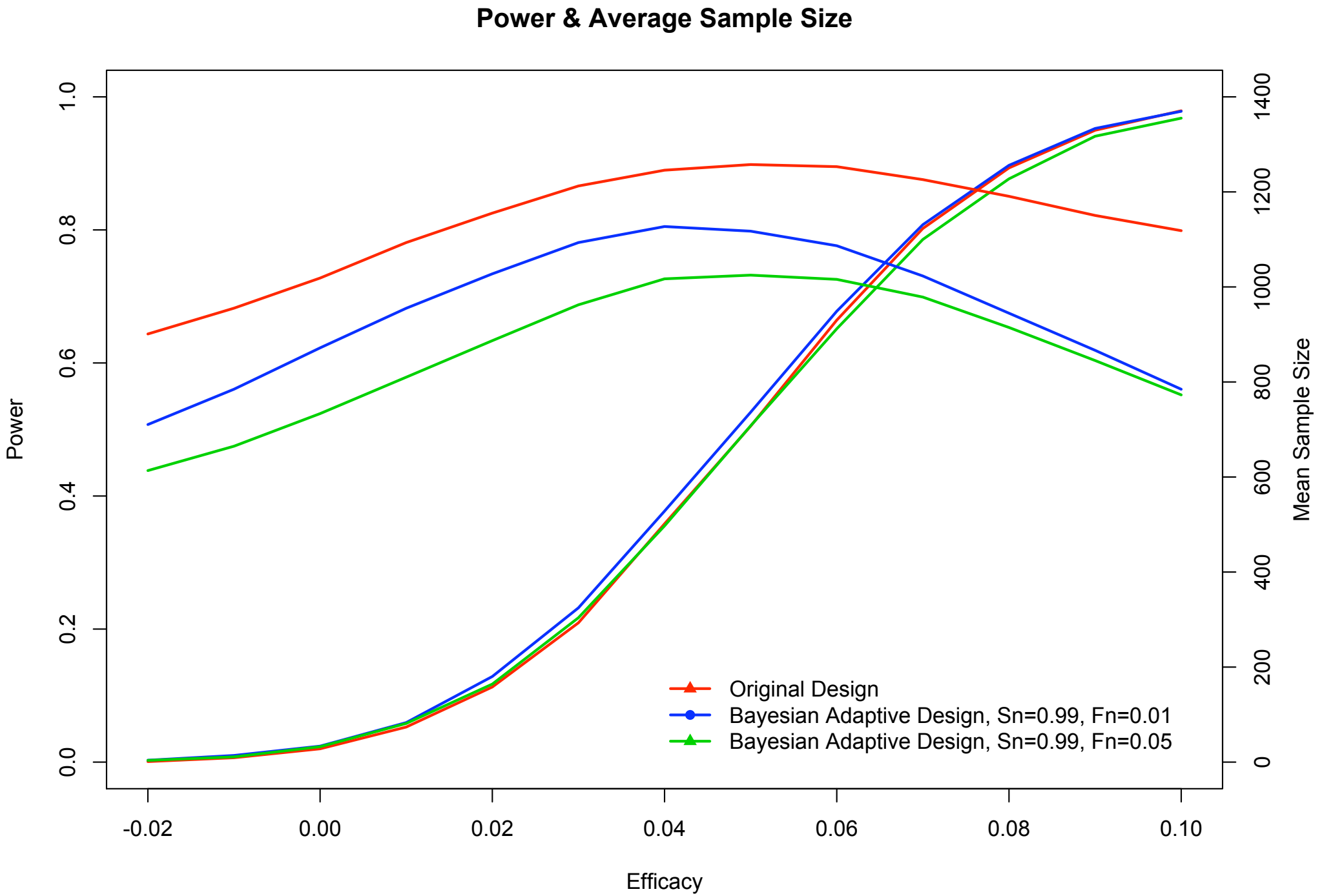
```
1000 simulations from pt = 0.8, pc = 0.6
                  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
Allow Stopping  0.034   0.172   0.208   0.212   0.253   0.453
   No Stopping  0.027   0.167   0.200   0.202   0.240   0.373

1000 simulations from pt = 0.6, pc = 0.6
                  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
Allow Stopping -0.293  -0.057  -0.010  -0.014   0.032  0.224
   No Stopping -0.173  -0.040   0.000   0.001   0.040  0.180
```

