

Early stopping for phase II cancer studies with time-to-event outcomes: a likelihood approach



Elizabeth Garrett-Mayer, PhD

*The Sidney Kimmel Comprehensive Cancer Center
Johns Hopkins University*

esg@jhu.edu

Motivation

- Oncology Phase II studies
 - Single arm
 - Evaluation of efficacy
- Historically,
 - 'clinical response' is the outcome of interest
 - Evaluated within several months (cycles) of enrollment
 - Early stopping often incorporated for futility

Early Stopping in Phase II studies:

Binary outcome

- Common design is Simon's two-stage (Simon, 1989)
 - Preserves type I and type II error
 - Procedure: Enroll N_1 patients (stage 1).
 - If x or more respond, enroll N_2 more (stage 2)
 - If fewer than x respond, stop.
 - Appropriate for **binary** responses

Early Stopping in Phase II studies: time-to-event outcomes

- Disease stabilization
- More common with novel treatments, targeted therapies
- Example: targeting stem cells
 - If treatment works, cancer does not progress
 - But, “bulk” may still remain
- Time-to-progression is relevant outcome
- But, takes a long time to evaluate...
- Examples where we commonly see PI's wanting to use PFS as primary outcome in phase II:
 - Pancreatic cancer (M Hidalgo, W Messersmith)
 - Lymphoma (Y Kasamon, R Ambinder)

One suggested approach

- Apply Simon's two-stage
- Example:
 - 1 year PFS of 0.30 versus 0.50 ($\alpha = \beta = 0.10$)
 - Enroll 20 patients
 - If 6 or more are PF at 1 year, enroll an additional 22 for a total of 42 patients.
- Study design
 - Assume trial will take 2 years to accrue (21 patients per year)
 - First 20 patients will be enrolled by end of year 1
 - 20th patient should be evaluable for 1 year PFS at end of year 2.

One suggested approach

- So, what's the problem?
 - Problem 1: By the end of year 2, almost all of the additional 22 patients will have been enrolled, yet the stage 1 patients have just become evaluable.
 - Problem 2: if the trial needs to be suspended after 20 patients (to wait for events), investigators may need to stop enrollment for 1 year.

Current approaches

- Bayesian approaches (Thall et al., 2005)
- Frequentist approaches (Case and Morgan, 2003)
- Ad hoc approaches
 - Use related outcome (e.g., clinical response)
 - Spend a little alpha early and evaluate:
 - At a prespecified time
 - When a prespecified number of patients have reached a landmark time (e.g. 1 year)
 - When a prespecified number of patients have been enrolled

Alternative approach

- Use likelihood-based approach (Royall (1997), Blume (2002))
- Not that different than Bayesian
 - Parametric model-based
 - No “penalties” for early looks
- But it is different
 - No prior information included
 - Early evaluations are relatively simple
 - Probability of misleading evidence controlled
 - Can make statements about probability of misleading evidence

