

Endpoints and Stopping Rules

Janet Wittes
Airlie 2005

Outline of talk

- Micro course in statistics
- Thoughts on endpoints
- Interim analysis

Intro to stat –Lesson 1

- 20 horses you know nothing-you put down \$1; what should the payoff be if you win?
- $(19/20)*\$0 + (1/20)*\$20 = \$1$
- You still want to put down \$1 but bet on 2 horses and still have a payoff of \$20.
- **Problem: multiplicity**

Lesson 2

- You look at the horses after $\frac{2}{3}$ of a lap and want to put down \$1 for \$20 payoff.
- **Problem: interim analysis**
- Look at horses after the race and want to choose the winner. Still \$1 for \$20 payoff.
- **Problem: ex post facto analysis**

Why is interim analysis a problem?

1. Sample size
2. Endpoints
3. Duration of trial
4. Multiplicity

1. Sample size

- If we have 90 percent power at end of study, what would our power be earlier?
- $N/4$ 40%
- $N/2$ 60%
- $3N/4$ 80%

2. Endpoints

- Statistical types
 - Mean (or median)
 - Proportion
 - Time to event
- Follow-up time

Example- alcohol dependence

- 84 day study with 3 month extension
- n=315
- Measure drinking each day
- Call a missing day “heavy drinking”
- Primary endpoint:
 - Cumulative days without heavy drinking

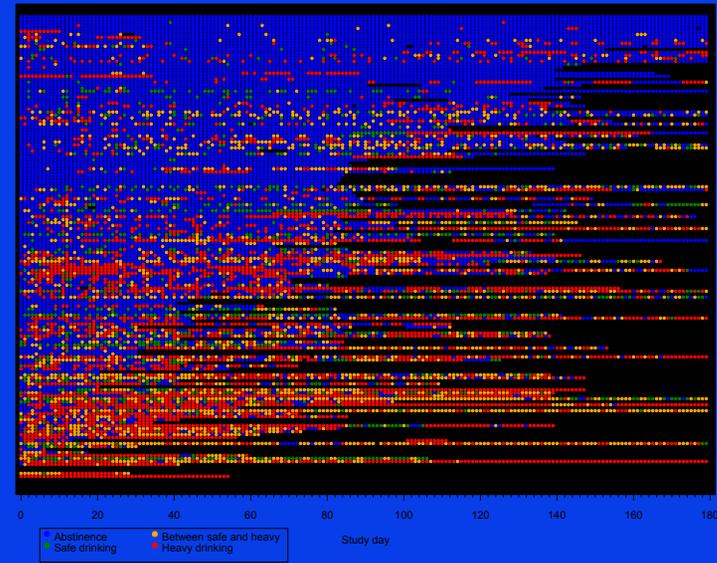
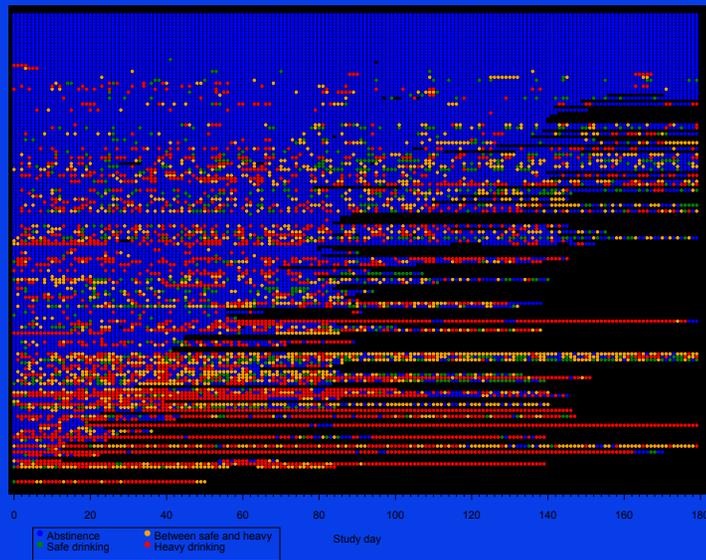
Mean days without heavy drinking

Intervention: 60 ± 2

Control: 59 ± 2

P-value: 0.6

Are these different?



