Frequentist Methods in Fixed Sample Tests

Example

Two-sided level .05 test of a normal mean (1 sample)

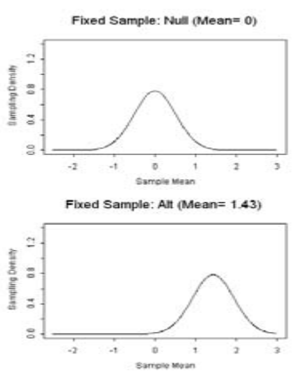
Hypotheses
- Null: Mean = 0
- Alt : Mean = 2

Sample size
- Variance = 26.02
- 100 subjects provide 97.5% power

Critical value (test statistic is the sample mean)
- Reject null if sample mean < -1 or > 1

Example

Sampling density is normal; alternative is simple shift



Statistical Issues

Design operating characteristics based on the sampling density.

Type 1 error (size of test)
- Probability of incorrectly rejecting the null hypothesis

Power (1 - type II error)
- Probability of rejecting the null hypothesis
- Varies with the true value of the measure of treatment effect

Statistical Issues

The type I error associated with a test design is found by integrating the sampling density under the null hypothesis.

Type 1 error (size of test) is the probability of observing a test statistic (estimate of treatment effect) more extreme than the critical value when the null hypothesis is true.

Example

Type I error: Null sampling density tails beyond crit value

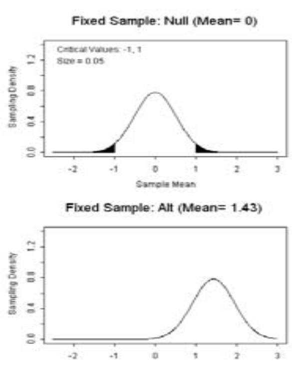
With a sample size of 100, when the mean is 0 and the variance is 26.02

- Probability of observing an estimate (sample mean) greater than 1 is 0.025
- Probability of observing an estimate (sample mean) less than -1 is 0.025

Two-sided type I error (size) is 0.05

Example

Type I error: Null sampling density tails beyond crit value



Statistical Issues

The statistical power associated with a test design is found by integrating the sampling density under particular alternative hypotheses.

Statistical power (1 - type II error) is the probability of observing a test statistic (estimate of treatment effect) more extreme than the critical value when the alternative hypothesis is true.

- Varies with the particular alternative
- In a two-sided test we consider one-sided power
  - lower power and/or
  - upper power

Example

Power: Alternative sampling density tail beyond crit value

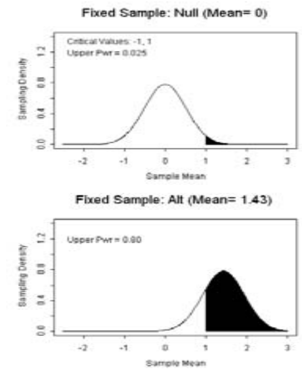
With a sample size of 100, when the variance is 26.02

- Probability of observing an estimate (sample mean) greater than 1 is 0.025 when the mean is 0
- Probability of observing an estimate (sample mean) greater than 1 is 0.800 when the mean is 1.43
- Probability of observing an estimate (sample mean) greater than 1 is 0.975 when the mean is 2

(Power under the null hypothesis is the size of the test.)

Example

Power: Alternative sampling density tail beyond crit value



Statistical Issues

Statistical inference at the end of a trial.

Upon completion of a clinical trial, we are interested in making inference based on an observed test statistic (estimate of treatment effect)

- Point estimate of treatment effect (single best estimate)
- Interval estimate of treatment effect (provides measure of precision of point estimate)
- Quantification of evidence for or against null hypothesis
- Binary decision about truth or falsity of null and alternative hypotheses

Example

Two-sided level .05 test of a normal mean (1 sample)

Suppose we observe a sample mean of 0.4

Questions of interest: Based on observed sample mean of 0.4

- What is the best estimate of treatment effect?
- What is reasonable range of estimates?
- What does this observation tell us about the null hypothesis of a true treatment effect of 0?
- Should we decide that the true treatment effect is not 0?

Statistical Issues

Statistical inference based on the sampling density.

Frequentist inferential measures

- Estimates which
  - minimize bias
  - minimize mean squared error
- Confidence intervals
- P values
- Classical hypothesis testing

Statistical Issues

The P value associated with an observed test statistic is found by integrating the sampling density under the null hypothesis.

P value is the probability (calculated under the null hypothesis) of observing a test statistic (estimate of treatment effect) more extreme than what was actually observed.

(How unusual is the observed data when the null hypothesis is true?)

Example

P value: Null sampling density tail beyond observed value

If the true treatment effect corresponds to a mean of 0

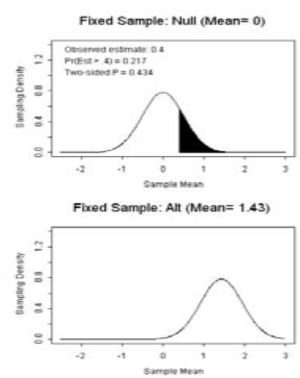
- the probability of observing a sample mean greater than 0.4 is 0.217, and
- the probability of observing a sample mean less than 0.4 is 0.783.

Two-sided P value is twice the smaller of these probabilities

- Two-sided P value: 0.434

Example

P value: Null sampling density tail beyond observed value



Statistical Issues

The confidence interval associated with an observed test statistic is found by integrating the sampling density under all hypotheses.

A particular hypothesized treatment effect is in a $100(1-\alpha)\%$ confidence interval for the observation if, based on the sampling density for that hypothesis, the probability of a test statistic lower (or greater) than the observed value is between $\alpha/2$ and $1-\alpha/2$

(For which hypothesized values of the treatment effect is the observed data not too unusual?)