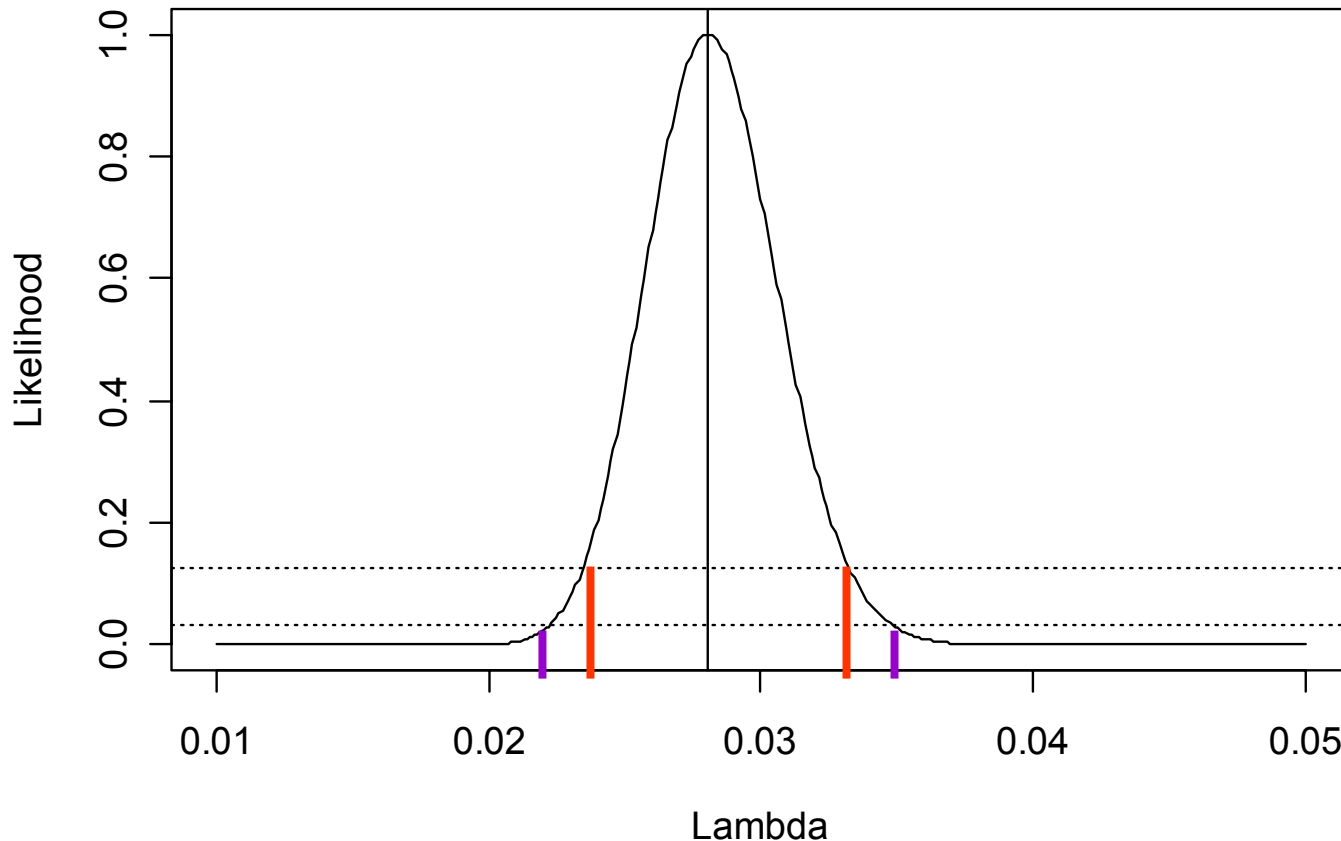
Law of Likelihood

If hypothesis A implies that the probability of observing some data X is $P_A(X)$, and hypothesis B implies that the probability is $P_B(X)$, then the observation $X=x$ is **evidence supporting A over B if $P_A(x) > P_B(x)$** , and the likelihood ratio, $P_A(x)/P_B(x)$, measures the strength of that evidence.

(Hacking 1965, Royall 1997)

Likelihood approach

- Determine “what the data say” about the parameter of interest
- Likelihood function: gives a picture of the data
- Likelihood intervals (LI): gives range of reasonable values for parameter of interest

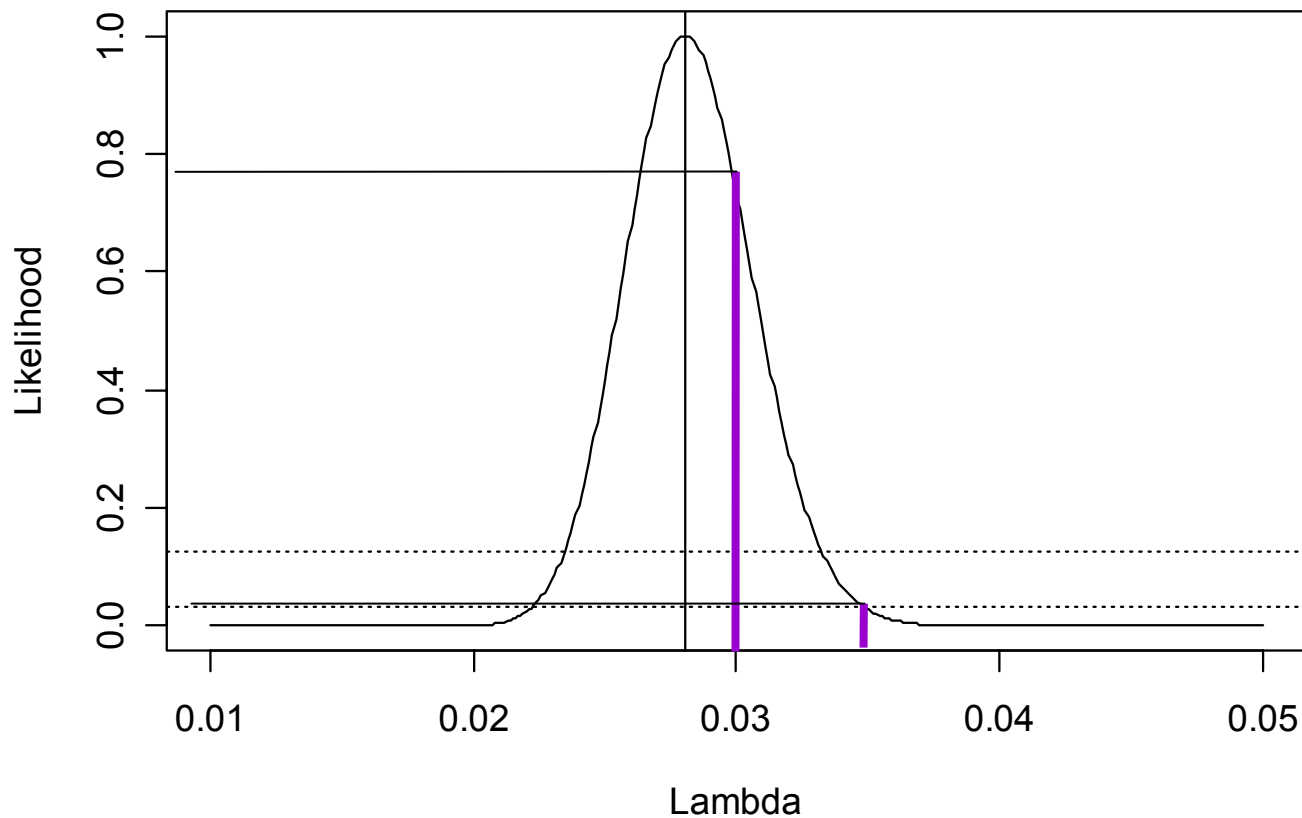


1/8
1/32

Likelihood approach

□ Likelihood ratios (LR)