With 3 month extension

- Cumulative days without heavy drinking

	3 mo	6 mo
○ Intervention:	60 ±2	100 ± 4
○ Control:	59 ±2	90 ± 4
○ P-value:	0.6	0.09

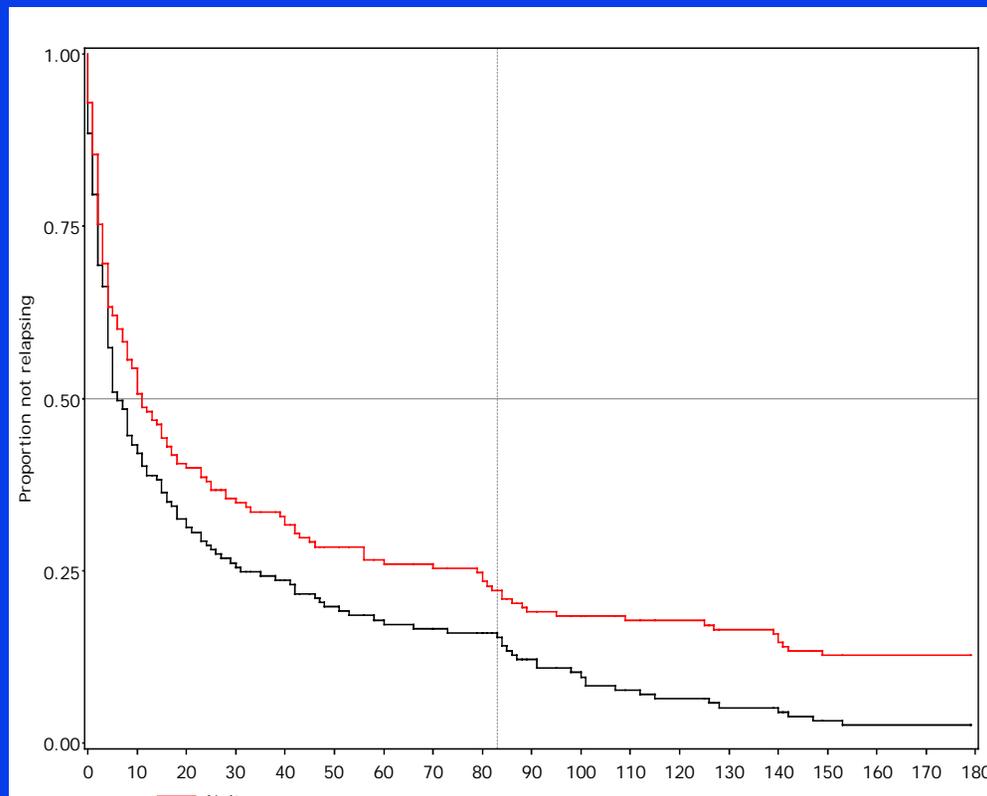
Proportion with no heavy drinking

Treated		Control		P-value (Fisher exact)
37/158	23%	25/157	16%	0.12

Proportion with no heavy drinking

Treated		Control		P-value (Fisher exact)
3 months				
37/158	23%	25/157	16%	0.12
6 months				
20/158	13%	4/157	3%	0.001

Time to heavy drinking



Median days to relapse

	Median (95% CI)	
Treated	11	(8,17)
Control	6	(4, 10)
P-value	3 mo	0.05

Median days to relapse

	Median (95% CI)	
Treated	11	(8,17)
Control	6	(4, 10)
P-value	3 mo	0.05
	6 mo	0.004

Equal p-value is not equal evidence

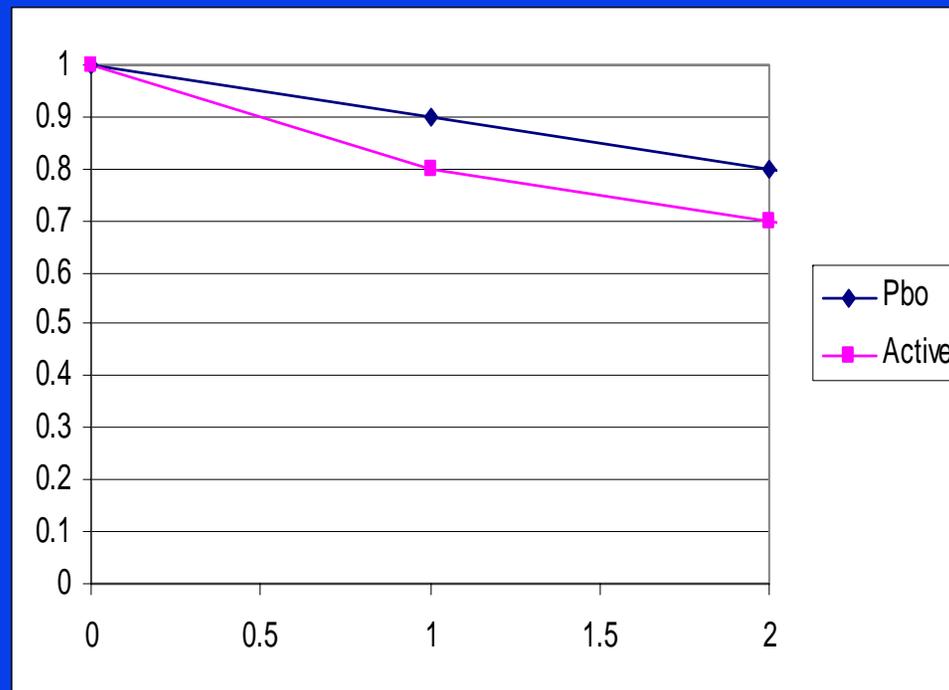
- Think of a p-value as the icing on the cake
- You need the cake!
- Think of your results in terms of the estimated effect first, the significance level second