Example

Conf int: Sampling density tail beyond observed value

We want a 95% CI for the observed sample mean of 0.4.

If the true treatment effect corresponds to a mean of 0, the probability of observing a sample mean greater than 0.4 is 0.217, which is between 0.025 and 0.975

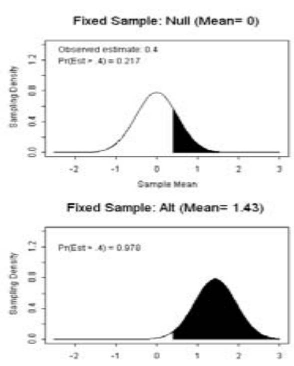
- Hence, 0 is in the 95% confidence interval

If the true treatment effect corresponds to a mean of 1.43, the probability of observing a sample mean greater than 0.4 is 0.978, which is not between 0.025 and 0.975

- Hence, 1.43 is not in the 95% confidence interval

Example

Conf int: Sampling density tail beyond observed value



Statistical Issues

Many point estimates of the true treatment effect are based on the sampling density.

Find the value of the treatment effect for which the observed test statistic is

- the mean of its sampling distribution
- the median of its sampling distribution
- the mode of its sampling distribution

Maximum likelihood estimates correspond to finding the value of the treatment effect for which the sampling <u>density</u> of the observed <u>data</u> is maximized. (Need to consider sufficiency of statistics.)

Statistical Issues

For all estimates, many measures of optimality are based on the sampling distribution.

Unbiasedness

- For the sampling distribution under <u>every</u> hypothesized treatment effect, the expected value of the estimate is the true value

Minimum mean squared error

- For the sampling distribution under <u>every</u> hypothesized treatment effect, the expected value of the squared difference between the estimate and the true value is as small as possible

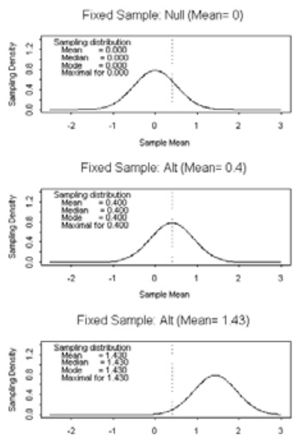
Example

Sampling density is normal; alternative is simple shift

For an observed sample mean of 0.4, this will be the mean, median, and mode of the sampling distribution only if the true treatment effect is 0.4.

Among all sampling distributions (as the true treatment effect varies), the sampling density that is highest at 0.4 is the one that corresponds to a treatment effect of 0.4.

Example



Frequentist Methods in Presence of a Stopping Rule

Statistical Issues

In monitoring a study, ethical considerations may demand that a study be stopped early.

The conditions under which a study might be stopped early constitutes a stopping rule

- At each analysis, the values that would cause a study to stop early are specified

The stopping boundaries might vary across analyses due to the imprecision of estimates

- At earlier analyses, estimates are based on smaller sample sizes and are thus less precise

Statistical Issues

The choice of stopping boundaries is typically governed by a wide variety of often competing goals.

The process for choosing a stopping rule is the substance of this course.

For the present, however, we consider only the basic framework for a stopping rule.

Statistical Issues

The stopping rule must account for ethical issues.

Early stopping might be based on

- Individual ethics
  - the observed statistic suggests efficacy
  - the observed statistic suggests harm
- Group ethics
  - the observed statistic suggests equivalence

Exact choice will vary according to scientific / clinical setting

Example

Two-sided level .05 test of a normal mean (1 sample)

Fixed sample design

- Null: Mean = 0; Alt : Mean = 2
- Maximal sample size: 100 subjects

Early stopping for harm, equivalence, efficacy according to value of sample mean

(Example stopping rule taken from a two-sided symmetric design (Pampallona & Tsiatis, 1994) with a maximum of four analyses and O'Brien-Fleming (1979) boundary relationships)

Example

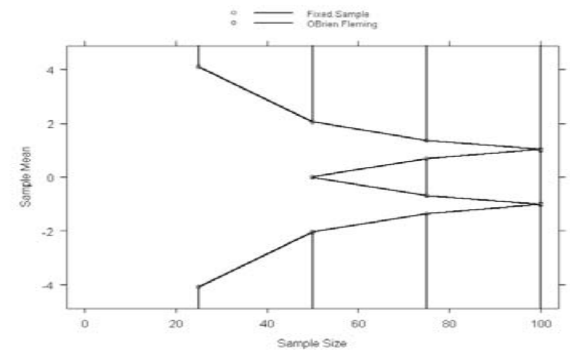
“O’Brien-Fleming” stopping rule

At each analysis, stop early if sample mean is indicated range

N	Harm	Equiv	Efficacy
25	< -4.09	----	> 4.09
50	< -2.05	(-0.006,0.006)	> 2.05
75	< -1.36	(-0.684,0.684)	> 1.36

Example

“O’Brien-Fleming” stopping rule



Statistical Issues

In sequential testing (1 or more interim analyses), more specialized software is necessary.