- Take ratio of heights of L for different values of λ
- $L(\lambda=0.030)=0.78$; $L(\lambda=0.035)=0.03$.
- $LR = 26$



Likelihood-Based Approach

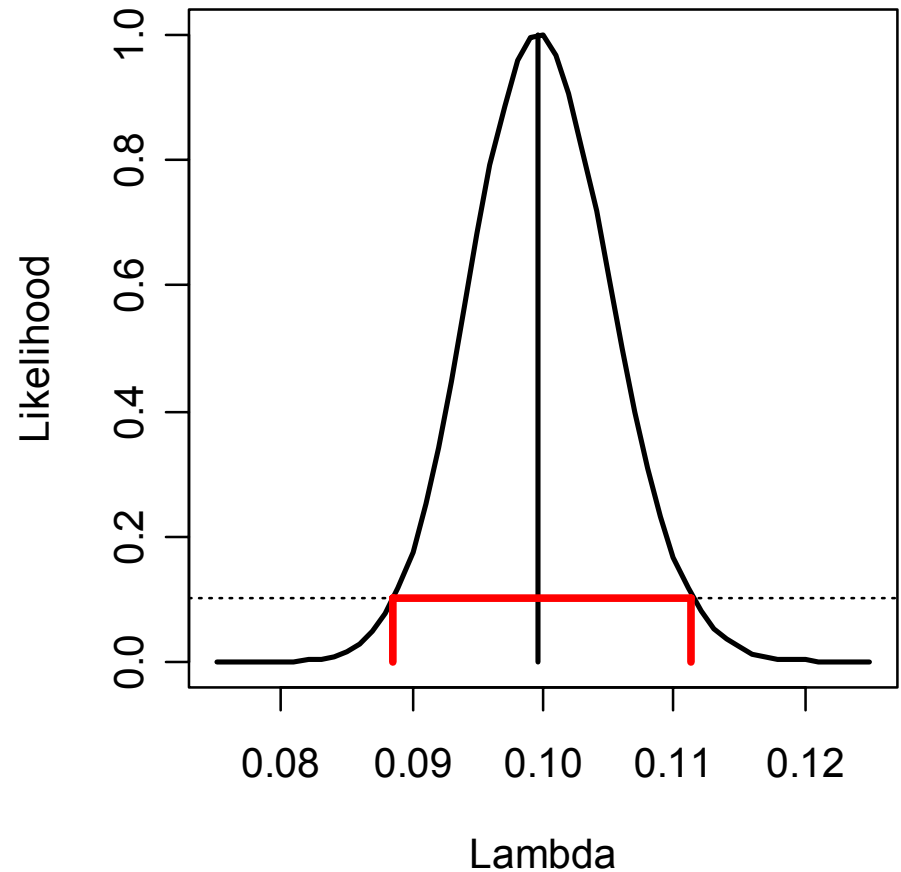
- Use likelihood ratio to determine if there is sufficient evidence in favor of the null or alternative hypothesis
- Can look at data as often as desired
- No penalty for multiple looks
- “Evidence-based”
- Error rates are bounded

Key difference in likelihood versus frequentist paradigm

- Consideration of the alternative hypothesis
- Frequentist p-values:
 - calculated assuming the null is true,
 - Have no regard for the alternative hypothesis
- Likelihood ratio:
 - Compares evidence for two hypotheses
 - Acceptance or rejection of null depends on the alternative

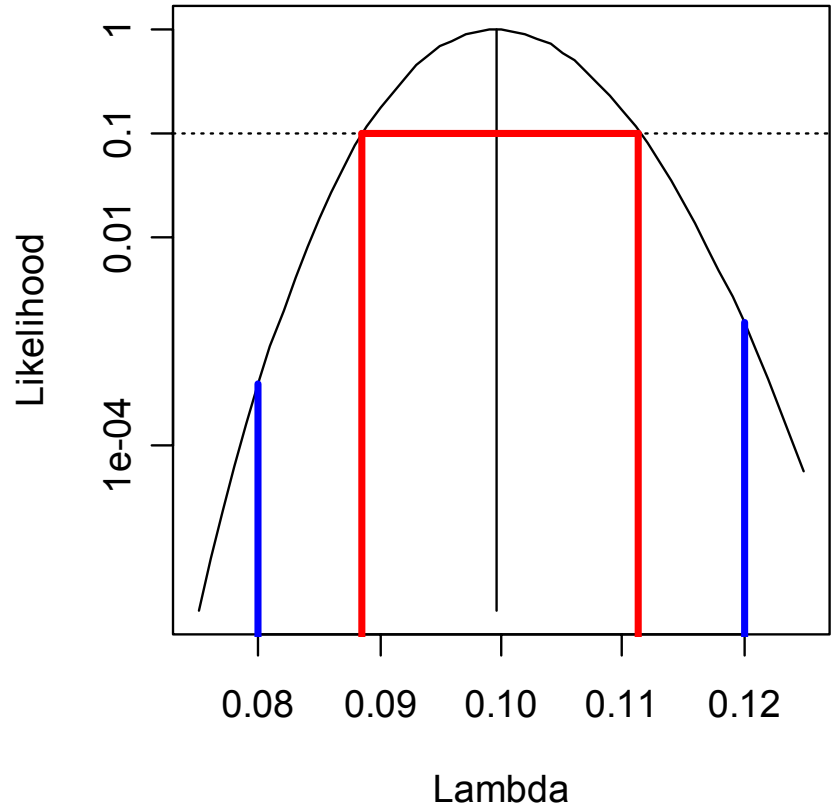
Example:

- Assume $H_0: \lambda = 0.12$
vs. $H_1: \lambda = 0.08$
- What if true $\lambda = 0.10$?
- Simulated data, $N=300$
- Frequentist:
 - $p = 0.01$
 - Reject the null
- Likelihood
 - $LR = 1/4$
 - Weak evidence for null



Example:

- Why?
- P-value looks for evidence against null
- LR compares evidence for both hypotheses
- When the “truth” is in the middle, which makes more sense?



