

Interim Analysis and Data Monitoring

Clinical trials are often longitudinal in nature. It is often impossible to enroll all subjects at the same time, so it can take a long time to complete a longitudinal study. Over the course of the trial one needs to consider administrative monitoring, safety monitoring and efficacy monitoring.

Efficacy monitoring can be performed by taking *interim looks* at the primary endpoint data (prior to all subjects being enrolled or all subjects completing treatment). This is because:

- It potentially stops the trial early if there is convincing evidence of benefit or harm of the new product
- It potentially stops the trial for futility, in other words the chance of significant beneficial effect by the end of the study is small given observed data
- Re-estimate final sample size required to yield adequate power to obtain a significant result

Interim analysis evaluates for early efficacy, early futility, safety concerns, or adaptive design with respect to sample size or power.

Group Sequential Design

A common type of study design for interim analysis is GSD, in which data are analyzed at regular intervals.

- Determine a priori the number of interim "looks"
- Let $K = \#$ of total planned analyses including final ($K \geq 2$)
- For simplicity, assume 2 groups and that subjects are randomized in a 1:1 manner
- After every $n = N/K$ subjects are enrolled and followed for a specific time period, perform an interim analysis on all subjects followed cumulatively
- If there is a significant treatment difference at any point, consider stopping the study

Due to multiple testing the probability of observing at least one significant interim result is much greater than the overall $\alpha = .05$, as a result the interim analyses should NOT be performed using the family-wise error rate. The data at each interim analysis contains data from the previous interim and thus are not independent.

Equivalently we would have K critical values for each interim:

- First interim analysis compare test statistic to critical value $Z_1 |> c_1$. If significant then stop the trial.
- Second interim analysis compare test statistic to critical value $Z_2 |> c_2$. If significant stop the trial
- ...
- Final analysis compare test statistic to critical value $X_k |> c_k$.

Pocock Approach (1977)

Derives constant critical values across all stages to maintain the overall significance level at .05. The critical value depends on the number of interim analyses, but is the same for each interim look.

Ex. When $K=5$, Z critical value = 2.413 for each interim and the final analysis. When $K=4$, Z critical value = 2.361.

O'Brien-Fleming Approach (1979)

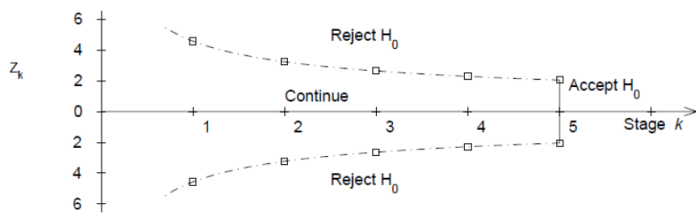
Proposed sequential testing procedure that has critical values (in absolute value) decrease over the stages. Z critical value depends on total number of interim analyses and the stage of the interim analysis. The critical Z value depends on the total number of interim analyses and the stage of the interim analysis.

Table 2.3 *O'Brien & Fleming tests: constants $C_B(K, \alpha)$ for two-sided tests with K groups of observations and Type I error probability α*

K	$C_B(K, \alpha)$		
	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$
1	2.576	1.960	1.645
2	2.580	1.977	1.678
3	2.595	2.004	1.710
4	2.609	2.024	1.733
5	2.621	2.040	1.751
6	2.631	2.053	1.765
7	2.640	2.063	1.776
8	2.648	2.072	1.786
9	2.654	2.080	1.794
10	2.660	2.087	1.801
11	2.665	2.092	1.807
12	2.670	2.098	1.813
15	2.681	2.110	1.826
20	2.695	2.126	1.842

Ex. For $K = 5$ after after 200 subjects completed:

- $C_5 = 2.040$ a constant determined by O'Brien-Fleming)
- First look critical value is: $\sqrt{5/1} * 2.04 = 4.56$
- Second look critical value: $\sqrt{5/2} * 2.04 = 3.23$
- ...
- Critical value final look: $\sqrt{5/5} * 2.04 = 2.04$



This makes it more difficult to declare superiority at "earlier" looks, but does lose much of the original alpha at the final look. This is more conservative than Pocock and the recommended approach by the FDA on their 2010 Guidance on Adaptive Designs.

Controlling the Overall Significance Level

Issues with group sequential procedures:

- Need to specify a priori for the number of planned interim analyses
- The timing of interim analyses is generally not exact calendar time but "information time" (i.e. based on sample size or number of events)
 - Assumes interim analyses performed after every n subjects complete follow-up (or y events occur)
 - Difficult to schedule formal review procedures for interim analyses after every n subjects (or y events); may want to allow more flexibility in scheduling
- So the main issue is: How do we allow for more flexibility for unscheduled interim analysis?

Alpha-Spending

From Lan and DeMets (1983) Biometrika: Adjust the levels via an "alpha-spending function". Think of it like each analysis spends a bit of the alpha power.

- $\alpha(s)$ denotes alpha spending function.
 - s = proportion of information (sample size or events) accrued
 - $s = 0$ at the start of the study (0% information); $\alpha(0) = 0$
 - $s = 1$ is the end of the study (100% information); $\alpha(1) = 1$
 - $\alpha(s_k)$ proportion of Type 1 error one is willing to spend up to time k
 - Not a significance level

This can work in conjunction with O'Brien-Fleming.