The sampling density at each stage depends on continuation from previous stage

Recursive numerical integration of convolutions

The sampling density is not so simple: skewed, multimodal, with jump discontinuities

The treatment effect is no longer a shift parameter

Example

“O’Brien-Fleming” stopping rule

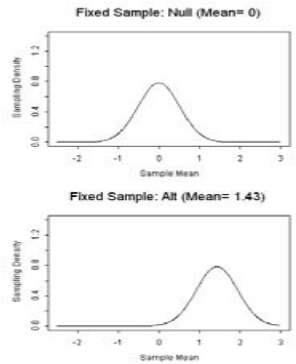
Possibility for early stopping introduces jump discontinuities at values corresponding to stopping boundaries

- Size of jump will depend upon true value of the treatment effect (mean)

N	Harm	Equiv	Efficacy
25	< -4.09	----	> 4.09
50	< -2.05	(-0.006,0.006)	> 2.05
75	< -1.36	(-0.684,0.684)	> 1.36

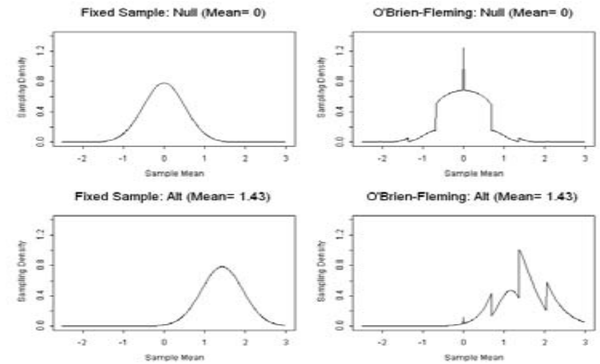
Example

Fixed sample (no interim analyses) sampling density



Example

Sampling density under stopping rule



Statistical Issues

Because the estimate of the treatment effect is no longer normally distributed in the presence of a stopping rule, the frequentist inference typically reported by statistical software is no longer valid

The standardization to a Z statistic does not produce a standard normal

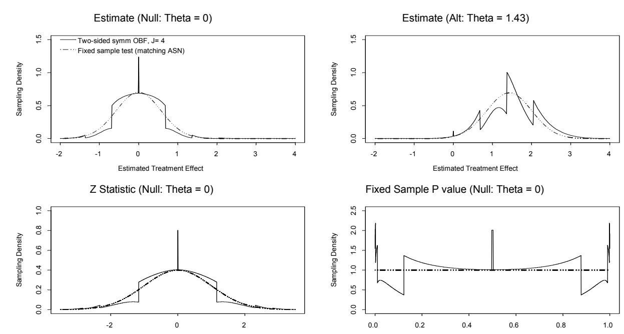
- The number 1.96 is now irrelevant

Converting that Z statistic to a fixed sample P value does not produce a uniform random variable under the null

- We cannot compare that fixed sample P value to 0.025

Sampling Densities for Z, Fixed P

Sampling densities for Z statistic, fixed sample P value in the presence of a stopping rule



Statistical Issues

Because a stopping rule changes the sampling distribution, the use of a stopping rule should change the computation of those design operating characteristics based on the sampling density.

Type 1 error (size of test)

- Probability of incorrectly rejecting the null hypothesis

Power (1 - type II error)

- Probability of rejecting the null hypothesis
- Varies with the true value of the measure of treatment effect

Example

Type I error: Null sampling density tails beyond crit value

Fixed sample test: Mean 0, variance 26.02, N 100

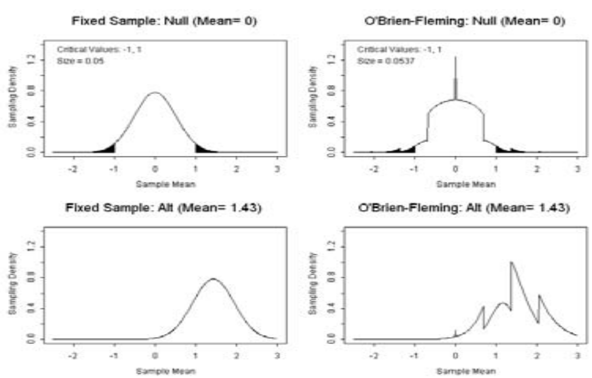
- Prob that sample mean is greater than 1 is 0.025
- Prob that sample mean is less than -1 is 0.025
- Two-sided type I error (size) is 0.05

O'Brien-Fleming stopping rule: Mean 0, variance 26.02, max N 100

- Prob that sample mean is greater than 1 is 0.0268
- Prob that sample mean is less than -1 is 0.0268
- Two-sided type I error (size) is 0.0537

Example

Type I error: Null sampling density tails beyond crit value



Example

Power: Alternative sampling density tail beyond crit value

Fixed sample test: variance 26.02, N 100

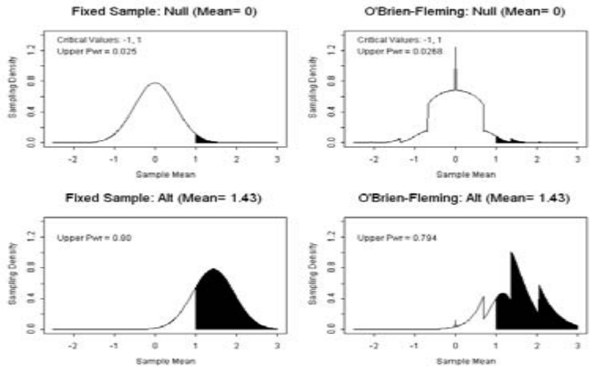
- Mean 0.00: Prob that sample mean > 1 is 0.025
- Mean 1.43: Prob that sample mean > 1 is 0.800
- Mean 2.00: Prob that sample mean > 1 is 0.975

O'Brien-Fleming stopping rule: variance 26.02, max N 100

- Mean 0.00: Prob that sample mean > 1 is 0.027
- Mean 1.43: Prob that sample mean > 1 is 0.794
- Mean 2.00: Prob that sample mean > 1 is 0.970

Example

Power: Alternative sampling density tail beyond crit value



Statistical Issues

Because a stopping rule changes the sampling distribution, the use of a stopping rule should change the computation of those measures of statistical inference based on the sampling density.