Motivating Example

- New treatment for cancer
- Standard of care: 6 month progression-free survival (PFS) is 50%
- We'd like to see if the new treatment has 60% or greater 6 month PFS:

H_0 : PFS = 50% (null)

vs.

H_1 : PFS = 60% (alternative)

- Design issues:
 - Single arm
 - Time-to-event outcome

Likelihood Inference

- Weak evidence: at the end of the study, there is not sufficiently strong evidence in favor of either hypothesis
 - This can be controlled by choosing a large enough sample size
 - But, if neither hypothesis is correct, can end up with weak even if N is seemingly large (appropriate)
- Strong evidence
 - Correct evidence: strong evidence in favor of correct hypothesis
 - Misleading evidence: strong evidence in favor of the incorrect hypothesis.
 - This is our interest today: what is the probability of misleading evidence?
 - This is analogous to the alpha (type I) and beta (type II) errors that frequentists worry about

Misleading Evidence in Likelihood Paradigm

- Universal bound: Under H_0 ,

$$P\left(\frac{L_1}{L_0} \geq k\right) \leq 1/k \quad (\text{Birnbaum, 1962; Smith, 1953})$$

- In words, the probability that the likelihood ratio exceeds k in favor of the wrong hypothesis can be no larger than $1/k$.
- In certain cases, an even lower bound applies (Royall, 2000)
 - Difference between normal means
 - Large sample size
- Common choices for k are 8 (strong), 10, 32 (very strong).

Implications

- Important result: For a sequence of independent observations, the universal bound still holds (Robbins, 1970)
- Implication: We can look at the data as often as desired and our probability of misleading evidence is bounded
- That is, if $k=10$, the probability of misleading strong evidence is $\leq 10\%$
- Not bad! Considering $\beta = 10\text{-}20\%$ and $\alpha = 5\text{-}10\%$ in most studies

Early stopping in phase II TTE study

- Motivating Example
 - Single arm
 - Time-to-event outcome
 - Early stopping for futility
- Standard frequentist approach
 - Non-parametric (i.e., no model)
 - Kaplan-Meier estimate of 6 mo. PFS
 - “Robust”, but not powerful!
- Likelihood approach
 - **Requires a parametric model (like the Bayesians!)**

Model Choice Considerations

- Trade-off: One-parameter vs. >One-parameter model
 - Parsimony versus fit
 - Bias versus variance
- Small amount of data: cannot tolerate many parameters
- Exponential (one-parameter) obvious choice
- Some other options:
 - Weibull
 - Log-normal
 - Cure-rate

Critical Issue

- ❑ Decision to stop must be robust to model misspecification
- ❑ “Robustifying” likelihood (Royall and Tsou, 2003)
- ❑ Not appropriate here: exponential with censoring does not meet criteria
- ❑ Further study needed to see early stopping behavior when model is misspecified

Exponential Model and Likelihood

probability density: $f(t|\lambda) = \lambda e^{-\lambda t}$

survival function: $S(t) = e^{-\lambda t}$

Log - likelihood function: $L(\lambda; t, d) = \sum_{i=1}^N d_i \log \lambda - \lambda t_i$

Maximum likelihood estimate: $\hat{\lambda} = \frac{\sum d_i}{\sum t_i}$

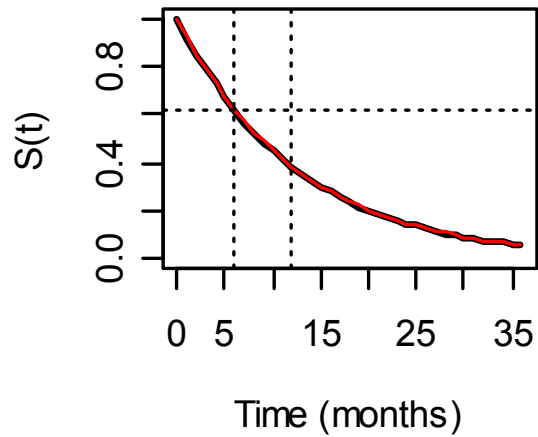
Simulations

- ❑ Need comparability across distributions of simulated data
- ❑ Chose underlying distributions with same 6 month survival
 - Exponential
 - Weibull: one with larger variance, one with smaller
 - Log-normal: one with larger variance, one with smaller
 - Cure-rate
- ❑ Working model: exponential distribution
- ❑ First set of simulations: data generated assuming treatment is **efficacious**
- ❑ Second set of simulations: data generated assuming treatment **lacks efficacy**

Comparison of underlying distributions

Black: true distn
Red: best exponential

Exponential, 0.08



Simulation Study 1

- Alternative is true: should not stop for futility most of the time
- Three (or more) ways to characterize hypotheses:

H_0 : 6 mo PFS = 49% vs. H_1 : 6 mo PFS = 62%

H_0 : $E(t) = 8.3$ mo vs. H_1 : $E(t) = 12.5$ mo

H_0 : $\lambda = 0.12$ vs. H_1 : $\lambda = 0.08$

- $N = 100$
- Starting with **25th** patient, analyze data every **5th** enrollment
- Censoring is assumed to be administrative
- 24 months of enrollment (assuming no early stopping)
- Total study time 36 months (24 month accrual, 12 month F.U.)
- Use likelihood ratio of 1/10 (i.e., $k=10$)

Frequentist Properties of Simulation Study 1

- $N=100$, $H_0: \lambda = 0.12$ vs. $H_1: \lambda = 0.08$
- Using exponential test and assuming exponential data:
 - Alpha = 5%
 - Power = 96%
- Using non-parametric test, and assuming exponential data:
 - Alpha = 5%
 - Power = 85%
- No interim analyses included

Stopping Rule

- Stop if the likelihood ratio $< 1/10$
- That is, if the ratio of the likelihood for the NULL to the ALTERNATIVE is 10, then stop.

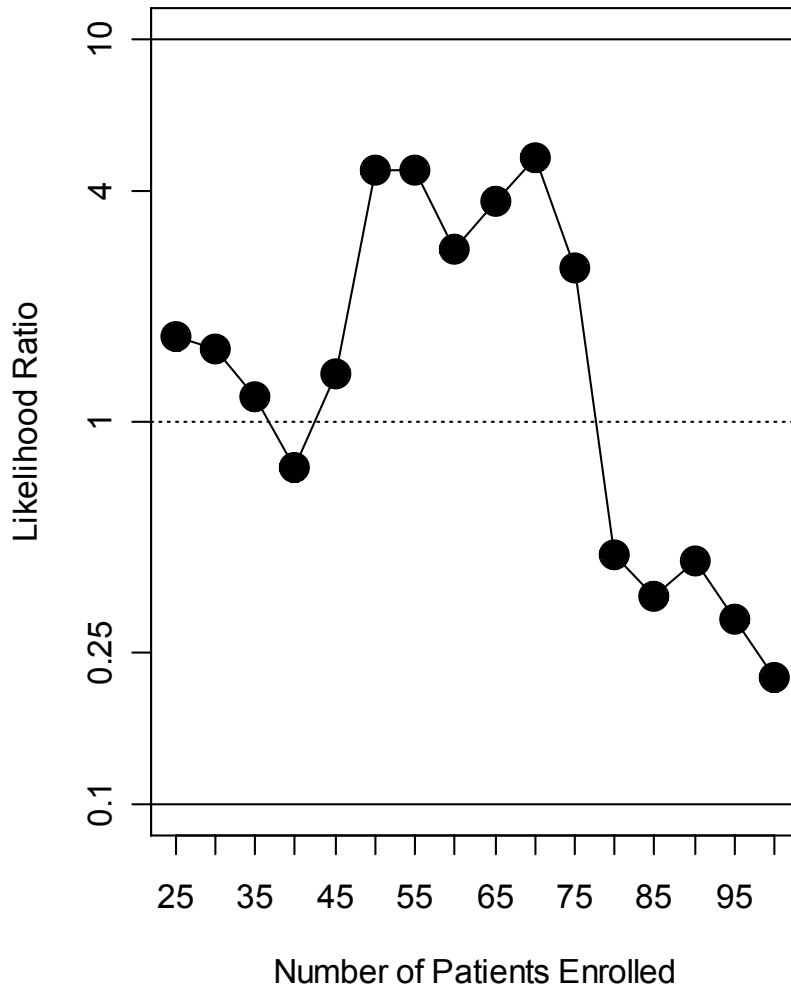
- Note 1: ONLY considering stopping for futility!
- Note 2: based on universal bound, we have a less than 10% chance of strong evidence in favor of the wrong hypothesis
- Note 3: based on Royall (2000), probably have even less than that....

Why not look before 25 patients?