Interpretation:

- If the first interim analysis occurs after 20% of information, reject treatment equality if p-value $< .000001$
- If the third interim analysis occurs after 60% of the subjects are in, reject treatment equality if p-value $< .0074$

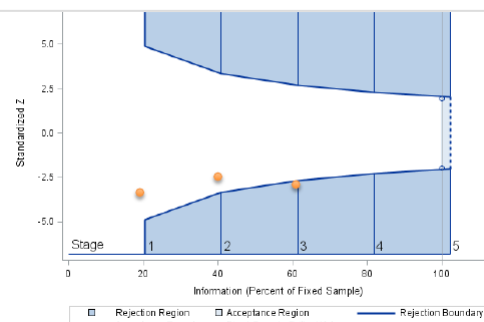
Focus on the significance levels, note the the alpha spending and significance level are two distinct values and not equal. Computing critical values and corresponding significant levels require knowledge of multivariate distributions. There is no "simple" equation to get from $\alpha(s)$ to Z. Typically obtained via numerical integration.

```
/*
plots=boundary - graph observed standardized test statistics
at each interim
errspend - give overall error spending
bscale=pvalue - instead of critical values print p-values
info=equal - equal intervals
stop=reject - trying to reject the null (default)
*/

proc seqdesign errspend plots=boundary;
TwoSidedObrienFleming: design nstages=5 alpha=0.05
alt=twosided info=equal method=errfuncobf
stop=reject;
run;
```

Null Reference = 0					
Stage	Information Level	Alternative Reference		Boundary Values	
	Proportion	Lower	Upper	Lower Alpha	Upper Alpha
1	0.2000	-1.46628	1.46628	-4.87688	4.87688
2	0.4000	-2.07364	2.07364	-3.35706	3.35706
3	0.6000	-2.53968	2.53968	-2.68028	2.68028
4	0.8000	-2.93256	2.93256	-2.28982	2.28982
5	1.0000	-3.27871	3.27871	-2.03103	2.03103

Critical values



Notes that the bscale=pvalue option gives one tail of the distribution, so to get the significance level we need to multiply the upper or lower boundary by 2.

Error Spending Information					
Stage	Information Level	Cumulative Error Spending			
	Proportion	Alpha	Beta	Beta	Alpha
1	0.2000	0.00000	0.00000	0.00000	0.00000
2	0.4000	0.00039	0.00000	0.00000	0.00039
3	0.6000	0.00381	0.00000	0.00000	0.00381
4	0.8000	0.01221	0.00000	0.00000	0.01221
5	1.0000	0.02500	0.10000	0.10000	0.02500

Overall error spending:
Take upper or lower
boundary and multiply by 2

Pocock Approach in SAS

```

?* Pocock alpha-spending for a two-sided test with
[Two-sided alpha spent of .05 by final analysis */
proc seqdesign errspend bscale=pvalue;
TwoSidedPocock: design nstages=5 alpha=0.05
alt=twosided info=equal method=poc stop=reject;
run;

```

Boundary Information (p-Value Scale) Null Reference = 0					
Stage	Information Level Proportion	Alternative Reference		Boundary Values	
		Lower	Upper	Alpha	Alpha
1	0.2000	-1.59237	1.59237	0.00791	0.99209
2	0.4000	-2.25195	2.25195	0.00791	0.99209
3	0.6000	-2.75807	2.75807	0.00791	0.99209
4	0.8000	-3.18474	3.18474	0.00791	0.99209
5	1.0000	-3.56065	3.56065	0.00791	0.99209

Error Spending Information					
Stage	Information Level Proportion	Cumulative Error Spending			
		Lower		Upper	
		Alpha	Beta	Beta	Alpha
1	0.2000	0.00791	0.00003	0.00003	0.00791
2	0.4000	0.01376	0.00003	0.00003	0.01376
3	0.6000	0.01827	0.00003	0.00003	0.01827
4	0.8000	0.02193	0.00003	0.00003	0.02193
5	1.0000	0.02500	0.10000	0.10000	0.02500

Interim Analyses For Safety

It is not necessarily easy from an administrative and study conduct perspective. We need to determine if data is still recent enough to be included. Weeks or months may pass between the last subject visit and generation of interim results due to data entry and cleaning. Depending on the size of the study, the ideal goal is to have a < 60-day lag between data collected at sites and the interim analysis report. Otherwise, interim analysis be be obsolete by the time analysis is completed.

Inspection of adverse events and serious adverse events is primary concern. Labs, vital signs, etc. need to be inspected. Unlike efficacy there is often no formal stopping rules based on p-values or a parametric test. If it is felt there is a safety concern, the study may be stopped regardless of significance between treatments.

The results of the interim analysis are also inspected by a Data Safety and Monitoring Board (DSMB), sometimes called the Data Monitoring Committee. Usually these consist of:

- ≥ 1 Statistician with expertise in interim analyses
- 2-4 clinicians with experience in the topic
- Maybe an ethicist (especially in government-sponsored trials)
- No member can be a study investigator
- DSMB is independent of the sponsor and all study activities
- DSMB can recommend to sponsor early stoppage when there is evidence of clear risk, harm or futility. But they cannot stop the trial themselves

The sponsor will often hire an outside group (Contract Research Organization or CRO) to perform the interim analyses. The statistician at CRO has a randomization schedule, and the analysis group

cannot divulge ANY information to the sponsor or to any personnel involved in the study; It is presented to the independent Data Safety Monitoring Board.

Revision #4

Created 31 March 2023 14:12:02 by Elkip

Updated 31 March 2023 16:07:53 by Elkip