- Frequentist inferential measures
  - Estimates which
    - minimize bias
    - minimize mean squared error
  - Confidence intervals
  - P values
  - Classical hypothesis testing

Example

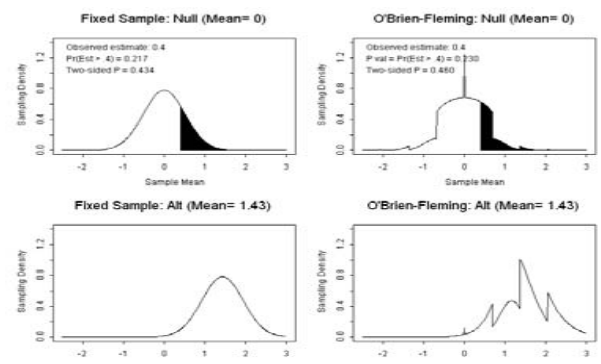
P value: Null sampling density tail beyond observed value

- Fixed sample: Obs 0.4, Mean 0, variance 26.02, N 100
  - Prob that sample mean is greater than 0.4 is 0.217
  - Prob that sample mean is less than 0.4 is 0.783
  - Two-sided P value is 0.434

- O'Brien-Fleming stopping rule: Obs 0.4, Mean 0, variance 26.02, max N 100
  - Prob that sample mean is greater than 0.4 is 0.230
  - Prob that sample mean is less than 0.4 is 0.770
  - Two-sided P value is 0.460

Example

P value: Null sampling density tail beyond observed value



Example

Conf int: Sampling density tail beyond observed value

Fixed sample: 95% CI for Obs 0.4, variance 26.02, N 100

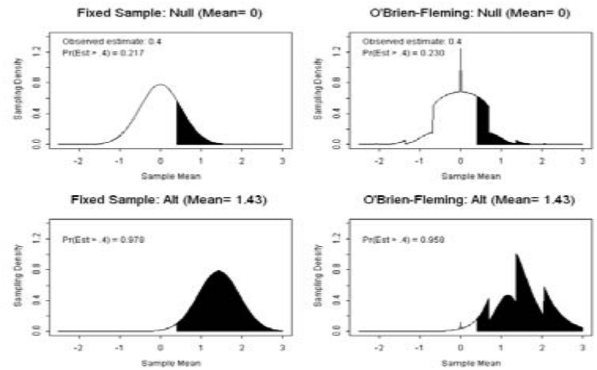
- Mean 0.00: Prob that sample mean > 0.4 is 0.217
- Mean 1.43: Prob that sample mean > 0.4 is 0.978
- 95% CI should include 0, but not 1.43

O'Brien-Fleming stopping rule: 95% CI for Obs 0.4, variance 26.02, max N 100

- Mean 0.00: Prob that sample mean > 0.4 is 0.230
- Mean 1.43: Prob that sample mean > 0.4 is 0.958
- 95% CI should include 0 and 1.43

Example

Conf int: Sampling density tail beyond observed value



Example

Effect of sampling distribution on estimates

For observed sample mean of 0.4, some point estimates are computed based on summary measures of the sampling distribution.

We can examine how the stopping rule affects the summary measures for sampling distribution

- If they differ, then the corresponding point estimates should differ
- (In session 4 we will give precise comparisons for various estimates)

Example

Effect of sampling distribution on estimates

Sampling distribution summary measures for variance 26.02, max N 100

True treatment effect: Mean = 0.000

Sampling Dist Summary Measure	Fixed Sample	O'Brien-Fleming
Mean	0.000	0.000
Median	0.000	0.000
Mode	0.000	0.000
Maximal for	0.000	0.000

Example

Effect of sampling distribution on estimates (cont.)

Sampling distribution summary measures for variance 26.02, max N 100

True treatment effect: Mean = 0.400

Sampling Dist Summary Measure	Fixed Sample	O'Brien-Fleming
Mean	0.400	0.380
Median	0.400	0.374
Mode	0.400	0.000
Maximal for	0.400	0.400

Example

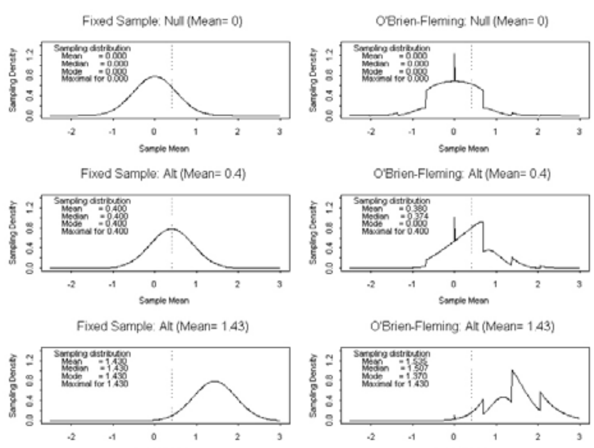
Effect of sampling distribution on estimates (cont.)

Sampling distribution summary measures for variance 26.02, max N 100

True treatment effect: Mean = 1.430

Sampling Dist Summary Measure	Fixed Sample	O'Brien-Fleming
Mean	1.430	1.535
Median	1.430	1.507
Mode	1.430	1.370
Maximal for	1.430	1.430

Example



Statistical Issues

The choice of stopping rule will vary according to the exact scientific and clinical setting for a clinical trial

Each clinical trial poses special problems

Wide variety of stopping rules needed to address the different situations

(One size does not fit all)

Statistical Issues

When using a stopping rule, the sampling density depends on exact stopping rule

This is obvious from what we have already seen.