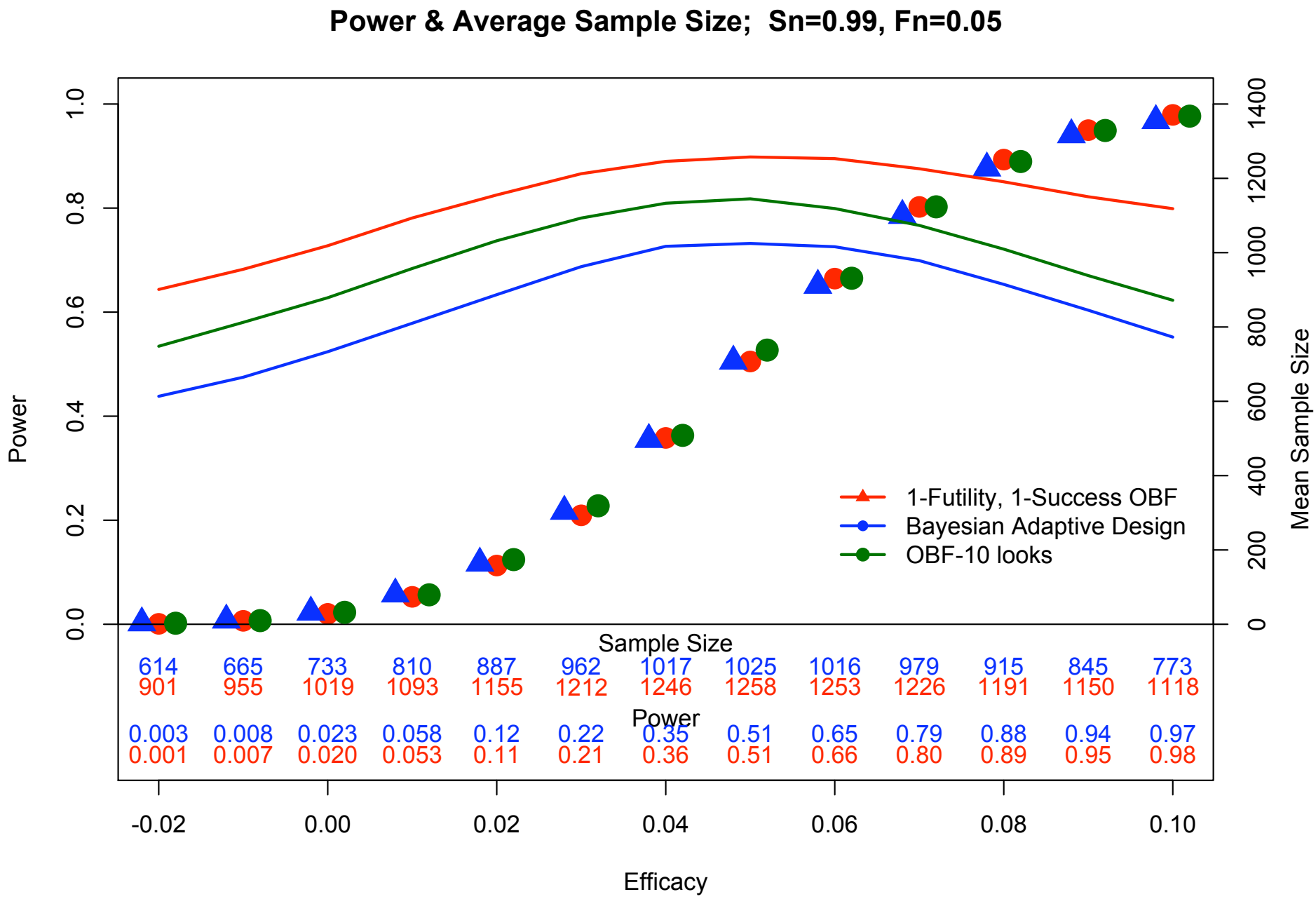
# Another Example

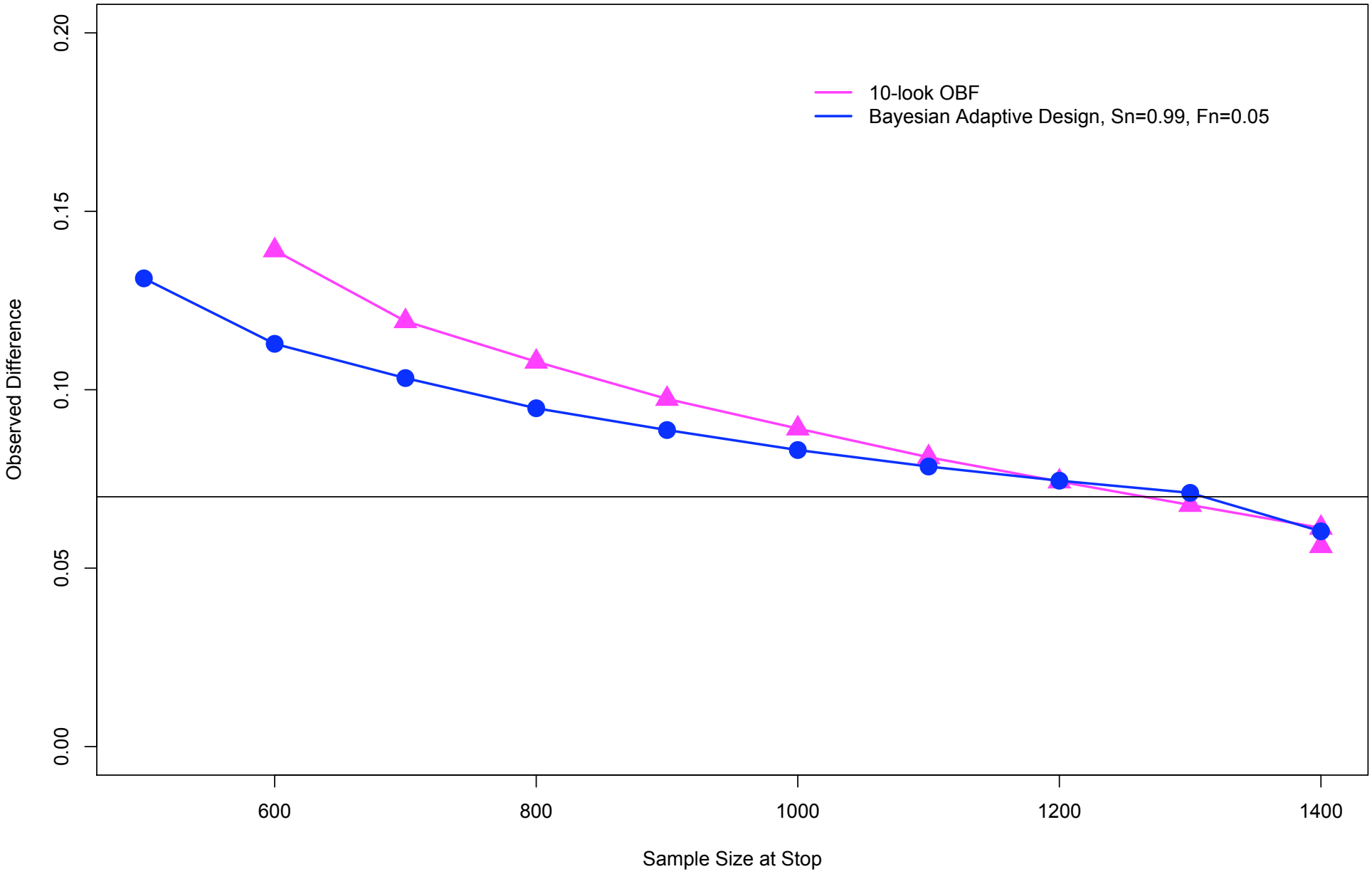
- SHINE Trial
  - Tight glycemic control in stroke
  - Designed using 1*-look OBF
  - Redesigned (NIH grant) using Bayesian adaptive trial
  - Decided to execute using 5-look OBF
  - Stored datasets for Bayesian re-evaluation
    - Connor JT, Broglio KB, Durkalski V, Meurer WJ, and Johnston KC. The Stroke Hyperglycemia Insulin Network Effort (SHINE) Trial. An Adaptive Trial Design Case Study. Trials. March 2015, Vol 16, No 72.
  - Final negative results just announced
    - Bayesian re-analysis forthcoming


Power & Average Sample Size; Sn=0.99, Fn=0.01
Original Design
Bayesian Adaptive Design
Power
Mean Sample Size
0.0
0.2
0.4
0.6
0.8
1.0
0
200
400
600
800
1000
1200
1400
Sample Size
710 785 872 955 1028 1093 1127 1117 1087 1023 945 867 785
901 955 1019 1093 1155 1212 1246 1258 1253 1226 1191 1150 1118
Power
0.003 0.010 0.024 0.059 0.13 0.23 0.38 0.53 0.68 0.81 0.90 0.95 0.98
0.001 0.007 0.020 0.053 0.11 0.21 0.36 0.51 0.66 0.80 0.89 0.95 0.98
-0.02
0.00
0.02
0.04
0.06
0.08
0.10
Efficacy


