Lecture 6: Handling Missing Data

Outline for lecture

- Review of concepts
- Intent-to-treat principle
- Approaches to analysis in presence of missing data
- Misconceptions about missing data

Overview: Impact of Missing data

- Introduction of bias
- Loss of precision
- Lack of internal and face validity, particularly if a large amount of data are missing

Key Questions

- What are some ways that missing data can cause bias in a study analysis?
- What are some ways to prevent missing data?
- How do we preserve intent-to-treat principle with missing data?
- What are the differences between these types of missing data: MCAR, MAR, MNAR?
- What are the methods of analysis to handle missing data and the assumptions they rely on to be appropriate?

IDEAL Clinical STUDY

- All surveys complete (no skipped questions)
- No dropouts
- No loss to follow-up
- No noncompliance (i.e. no missed visits)
- No errors

Common Causes of Missing Data

- Dropout
- Loss to follow-up
- Noncompliance with measurement procedure
  - missed visit
  - refused procedure/skipped questions or surveys
- Error
  - test not done
  - test done improperly
  - results lost
- Deliberate exclusion from analysis

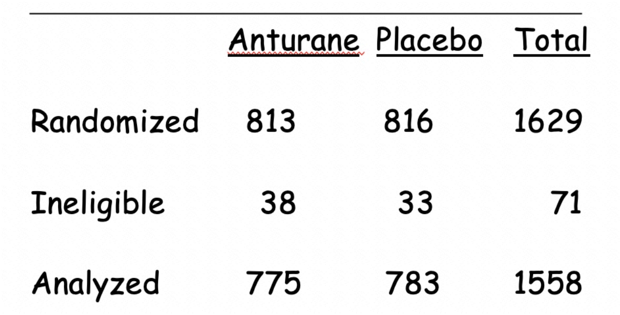
THE BIG WORRY ABOUT MISSING DATA

- Missingness may be associated with outcome and confound analysis
- We don't know the form of this association
  - Any fix we apply relies on untestable assumptions
- Nevertheless, if we fail to account for the (true) association, we may bias our results

ANTURANE REINFARCTION TRIAL (ART)

Temple and Pledger (1980)

- Randomized, double-blind, placebo-controlled trial in post-MI subjects
- Primary endpoint: mortality



	Anturane	Placebo	Total
Randomized	813	816	1629
Ineligible	38	33	71
Analyzed	775	783	1558

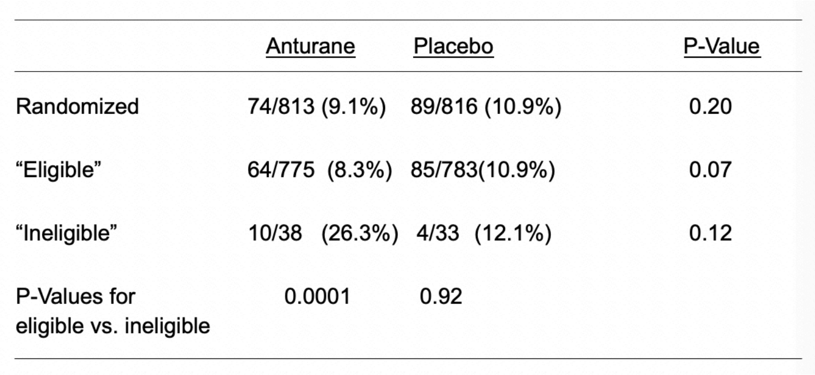
REASONS FOR INELIGIBILITY

- 1/3 - time since MI: $<$ 25 days or $>$ 35 days
- 1/3 - enzymes not elevated
- 1/3 - other: age, enlarged heart, prolonged hospitalization, ...
- Number ineligible about the same in each treatment group (38 vs 33)

But....

ANTURANE MORTALITY RESULTS

Temple & Pledger (1980) NEJM, p. 1488



	Anturane	Placebo	P-Value
Randomized	74/813 (9.1%)	89/816 (10.9%)	0.20
“Eligible”	64/775 (8.3%)	85/783(10.9%)	0.07
“Ineligible”	10/38 (26.3%)	4/33 (12.1%)	0.12
P-Values for eligible vs. ineligible	0.0001	0.92	

FDA CONCERNS

- Apparently haphazard application of “eligibility” criteria by adjudication group
- Some deaths excluded on drug arm were very similar to deaths not excluded on placebo arm
- Overall, cause of death classification deemed very unreliable
- No pre-specified plan to exclude any randomized subjects from analysis
- No issue of deliberate bias; adjudicators were blinded to treatment arm

HISTORICAL IMPORTANCE OF ART

- Anturane example established the ITT principle
  - Account for all participants randomized
  - Account for all events during follow up
- Extremely influential in FDA approach to review

INTENT-TO-TREAT (ITT) PRINCIPLE

All randomized patients should be included in the (primary) analysis, in their assigned treatment groups

THE RATIONALE FOR ITT

Randomization ensures that there are no systematic differences between the treatment groups.