3. Duration

- We cannot extrapolate beyond what we see

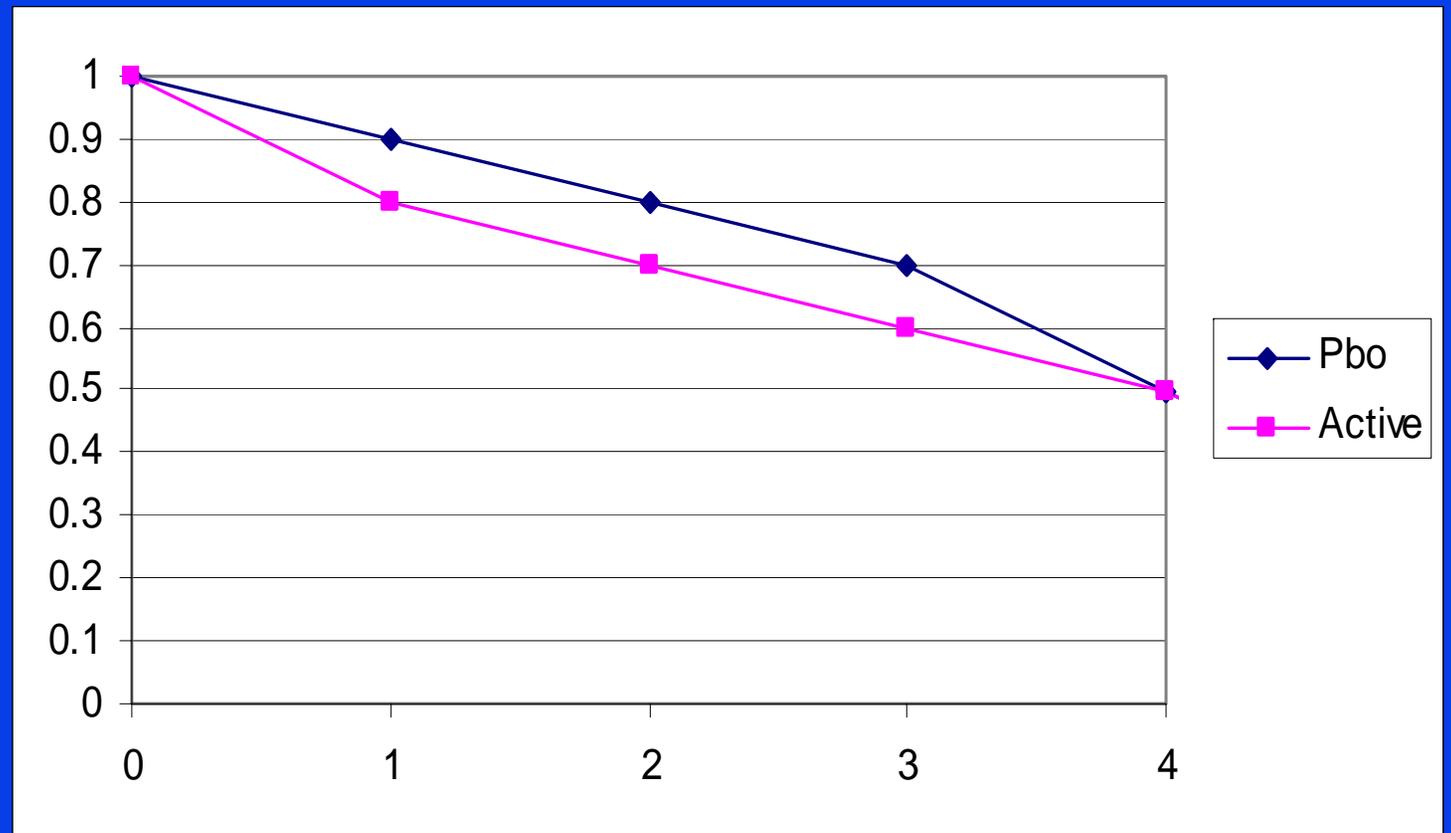


Example: HERS-type

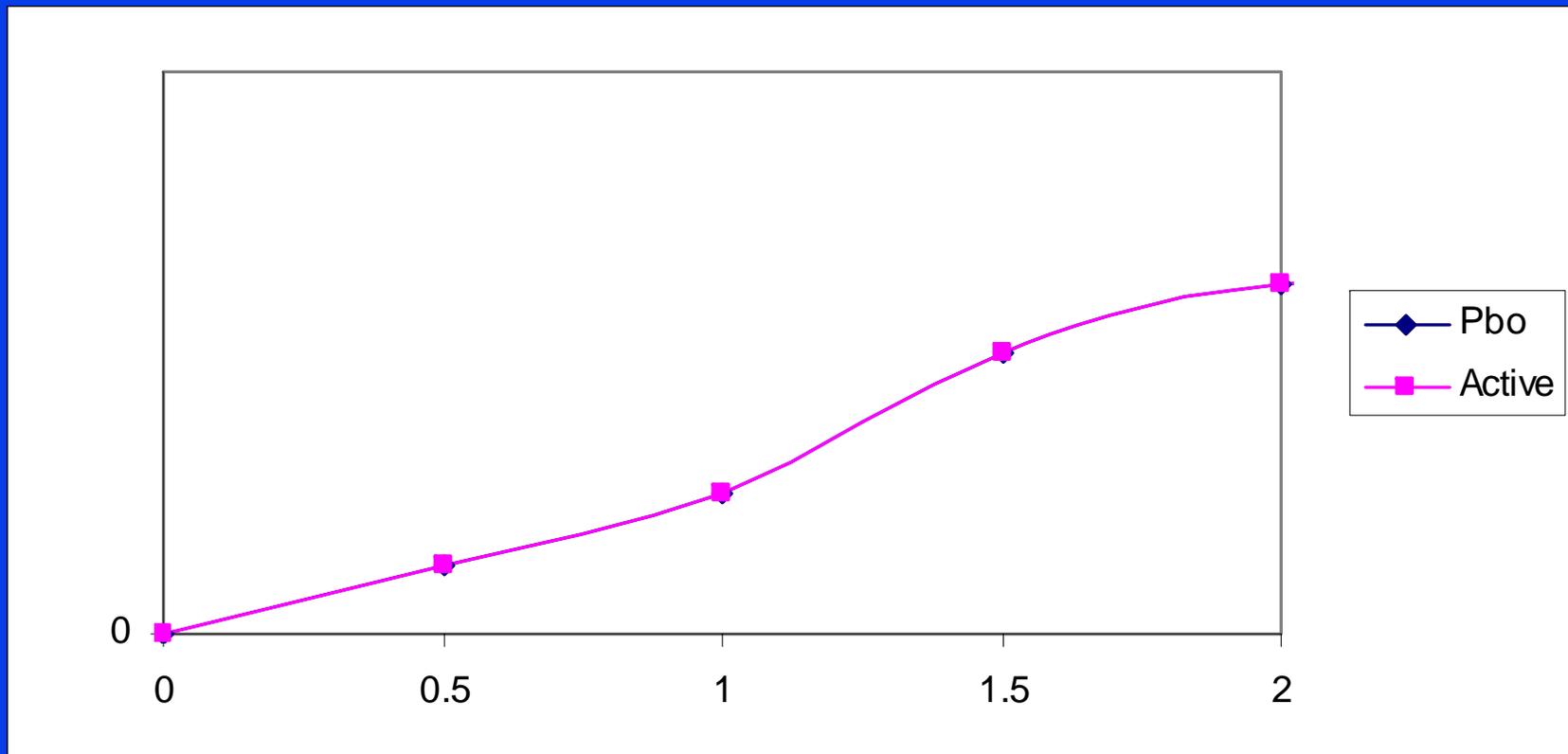


Duration

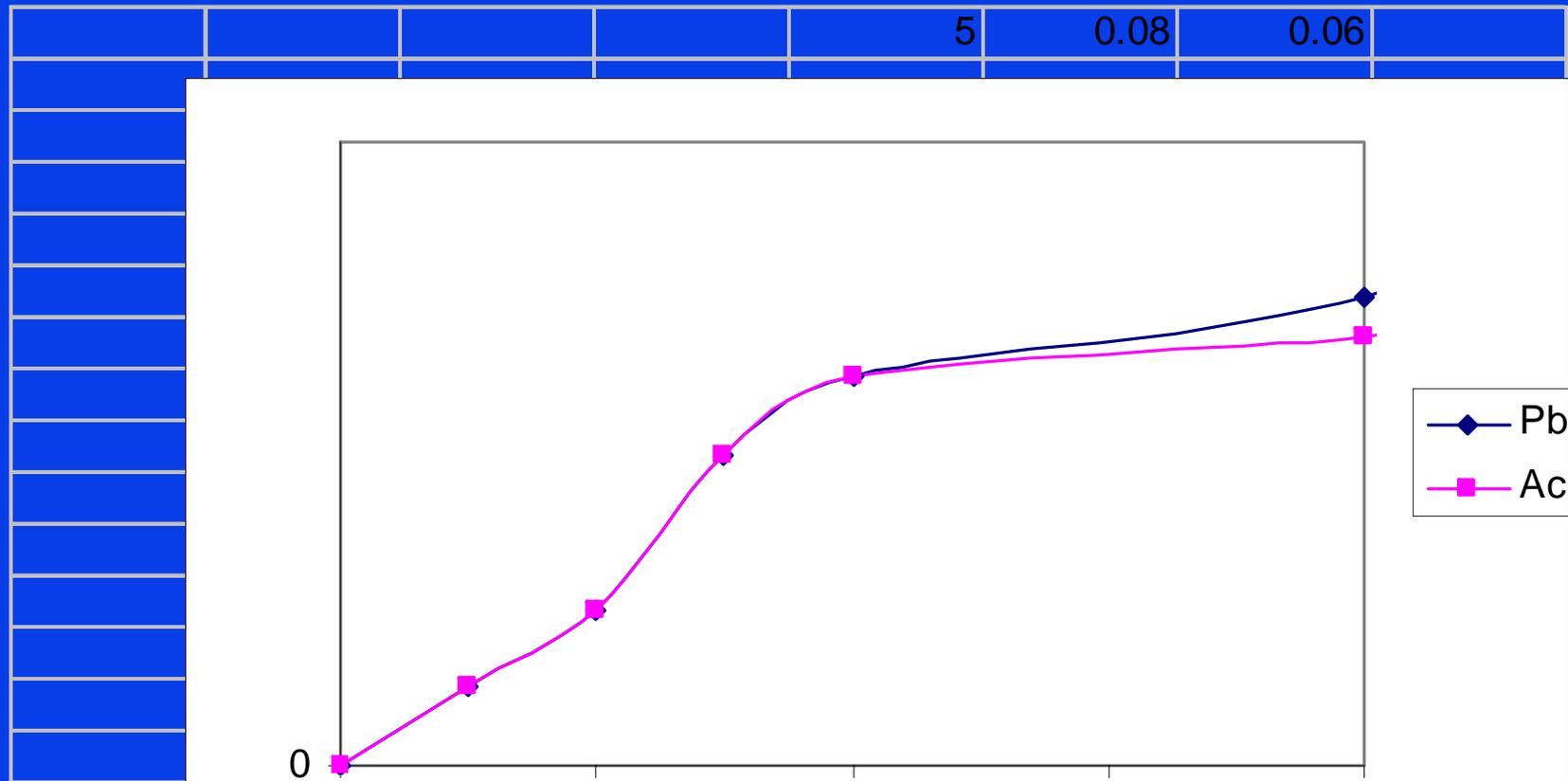
- What happened



VA HDL Intervention Trial



VA HDL Intervention Trial



4. Multiplicity: Why not look over and over?

- In the long run, everyone is dead.-Keynes

# of tests	Overall Type I error rate
1	=0.05
2	$1-(1-.05)^2$ =0.08
10	$1-(1-.05)^{10}$ =0.19
100	$1-(1-.05)^{100}$ =0.37
Infinity	$1-(1-.05)^\infty$ =1

Ensuring trial's integrity

- Threats from interim analyses
 - Investigators/subjects see interim results
 - Not accounting for multiplicity prospectively
 - If trial managers/planners/sponsors know or meet with those who know interim results