Power & Average Sample Size;  Sn=0.99, Fn=0.05
Power
Mean Sample Size
Original Design
Bayesian Adaptive Design
w/o Longitudinal Model
Sample Size
614 665 733 810 887 962 1017 1025 1016 979 915 845 773
901 955 1019 1093 1155 1212 1246 1258 1253 1226 1191 1150 1118
Power
0.003 0.008 0.023 0.058 0.12 0.22 0.35 0.51 0.65 0.79 0.88 0.94 0.97
0.001 0.007 0.020 0.053 0.11 0.21 0.36 0.51 0.66 0.80 0.89 0.95 0.98
Efficacy


Power & Average Sample Size
Power
Mean Sample Size
Efficacy
0.0
0.2
0.4
0.6
0.8
1.0
0
200
400
600
800
1000
1200
1400
-0.02
0.00
0.02
0.04
0.06
0.08
0.10
Original Design
Bayesian Adaptive Design, Sn=0.99, Fn=0.01
Bayesian Adaptive Design, Sn=0.99, Fn=0.05


Power & Average Sample Size;  Sn=0.99, Fn=0.05
Power
Mean Sample Size
Efficacy
1-Futility, 1-Success OBF
Bayesian Adaptive Design
OBF-10 looks
Sample Size
614 665 733 810 887 962 1017 1025 1016 979 915 845 773
901 955 1019 1093 1155 1212 1246 1258 1253 1226 1191 1150 1118
Power
0.003 0.008 0.023 0.058 0.12 0.22 0.35 0.51 0.65 0.79 0.88 0.94 0.97
0.001 0.007 0.020 0.053 0.11 0.21 0.36 0.51 0.66 0.80 0.89 0.95 0.98


10-look OBF
Bayesian Adaptive Design, Sn=0.99, Fn=0.05
Observed Difference
0.00
0.05
0.10
0.15
0.20
600
800
1000
1200
1400
Sample Size at Stop

# Combining Features

- Frequentist design uses 5 OBFs looks
  - Well understood
- Added blind sample size re-estimation prior to first OBF interim analysis
  - Well understood, Gould & Shih *Stats in Med* 1998
  - Pc = 0.25 vs. Pt = 0.32 Power = 0.83
  - Pc = 0.46 vs. Pt = 0.53 Power = 0.75
  - Increase sample size if pooled rate > 31%
- What happens if there is a big effect?

# Be Careful Combining Features

- Large effect size → High pooled rate
  - 30% vs. 50% (but analysis is unblinded, observe 40%)

# Be Careful Combining Features

- Large effect size → High pooled rate
  - 30% vs. 50% (but analysis is unblinded, observe 40%)
- High pooled rate → Increase in sample size
  - From 1400 to 1650

# Be Careful Combining Features

- Large effect size → High pooled rate
  - 30% vs. 50% (but analysis is unblinded, observe 40%)
- High pooled rate → Increase in sample size
  - From 1400 to 1650
- Increase in sample size → Delay 1st interim look
  - From 700 with data to 825 with data
  - About 4 months

# Be Careful Combining Features

- Large effect size → High pooled rate
  - 30% vs. 50% (but analysis is unblinded, observe 40%)
- High pooled rate → Increase in sample size
  - From 1400 to 1650
- Increase in sample size → Delay 1st interim look
  - From 700 with data to 825 with data
  - About 4 months
- Delay 1st interim look → Delay early stopping

# Be Careful Combining Features

- Large effect size → High pooled rate
  - 30% vs. 50% (but analysis is unblinded, observe 40%)
- High pooled rate → Increase in sample size
  - From 1400 to 1650
- Increase in sample size → Delay 1st interim look
  - From 700 with data to 825 with data
  - About 4 months
- Delay 1st interim look → Delay early stopping

- UNDERSTAND effects of combining features
- SIMULATE trials

# Summary

# Lessons

- Ensure minimum sample size will suffice
  - Not just statistical, but impactful
  - Company did a continue access protocol to get more patients during review, leading to panel
- Ensure data isn't coded optimistically
- Ensure inclusion / exclusion criteria rigorously followed
- Goldilocks trials gets the size 'just right' but that means you can be close to 'just wrong' if some data changes post hoc