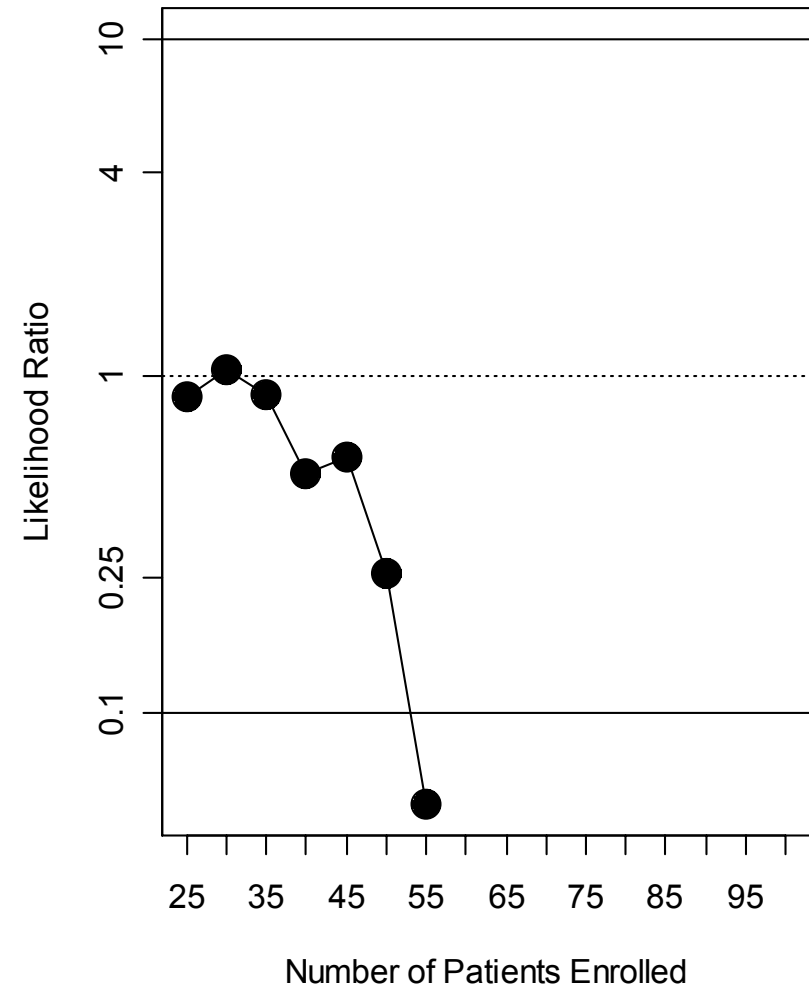
	End month 1	End month 2	End month 3	End month 4	End month 5	End month 6
Total enrolled	4	8	12	16	21	25
≥ 1 month f.u	0	4	8	12	16	21
≥ 2 month f.u	0	0	4	8	12	16
≥ 3 month f.u	0	0	0	4	8	12
≥ 4 month f.u	0	0	0	0	4	8
≥ 5 month f.u	0	0	0	0	0	4
≥ 6 month f.u	0	0	0	0	0	0

Simulated Data Examples

No stopping



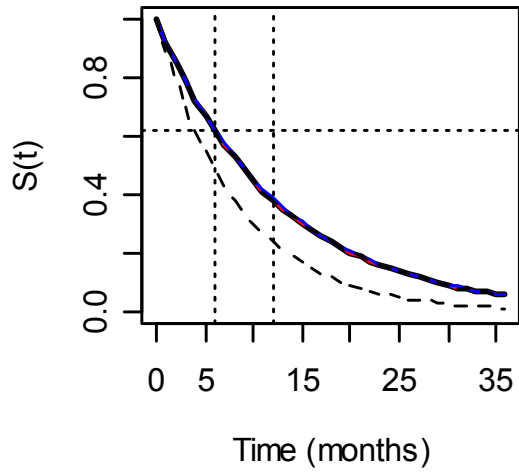
Stop at N=55



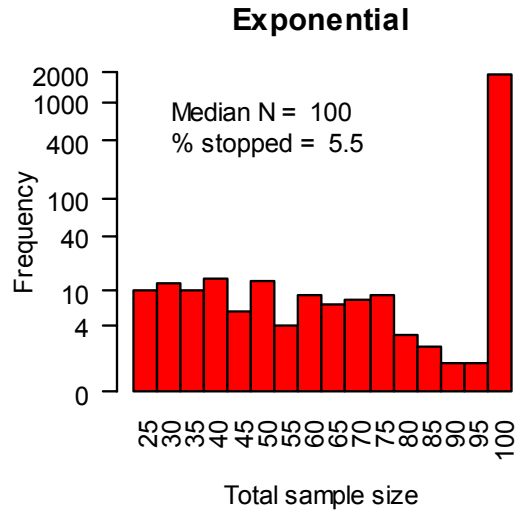
First set of simulations

Blue: 6 month estimate solid black: true distn
Red: 36 month estimate dashed: hypotheses

Exponential, 0.08



Early Stopping



Likelihood ratios:

	<1/32	[1/32, 1/10)	[1/10, 1)	[1,10)	[10,32)	>32
Exponential*	<0.01	0.05	0.01	0.05	0.05	0.82
Weibull 1	0.31	0.29	0.25	0.11	0.02	0.01
Log-Normal 1	0.30	0.24	0.27	0.14	0.03	0.02
Cure Rate	<0.01	0.02	<0.01	<0.01	<0.01	0.98
Weibull 2	<0.01	0.07	<0.01	<0.01	<0.01	0.92
Log-Normal 2	<0.01	0.05	<0.01	<0.01	<0.01	0.94

Frequentist Approach: Exponential Data

- Based on observed data (stopped and completed trials)
 - 84% of trials showed significant p-value (versus 87% with $LR > 10$)
 - Agreement of 96% for chosen hypothesis
 - Remaining 4% (81 trials)
 - LRT was significant
 - P-value was insignificant
 - All p-values that disagreed were between 0.051 and 0.0755

Simulation study 2

- Null hypothesis is true: should stop early for futility in most trials
- Three ways to characterize hypotheses:

H_0 : 6 mo PFS = 62% vs. H_1 : 6 mo PFS = 74%

H_0 : $E(t) = 12.5$ mo vs. H_1 : $E(t) = 20$ mo

H_0 : $\lambda = 0.08$ vs. H_1 : $\lambda = 0.05$

- $N = 100$
- Starting with **25th** patient, analyze data every **5th** enrollment
- Censoring is assumed to be administrative
- 24 months of enrollment (assuming no early stopping)
- Total study time 36 months (24 month accrual, 12 month F.U.)
- Use likelihood intervals of 1/10

