Multiple Testing

Zitong (Cathy) Qi
Mentor: Anna Neufeld

UNIVERSITY *of* WASHINGTON

Outline

- Brief Review of Hypothesis Testing
- Motivation for Interim Analysis & Multiple Testing Techniques
- Alpha-spending function and FWER
- Simulation Study: Comparing Methods
- Extensions

Motivating Example:

In clinical trial, with a treatment and a control group

Null hypothesis:
Mean of blood pressure (treatment)
= Mean of blood pressure(control)
$\mu_T = \mu_C$

Alternative hypothesis:
Mean blood pressure (treatment)
≠ Mean blood pressure(control) (treatment effect)
$\mu_T \neq \mu_C$

BRIEF review of NHST - null hypothesis significance testing for single test

P-value : Assuming the null hypothesis is true, how extreme is our observed statistic
(is our result simply due to random fluctuations)

Alpha: We choose a cutoff called alpha. If p-value is less than alpha, we reject the null and we call the result statistically significant

Type I error: when we conclude that the treatment and the control groups are different, even though in reality they are the same (wrongly reject the null hypothesis)

ALPHA: Probability of making a Type I error when conducting 1 SINGLE TEST

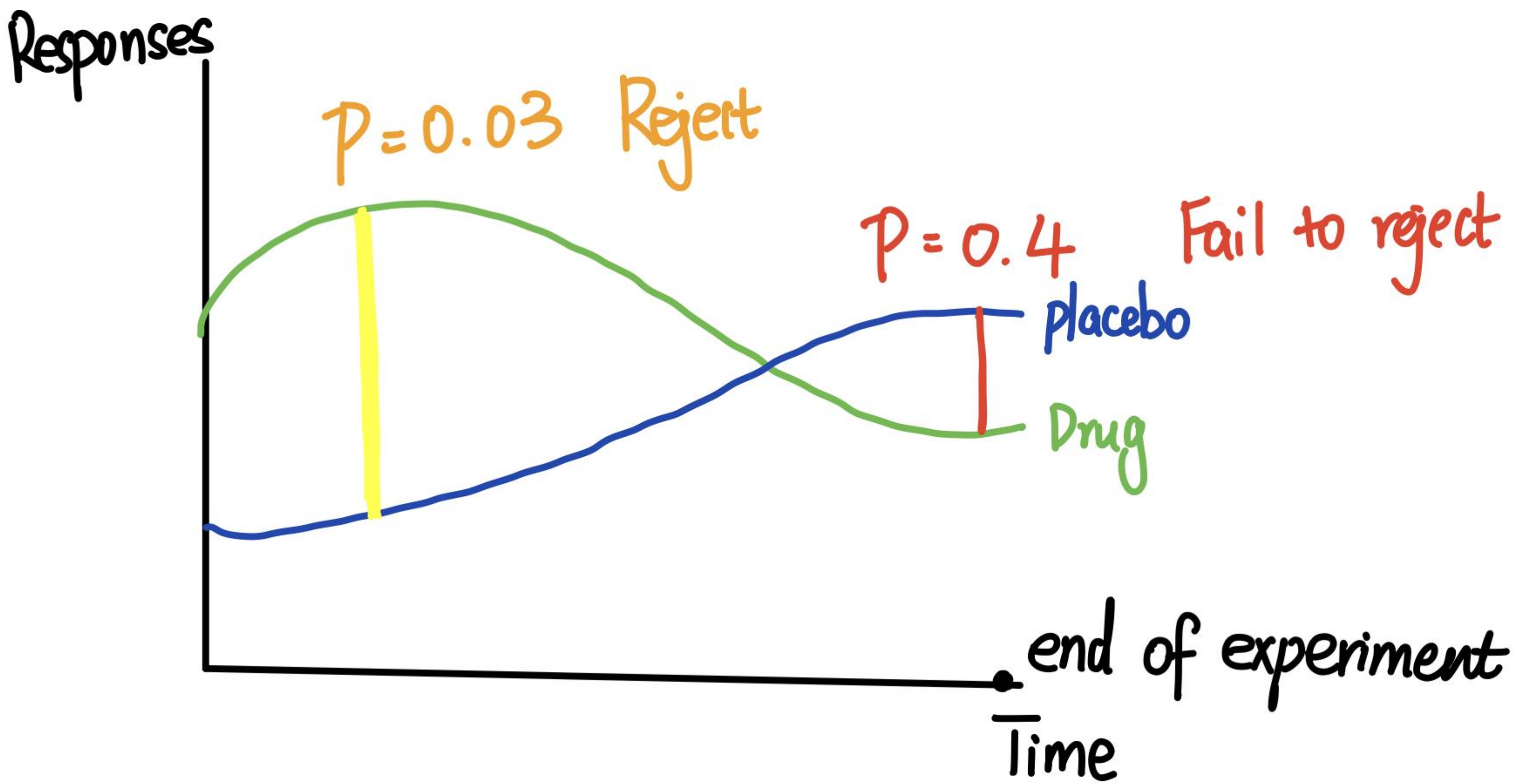
Family_Wise Error Rate

- Test every week as we recruit new patients to the trial?
- When scientists want to do repeated tests and follow treatment and control group over time, the probability of making a Type 1 error is no longer controlled!

FWER -> probability of making at least one Type I error at a specific significance level(Alpha) among multiple tests