A fixed sample test is merely a particular stopping rule:

- Gather all N subjects' data and then stop

Statistical Issues

The magnitude of the effect of the stopping rule on trial design operating characteristics and statistical inference can vary substantially

Rule of thumb:

- The more conservative the stopping rule at interim analyses, the less impact on the operating characteristics and statistical inference when compared to fixed sample designs.

Example

"Pocock" stopping rule

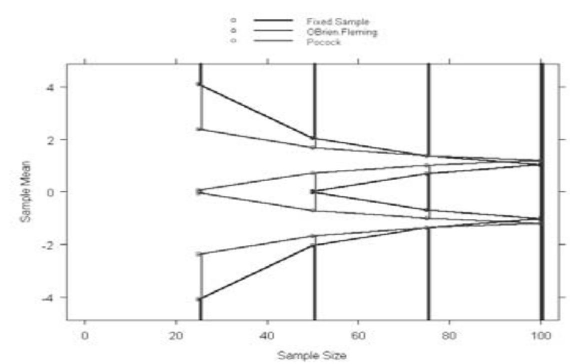
We can consider an alternative stopping rule that is less conservative at the interim analyses

- (This stopping rule is similar to the previous one <u>except</u> it uses Pocock (1977) boundary relationships)

N	Harm	Equiv	Efficacy
25	< -2.37	(-0.048,0.048)	> 2.37
50	< -1.68	(-0.715,0.715)	> 1.68
75	< -1.37	(-1.011,1.011)	> 1.37

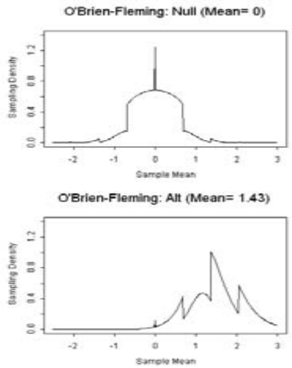
Example

"Pocock" vs "O'Brien-Fleming" stopping rules



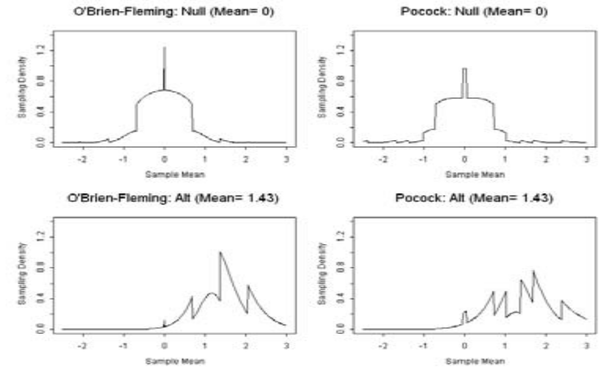
Example

O'Brien-Fleming sampling density



Example

Pocock vs O'Brien-Fleming sampling densities



Example

Type I error: Null sampling density tails beyond crit value

O'Brien-Fleming stopping rule: Mean 0, variance 26.02, max N 100

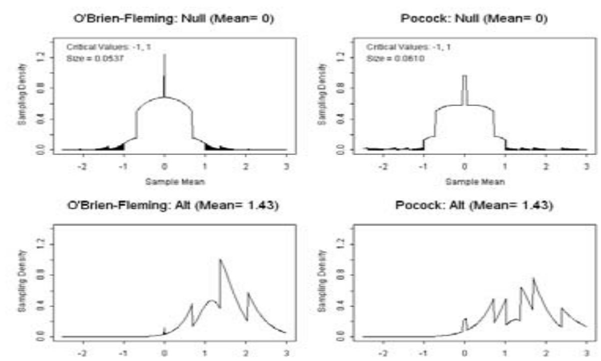
- Prob that sample mean is greater than 1 is 0.0268
- Prob that sample mean is less than -1 is 0.0268
- Two-sided type I error (size) is 0.0537

Pocock stopping rule: Mean 0, variance 26.02, max N 100

- Prob that sample mean is greater than 1 is 0.0305
- Prob that sample mean is less than -1 is 0.0305
- Two-sided type I error (size) is 0.0610

Example

Type I error: Null sampling density tails beyond crit value



Example

Power: Alternative sampling density tail beyond crit value

O'Brien-Fleming stopping rule: variance 26.02, max N 100

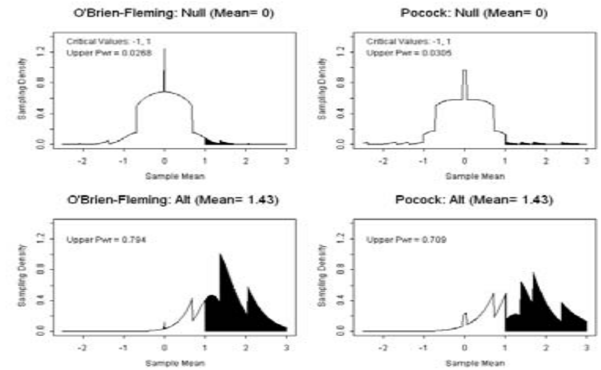
- Mean 0.00: Prob that sample mean > 1 is 0.027
- Mean 1.43: Prob that sample mean > 1 is 0.794
- Mean 2.00: Prob that sample mean > 1 is 0.972

Pocock stopping rule: variance 26.02, max N 100

- Mean 0.00: Prob that sample mean > 1 is 0.031
- Mean 1.43: Prob that sample mean > 1 is 0.709
- Mean 2.00: Prob that sample mean > 1 is 0.932