The exclusion of patients from the analysis on a systematic basis (e.g., lack of compliance with assigned treatment) can introduce systematic differences between treatment groups, thereby biasing the comparison.

DEALING WITH NONCOMPLIANCE

A PERENNIAL PROBLEM

- There will always be noncompliant subjects in clinical trials
- There will always be people who don’t use medical treatment as prescribed
- Conflict between the question we may WANT to ask, and the question we CAN ask

DIFFERENT QUESTIONS

- Is this treatment safe and effective when used as directed in a well-defined population?
- Is this drug safe and effective when prescribed for its generally intended purpose by practicing physicians?

PRAGMATIC VS EXPLANATORY

- “Explanatory” question: is the product safe and effective when used appropriately in carefully defined population
  - Scientific question, what pharmaceutical companies and FDA are trying to establish
  - Evaluation of treatment product (treatment efficacy)
- “Pragmatic” question: if we put this product out there, what are the benefits and what are the risks?
  - Public health perspective
  - Evaluation of treatment policy (treatment effectiveness)

So why cant we just exclude those who did not take the treatment?

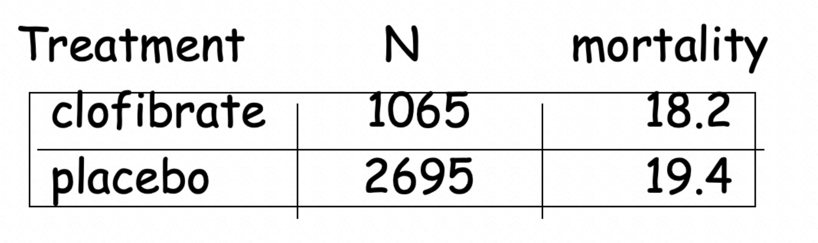
- Better outcomes in those who follow prescribed regimen could be due to the fact that they actually took the treatment
- Could be the reverse: those who are better prognostically may be more likely to be good adherers

CLASSIC EXAMPLE

- Coronary Drug Project (CDP)
    - Large RCT conducted by NIH in 1970's
    - Compared several treatments to placebo
    - Goal: improve survival in patients at high risk of death from heart disease
- Results disappointing
- Investigators recognized that many subjects did not fully comply with treatment protocol

CDP: Five-Year Mortality by Treatment Arm

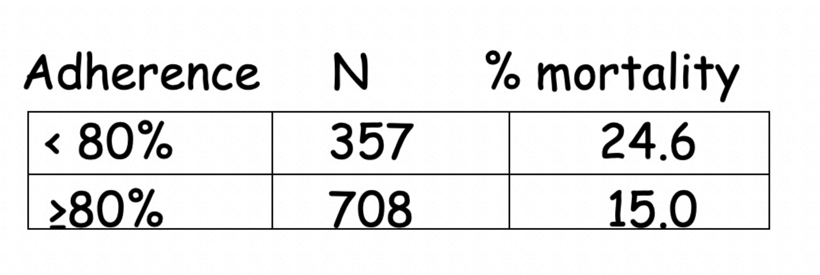
Coronary Drug Project Research Group (1975)



Treatment	N	mortality
clofibrate	1065	18.2
placebo	2695	19.4

CDP: Five-Year Mortality by Adherence to Clofibrate

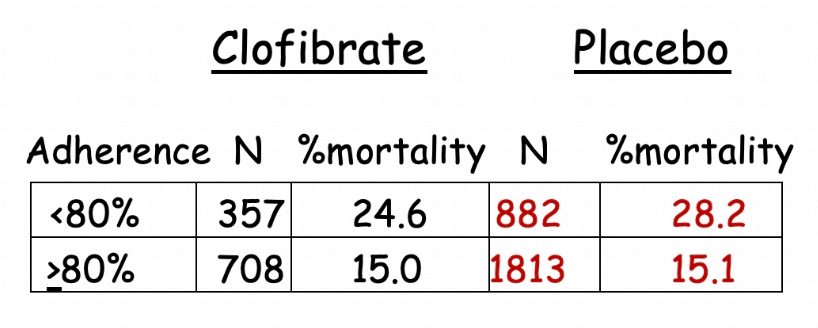
Coronary Drug Project Research Group, JAMA, 1975