Frequentist Properties of Simulation Study 2

- $N=100$, $H_0: \lambda = 0.08$ vs. $H_1: \lambda = 0.05$
- Using exponential test and assuming exponential data:
 - Alpha = 5%
 - Power = 98%
- Using non-parametric test, and assuming exponential data:
 - Alpha = 5%
 - Power = 78%
- No interim analyses included

Second set of simulations

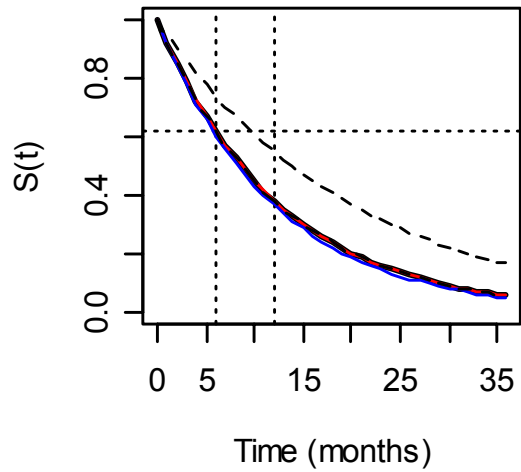
Blue: 12 month estimate

Red: 60 month estimate

solid black: true distn

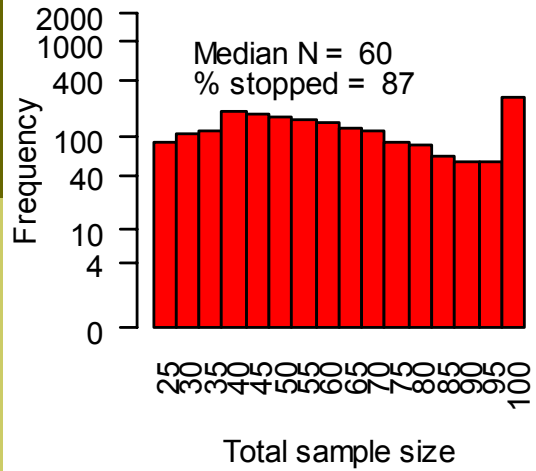
dashed: hypotheses

Exponential, 0.08



Early Stopping

Exponential



Likelihood ratios:

	<1/32	[1/32, 1/10)	[1/10, 1)	[1,10)	[10,32)	>32
Exponential*	0.20	0.76	0.03	0.01	<0.01	<0.01
Weibull 1	0.47	0.53	<0.01	<0.01	<0.01	<0.01
Log-Normal 1	0.27	0.73	<0.01	<0.01	<0.01	<0.01
Cure Rate	<0.01	0.04	<0.01	<0.01	<0.01	0.96
Weibull 2	0.18	0.80	0.01	0.01	<0.01	0.01
Log-Normal 2	0.06	0.55	<0.01	<0.01	0.01	0.37

Frequentist Approach: Exponential Data

- Based on observed data (stopped and completed trials)
 - 0.55% of trials showed significant p-value (versus 0.45% with $LR > 10$)
 - Agreement of 99.6% for hypothesis testing decision
- High agreement in inferences

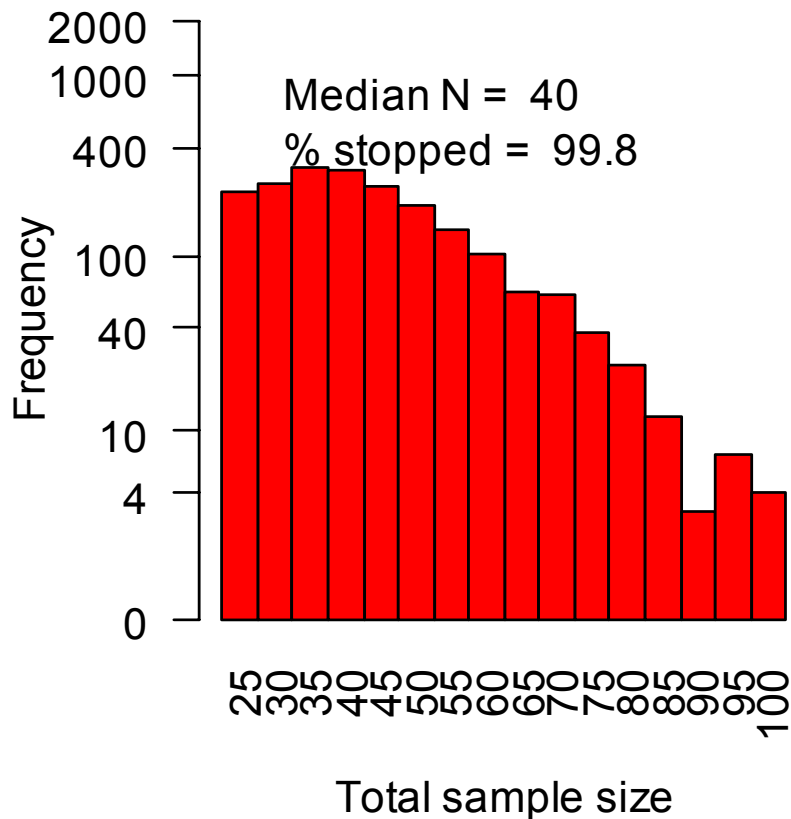
Last set of simulations (for today)

- Early stopping is critical when we have a rate that is even WORSE than the null
- Example:
 - We are testing 62% vs. 74% 6 month PFS
 - What if true 6 month PFS based on our regimen is only 55%? Or 49%?
 - What is the chance of early stopping in these cases?
- Simple scenario: exponential data, exponential model