Example

Power: Alternative sampling density tail beyond crit value



Example

P value: Null sampling density tail beyond observed value

O'Brien-Fleming stopping rule: Obs 0.4, Mean 0, variance 26.02, max N 100

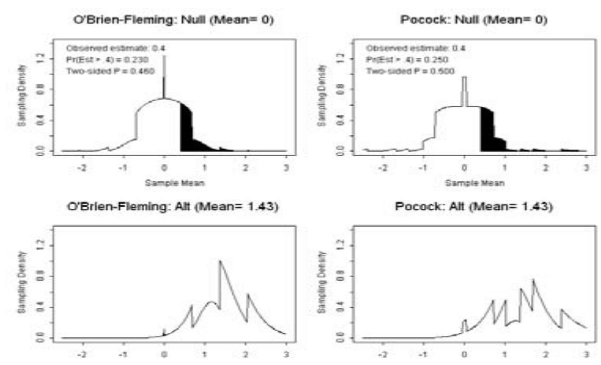
- Prob that sample mean is greater than 0.4 is 0.230
- Prob that sample mean is less than 0.4 is 0.770
- Two-sided P value is 0.460

Pocock stopping rule: Obs 0.4, Mean 0, variance 26.02, max N 100

- Prob that sample mean is greater than 0.4 is 0.250
- Prob that sample mean is less than 0.4 is 0.750
- Two-sided P value is 0.500

Example

P value: Null sampling density tail beyond observed value



Example

Conf int: Sampling density tail beyond observed value

O'Brien-Fleming stopping rule: 95% CI for Obs 0.4, variance 26.02, max N 100

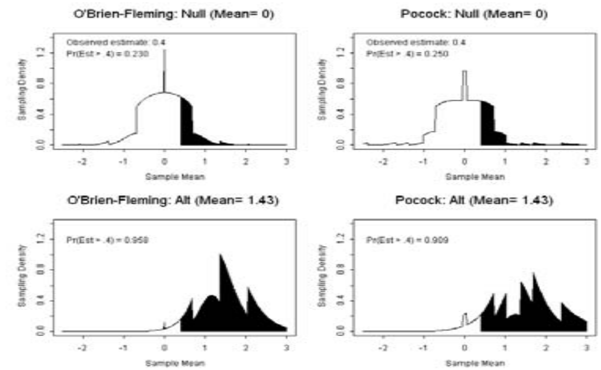
- Mean 0.00: Prob that sample mean > 0.4 is 0.230
- Mean 1.43: Prob that sample mean > 0.4 is 0.958
- 95% CI should include 0 and 1.43

Pocock stopping rule: 95% CI for Obs 0.4, variance 26.02, max N 100

- Mean 0.00: Prob that sample mean > 0.4 is 0.250
- Mean 1.43: Prob that sample mean > 0.4 is 0.909
- 95% CI should include 0 and 1.43

Example

Conf int: Sampling density tail beyond observed value



Example

Effect of sampling distribution on estimates

Sampling distribution summary measures for variance 26.02, max N 100

True treatment effect: Mean = 0.000

Sampling Dist Summary Measure	O'Brien-Fleming	Pocock
Mean	0.000	0.000
Median	0.000	0.000
Mode	0.000	0.000
Maximal for	0.000	0.000

Example

Effect of sampling distribution on estimates (cont.)

Sampling distribution summary measures for variance 26.02, max N 100

True treatment effect: Mean = 0.400

Sampling Dist Summary Measure	O'Brien-Fleming	Pocock
Mean	0.380	0.372
Median	0.374	0.333
Mode	0.000	0.040
Maximal for	0.400	0.400

Example

Effect of sampling distribution on estimates (cont.)

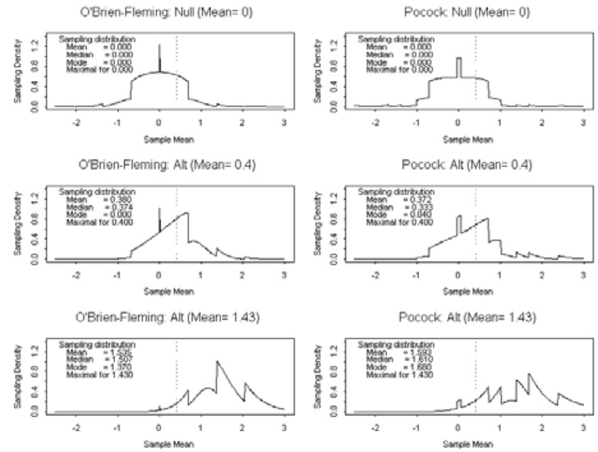
Sampling distribution summary measures for variance 26.02, max N 100

True treatment effect: Mean = 1.430

Sampling Dist Summary Measure	O'Brien-Fleming	Pocock
Mean	1.535	1.593
Median	1.507	1.610
Mode	1.370	1.680
Maximal for	1.430	1.430

February, 2003

Example



February, 2003

Stopping Rules as Group Sequential Tests

February, 2003

Statistical Issues

We can of course maintain the type I error when using a stopping rule by altering the critical value used to declare statistical significance

This only involves finding the correct quantiles of the true sampling density to use at the final analysis

February, 2003

Example

“O’Brien-Fleming” stopping rule

At each interim analysis, stop early if sample mean is indicated range

At the final analysis, the stopping must occur

N	Harm	Equiv	Efficacy
25	< -4.09	----	> 4.09
50	< -2.05	(-0.006,0.006)	> 2.05
75	< -1.36	(-0.684,0.684)	> 1.36
100	< -1.023	(-1.023,1.023)	> 1.023