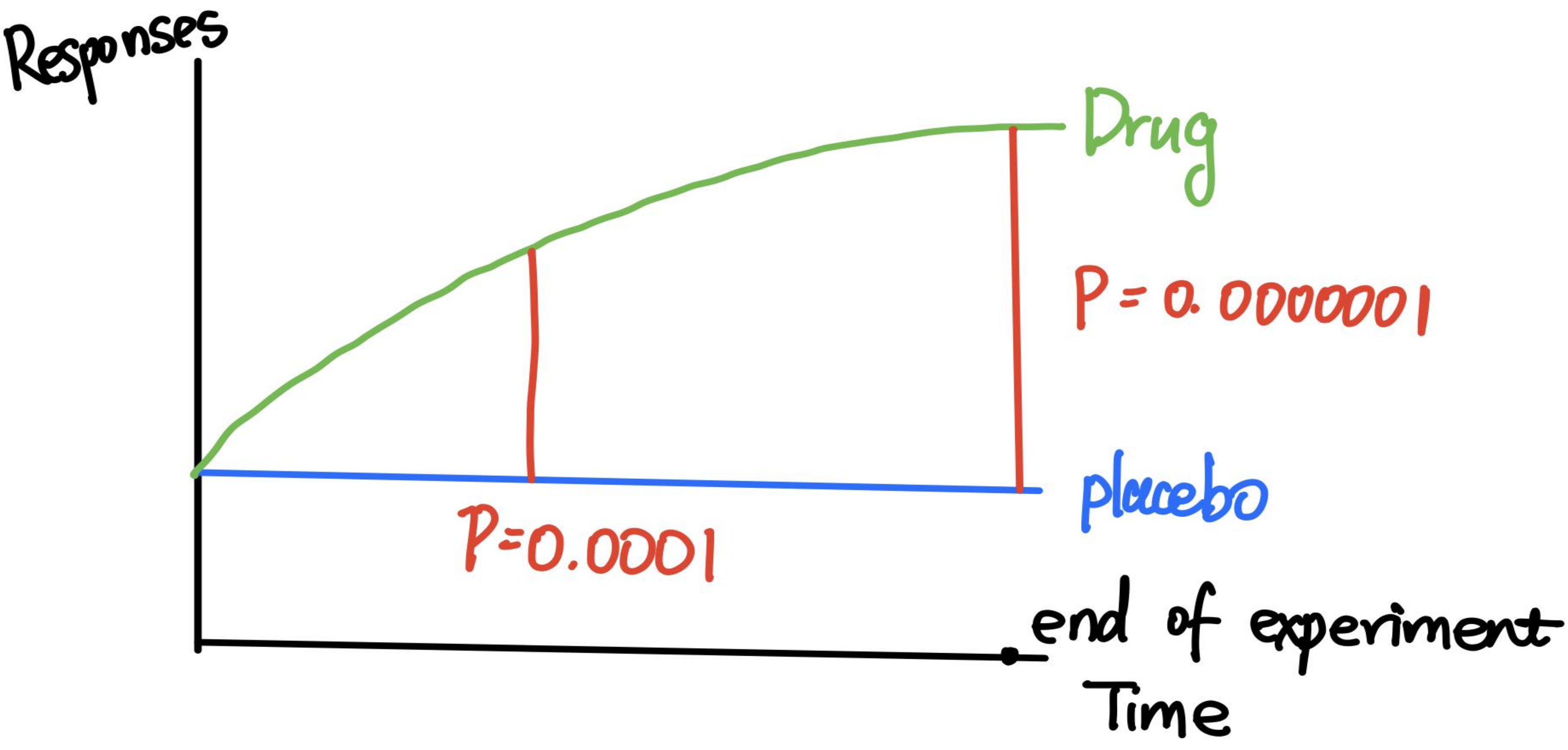
Interim Analyses

Null is true, alpha = 0.05



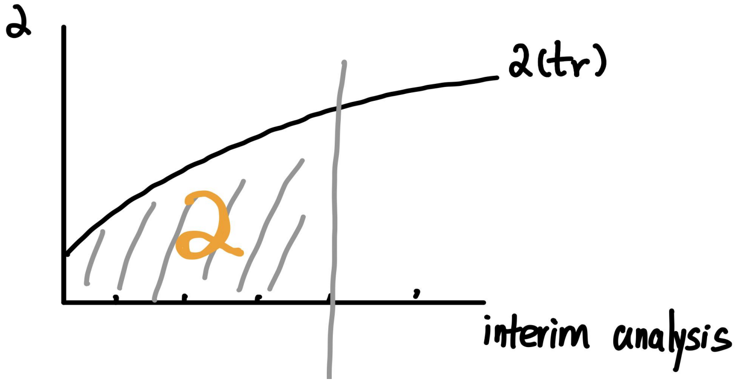
Interim Analyses

Null is false, alpha = 0.05



Multiple Testing Techniques: Alpha-spending functions

- In interim analyses, Pr(FWER) = Alpha
- Group sequential boundary: Allocate Alpha over k interim analyses
- Alpha -> an increasing function, alpha(t_r)
  (t_r) -> information fraction, 0-1



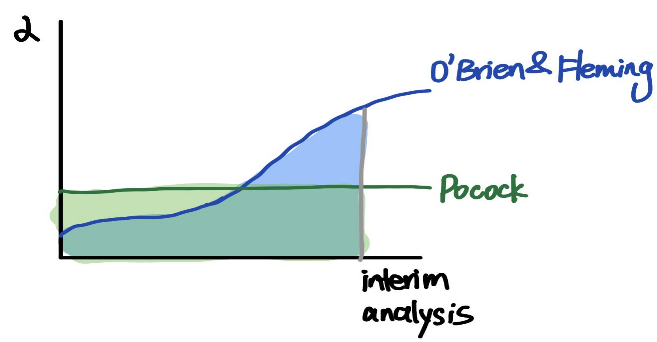
Alpha-spending functions:

- **Bonferroni Correction: (most general technique)**

fixed alpha for each analysis (alpha/m)

- **Sequential monitoring (DEPENDENCE)**

- **O'Brien and Fleming:** **more conservative stopping boundaries at early stages, larger power at the END**
- **Pocock:** **same significance level at each interim analysis, being able to stop early**



R simulation

- Verify power, interim analyses properties of Alpha-spending functions

- Sequentially monitor trials both under null (same mean for treatment and control) and under the alternative (different means, treatment_effect)

R Simulation Results (Null is True)

control at level 0.05

	FWER(probability of stopping the trial and concluding treatment and control are different)	K(average stopping time, among trials where we stopped)
No correction	0.20	3.79
Bonferroni	0.02	4.20
O'Brien & Felming	$\boxed{0.05}$	8.33
Pocock	$\boxed{0.05}$	4.24