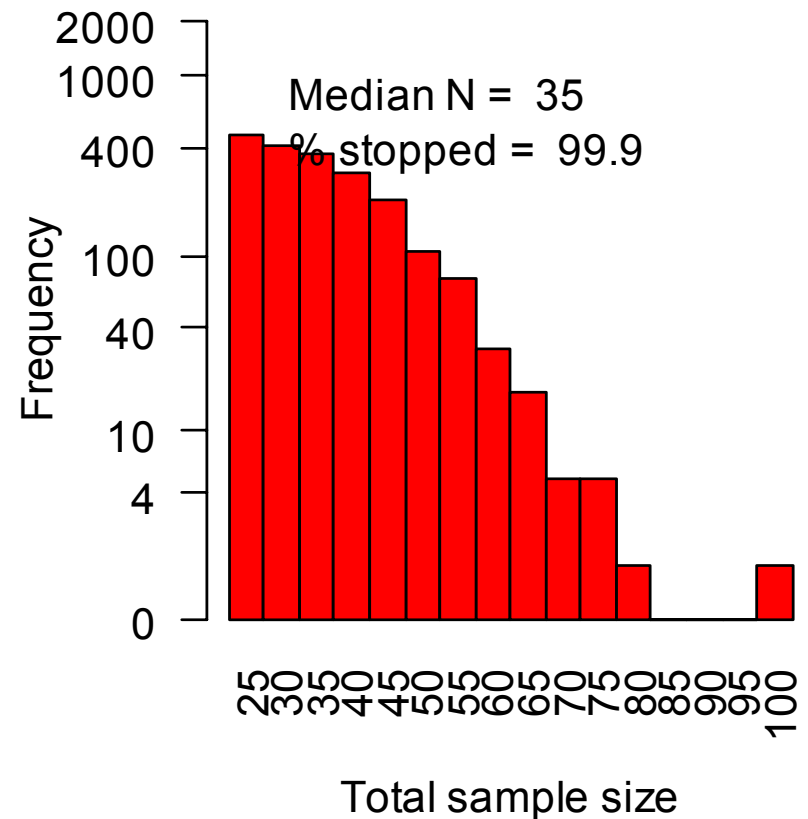
Early Stopping:

H_0 : 6 mo PFS = 62% vs. H_1 : 6 mo PFS = 74%

6 mo PFS = 55%



6 mo PFS = 49%



Likelihood Ratios

	<1/32	[1/32, 1/10)	[1/10, 1)	[1,10)	[10,32)	>32
55% 6 mo PFS	0.19	0.81	<0.01	<0.01	<0.01	<0.01
49% 6 mo PFS	0.26	0.74	<0.01	<0.01	<0.01	<0.01

Conclusions (1)

- **Yay! It works!**
 - **When we have exponential data and $k=10$:**
 - **We stop early OFTEN when we should**
 - **We RARELY stop early when we shouldn't**
- **But, we need to be careful...**
 - **We need a full understanding of the expected and observed survival distribution**
 - **If we have model misspecification, we could run into trouble**
 - **Not unrealistic: breast cancer—cure rate might be best-fitting**
 - **Quantifying simply by one point in time (e.g. 6 month PFS) could be dangerous**
 - **Should elicit several PFS at several times from Investigator**

Conclusions (2)

- **This is the perfect example of why we need to work in close collaboration with oncologists**
 - **Need to get a good appreciation for the anticipated distribution**
 - **Early stopping should be carefully considered based on observed data**
- **Implementation issues**
 - **Probably will not be able to do this in an “off-the-shelf” way**
 - **High-maintenance for the statistician**
 - **Better for patients**
 - **Better for Cancer Center (resources)**

Future work

- Feasibility of 2-parameter models
- Improvement in consistency of $\hat{\lambda}$? $\text{Log}(S(t))$?
- Different censoring mechanisms
- Larger deviations from exponential (how common?)
- Looks: when to start and how often?
- Study design guidelines (e.g. sample size)

References

- **Case and Morgan** (2003) Design of Phase II cancer trials evaluating survival probabilities. *BMC Medical Research Methodology*; v. 3.
- **Birnbaum** (1962) On the Foundations of Statistical Inference (with discussion). *JASA*, 53, 259-326.
- **Blume** (2002) Likelihood Methods for Measuring Statistical Evidence, *Statistics in Medicine*, 21, 2563-2599.
- **Hacking** (1965) *Logic of Statistical Inference*, New York: Cambridge Univ Press.
- **Royall** (1997) *Statistical Evidence: A Likelihood Paradigm*, London, Chapman & Hall.
- **Royall** (2000) On the Probability of Misleading Statistical Evidence, *JASA*, 95; 760-768.
- **Royall and Tsou** (2003) Interpreting statistical evidence by using imperfect models: robust adjusted likelihood functions. *JRSS-B*; 65(2), 391-404.
- **Simon** (1989) Optimal Two-Stage Designs for Phase II Clinical Trials. *Controlled Clinical Trials*; 10,1-10.
- **Smith** (1953) The Detection of Linkage in Human Genetics. *JRSS-B*, 15, 153-192.
- **Thall, Wooten and Tannir** (2005) Monitoring event times in early phase clinical trials: some practical issues. *Clinical Trials*; v. 2, 467-478.