Example

“Pocock” stopping rule

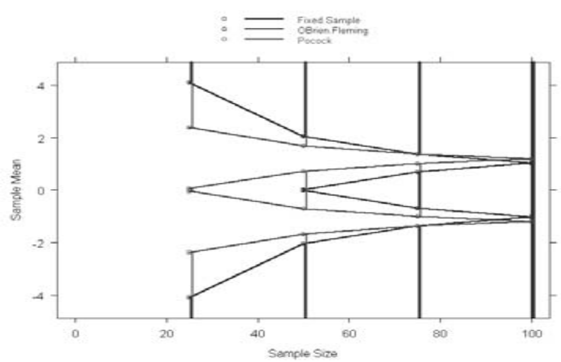
At each interim analysis, stop early if sample mean is indicated range

At the final analysis, the stopping must occur

N	Harm	Equiv	Efficacy
25	< -2.37	(-0.048,0.048)	> 2.37
50	< -1.68	(-0.715,0.715)	> 1.68
75	< -1.37	(-1.011,1.011)	> 1.37
100	< -1.187	(-1.187,1.187)	> 1.187

Example

“Pocock” vs “O’Brien-Fleming” stopping rules



Example

Power: Alternative sampling density tail beyond crit value

O’Brien-Fleming stopping rule: variance 26.02, max N 100

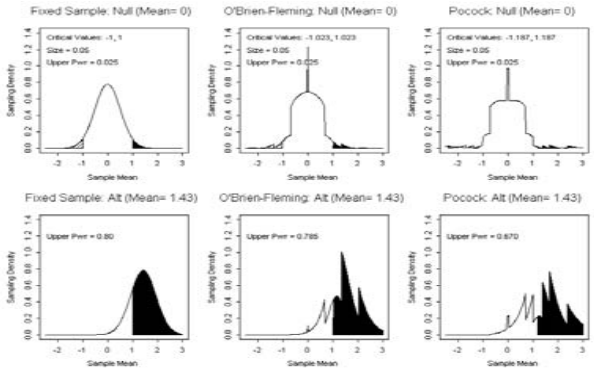
- Mean 0.00: Prob that sample mean > 1.023 is 0.025
- Mean 1.43: Prob that sample mean > 1.023 is 0.785
- Mean 2.00: Prob that sample mean > 1.023 is 0.970

Pocock stopping rule: variance 26.02, max N 100

- Mean 0.00: Prob that sample mean > 1.187 is 0.025
- Mean 1.43: Prob that sample mean > 1.187 is 0.670
- Mean 2.00: Prob that sample mean > 1.187 is 0.922

Example

Power: Alternative sampling density tail beyond crit value



Statistical Issues

The use of a stopping rule allows greater efficiency on average

Sample size requirements are a random variable

- Efficiency characterized by some summary of the sample size distribution
  - Average sample N (ASN)
  - Median, 75%ile of sample size distribution
  - Stopping probabilities at each analysis

Sample size distribution depends on true treatment effect

- (This was the goal of using a stopping rule)

Example

Sample size distribution for designs considered here

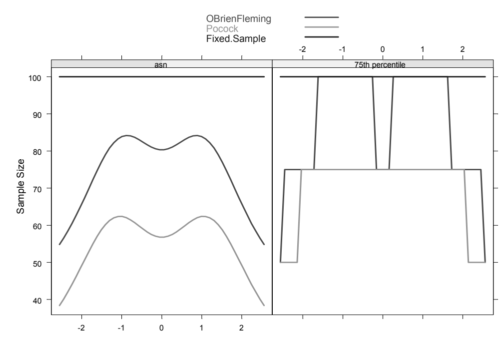
Fixed sample design requires 100 subjects no matter how effective (or harmful) the treatment is

O'Brien-Fleming stopping rule requires fewer subjects on average (worst case: about 84)

Pocock stopping rule requires even fewer subjects on average over a wide range of alternatives (worst case: about 62)

Example

Sample size distribution as a function of treatment effect



Example

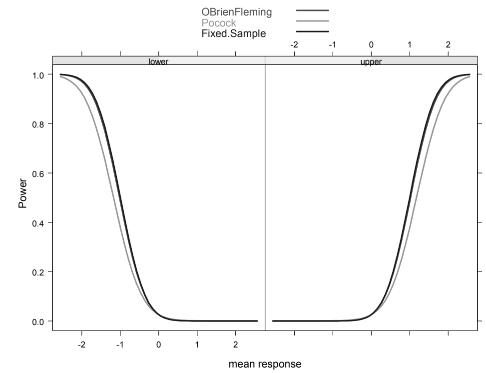
Failure to adjust the maximal sample size does affect the power of the clinical trial design

The introduction of the stopping rule will decrease the power of the design relative to a fixed sample design with the same maximal sample size

In the examples considered so far, we maintained the maximal sample size at 100 subjects

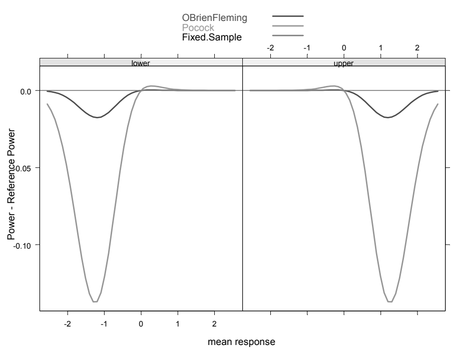
Example

Power as a function of treatment effect



Example

Power as a function of treatment effect relative to fixed sample design



Statistical Issues

We can maintain both the type I error and power when using a stopping rule by altering the critical value used to declare statistical significance and maximal sample size

This involves a search for the sample size that will provide the power.