Role of statistical stopping rules

- How do they preserve the Type I error rate?
- What is a “spending function”?
- The arguments for/against formal boundaries

Correcting for multiplicity

- Apportion error rate over the analysis
- E.g., declare success

at interim analysis only if $p < 0.01$ (1% error)

or

at final analysis if $p < 0.04$ (4% additional error for total 5% error)

($p < 0.045$ may be used because of statistical dependence between the two looks)

More on multiplicity

- Plan all interim analyses, even “administrative looks,” in advance.
- Allocate error for EACH analysis
 - For administrative looks, a very small error allocation, e.g. 0.01% or $p < 0.0001$, may be appropriate, leaving most error (e.g. 4.99%) for other analyses
 - If a need arises for an additional analysis, consult trialists in advance.

And still more

- Can ONLY be accomplished if planned prospectively.
 - If an interim analysis has occurred without a pre-specified stopping rule (e.g., $p < 0.01$),
 - there is no way to determine the critical p value for any remaining analyses
 - there may be no way to interpret the statistical significance of the results

When are boundaries necessary?

- Safety
 - I believe they are inappropriate here. (Not everyone agrees.)
 - A strong, experienced board that won't get nervous when it sees strong trends early
- Efficacy
 - I believe they are necessary here. (Not everyone agrees.)
 - Each member of the IDMC must understand the boundaries.

Monitoring for efficacy

- Statistical issues “solved”
- Boundaries for monitoring efficacy require unequivocal evidence in order to stop early

How should we use our α ?

- All at the end: critical value is 1.96
- Use up our whole wad at the first look
- Use it very sparingly
- Be profligate but leave a bit for the end

Why not use a lot early?

- We lose power
- Our sample size is low early, so we need to find a huge effect to see a statistically significant difference

Commonly used boundaries

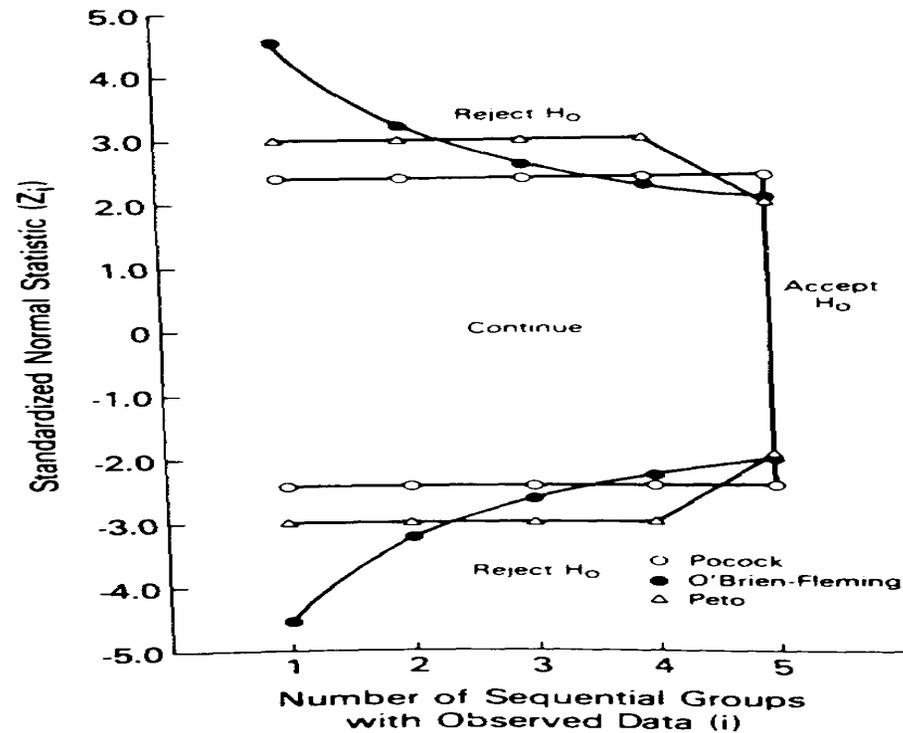


Figure 15-6 Three group sequential stopping boundaries for the standardized normal statistic (Z_i) for up to five sequential groups with two-sided significance level of 0.05.

Rules or guidelines?

- What does it mean to:
 - stop without crossing?
 - fail to stop if we cross?

Conclusion

- Think hard about your endpoint
 - Question
 - Power
 - Time
- Monitor in a way to preserve validity
- Don't be afraid to look at your data
 - But don't cheat