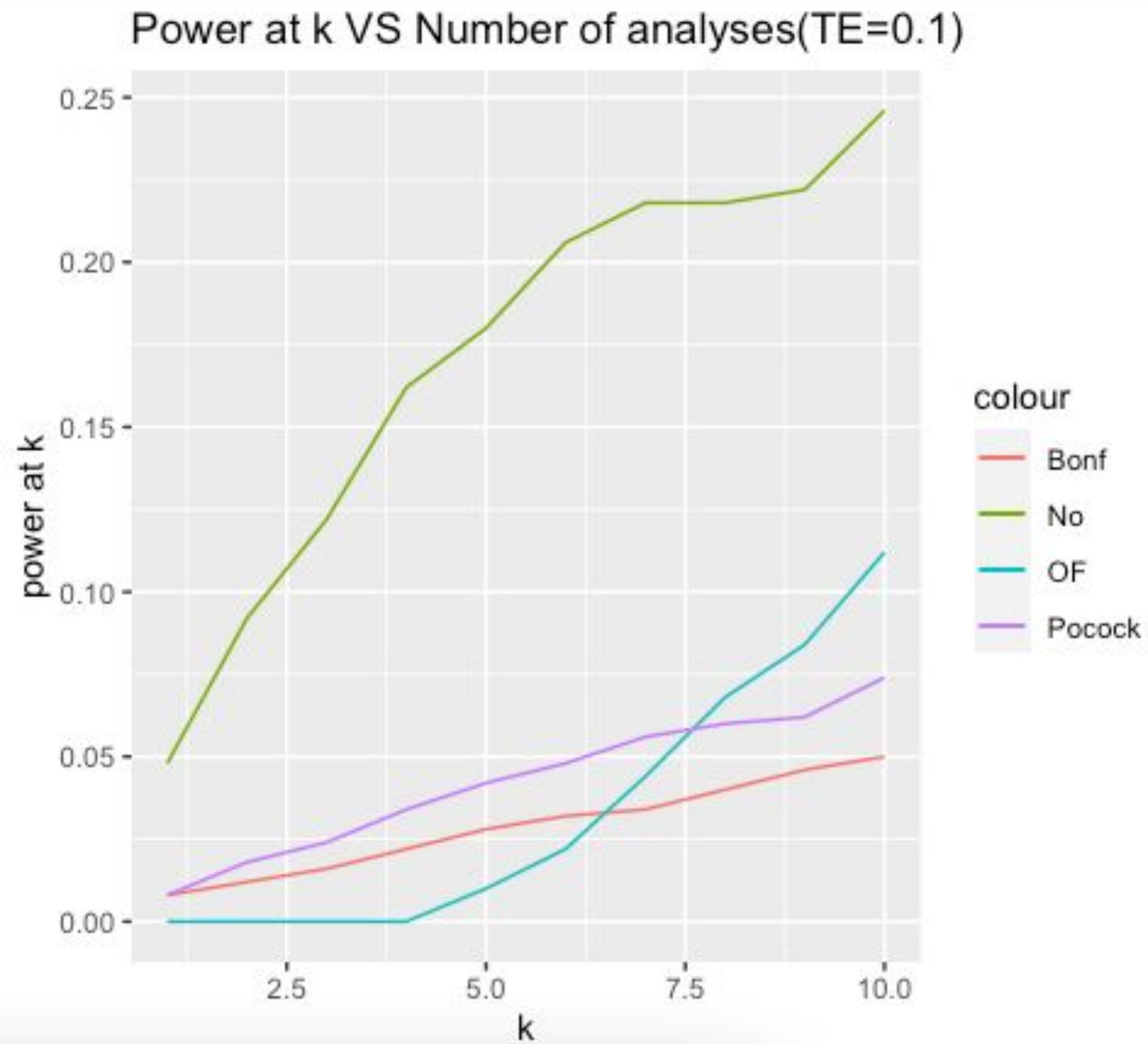
similar constant threhold

← 0.005

← 0.0106

increasing thresholds
hard to reject
at the beginning

R Simulation Results (Null is False)



Extensions:

- Pocock is more powerful than Bonferroni (dependence)

Can we do better?

- mFDR -> Reject as many null as possible while guaranteeing no more than alpha% of those rejected null are false positives

Extensions:

Alpha-spending functions:

Fixed boundary
--number of planned analyses
--initial alpha

Goal:
Control probability of making at least one type I error (FWER)

Alpha-investing functions:

Advanced boundary
--change based on results of previous test

Goal:
control a rate that depends on number of all rejected null, and number of rejected true nulls (mFDR)

THANK YOU

Acknowledgement:
-- THANK YOU Anna for guiding me through!!!
-- Appreciate the opportunity offered by DRP