Example

"O'Brien-Fleming" stopping rule with desired power

At each interim analysis, stop early if sample mean is indicated range

At the final analysis, the stopping must occur

N	Harm	Equiv	Efficacy
26	< -4.01	----	> 4.09
52	< -2.01	(-0.006,0.006)	> 2.01
78	< -1.34	(-0.670,0.670)	> 1.34
104	< -1.003	(-1.003,1.003)	> 1.023

Example

"Pocock" stopping rule with desired power

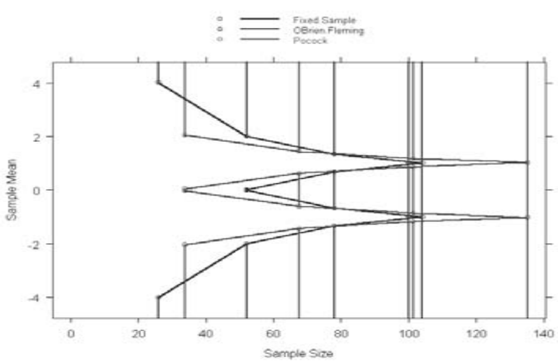
At each interim analysis, stop early if sample mean is indicated range

At the final analysis, the stopping must occur

N	Harm	Equiv	Efficacy
34	< -2.04	(-0.042,0.042)	> 2.04
68	< -1.44	(-0.615,0.615)	> 1.44
101	< -1.18	(-0.869,0.869)	> 1.18
135	< -1.021	(-1.021,1.021)	> 1.021

Example

"Pocock", "O'Brien-Fleming" with desired power



Example

Power: Alternative sampling density tail beyond crit value

O'Brien-Fleming stopping rule: variance 26.02, max N 104

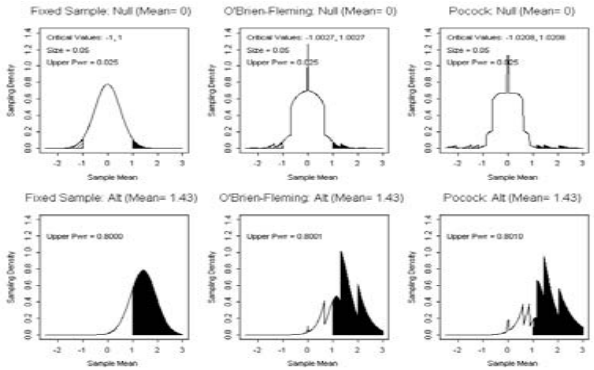
- Mean 0.00: Prob that sample mean > 1.003 is 0.025
- Mean 1.43: Prob that sample mean > 1.003 is 0.8001
- Mean 2.00: Prob that sample mean > 1.003 is 0.975

Pocock stopping rule: variance 26.02, max N 135

- Mean 0.00: Prob that sample mean > 1.021 is 0.025
- Mean 1.43: Prob that sample mean > 1.021 is 0.801
- Mean 2.00: Prob that sample mean > 1.021 is 0.975

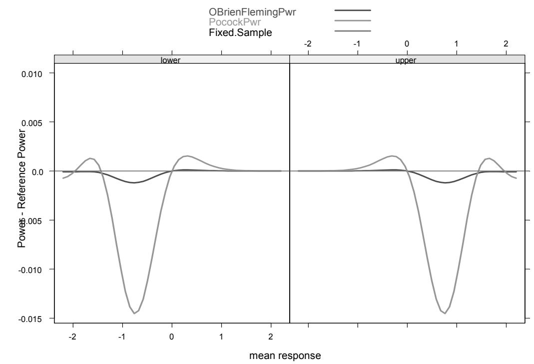
Example

Power: Alternative sampling density tail beyond crit value



Example

Power curves relative to fixed sample design



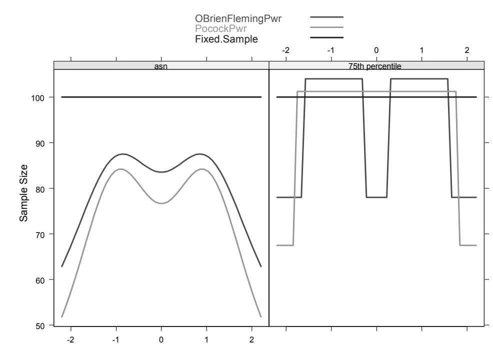
Example

The increased <u>maximal</u> sample size need not mean a less efficient design when using a stopping rule

- Fixed sample design requires 100 subjects no matter how effective (or harmful) the treatment is
- O'Brien-Fleming stopping rule requires fewer subjects on average (worst case: about 88) and the increase in the maximal sample size is only 4%
- Pocock stopping rule requires even fewer subjects on average over a wide range of alternatives, but requires a 35% increase in the maximal sample size
  - However, there is always less than a 25% chance that a trial would continue to the last analysis

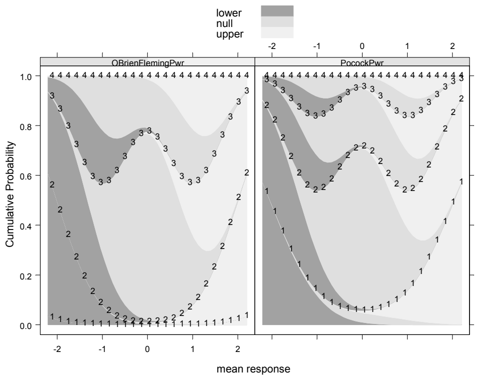
Example

Sample size distribution as a function of treatment effect



Example

Stopping probabilities as a function of treatment effect



Software

Finding an appropriate stopping rule requires access to appropriate software

Numerical integration of the sampling density

- (Simulation can be used in nonstandard settings)