Adherence	N	% mortality
< 80%	357	24.6
≥80%	708	15.0

CDP: Five-Year Mortality by Adherence to Clofibrate and Placebo

Coronary Drug Project Research Group, JAMA, 1975



	Clofibrate		Placebo	
Adherence	N	%mortality	N	%mortality
<80%	357	24.6	882	28.2
≥80%	708	15.0	1813	15.1

Implication of ITT for Study Design and Analysis

- Implication for design: collect all required data on all patients, regardless of compliance
  - Keep on study even if off treatment
- Implication for analysis: Need to model the missingness
  - Common approaches: Have to impute either the probability that each participant would have compete data or impute the missing outcome
  - To be successful, need to observe the variables associated with missingness
  - This is hard. Thus, avoid missing data

Estimands, Estimates, and Estimators

Little and Lewis (2021)

- Estimand- a target quantity; ie, what the study aspires to measure
- Estimator -a formula or algorithm used to estimate the target quantity from the clinical trial data
- Estimate - the numeric value obtained when the estimator is applied to the actual data from the trial.

Another good reference: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/e9r1-statistical-principles-clinical-trials-addendum-estimands-and-sensitivity-analysis-clinical

ITT and the Causal Estimand

- The ITT analysis addresses the treatment policy estimand– difference between randomized groups, regardless of compliance
- Some argue may ITT analysis may not be estimating the ideal causal estimand
  - When analyze outcome regardless of compliance, not getting at ”pure” treatment effect
- An estimand evaluating the causal treatment effect can be much more difficult to estimate
  - Have to address missing data
  - May involve a “Hypothetical” outcome

Alternative Causal Estimands

- Consider “Principal Stratum” – treatment effect for hypothetical scenarios
  - i.e. Treatment effect if everyone took treatment as directed (potential outcomes)
  - Inference may be for a hypothetical scenario not of clinical interest or practical to estimate. Need to live in a world with no toxicity, no other cause death, etc.
  - Anti-HIV Microbicide example those with toxicity were “non-adherent” to protocol and were at increased risk.

See Olarte Parra et al. (2021) for nice discussion of alternative estimands in clinical trials.

HANDLING MISSING VALUES IN ANALYSIS

MANY DIFFERENT APPROACHES

- Exclude subjects with missing values (“Completers” analysis or “complete case” analysis)
- Last Observation Carried Forward (LOCF)
- Baseline Observation Carried Forward (BOCF)
- Group means
- Inverse probability weighting
- Multiple imputation
- Causal modeling for desired estimand

Taxonomy for Missing Data

Little et al. (2012)

- Missing completely at random (MCAR)
  - Missingness mechanism depends neither on the observed or unobserved data
- Missing at random (MAR)
  - Missingness mechanism depends on data that are observed, but not data that are not observed
- Missing not at random (MNAR), or nonignorable missing

MISSING COMPLETELY AT RANDOM (MCAR)

- Missing data are independent of any covariates and independent of the true outcome
- Fact that the data are missing provides no information about outcome, nor could it be predicted from other measured variables
- Under MCAR, completers analysis will give unbiased estimate of treatment effect
- Rare that we can be confident that missing data in clinical trials are MCAR

MISSING AT RANDOM (MAR)

- Missingness is associated with other variables we measure, but once we account for those variables missingness is not associated with the missing value
- Implication: we can predict the missing outcome in an unbiased way from characteristics of subject, of other subjects, and on observed outcomes of other subjects
  - Two common approaches: multiple imputation, inverse probability weighting

MISSING NOT AT RANDOM (MNAR)

- Also referred to as “nonignorable” missing
- Missingness depends on unobserved events/characteristics; so one cannot make unbiased estimates of actual outcome using other measured variables
- Incorporation of subjects with missing data in the analysis requires modeling the missing data mechanism
  - Generally takes form of sensitivity analysis of results under different scenarios or a “pattern mixture” modeling

IMPLICATIONS FOR ANALYSIS

- MCAR
  - Inference for subjects with complete data will be same as for subjects with missing data; can just exclude dropouts
- MAR
  - Inference for subjects with missing data will be the same after accounting for baseline characteristics and observed data for other participants
- MNAR
  - Requires additional assumptions about missingness mechanism

EXCLUSIONS OF SUBJECTS

- Excluding subjects with missing values “completers analysis” is simplest approach
  - requires assumption that excluded subjects are a random subset of all randomized subjects: MCAR
- Approach generally not recommended unless amount of missing data is minimal ($<$5%)
  - Assumptions too difficult to justify
  - Complete case data does not take advantage of the partial data available on excluded participants
- Unfortunately very common approach; often, no acknowledgement of basic assumptions

IMPUTATION

- Determine a value that is “best guess” of true value of missing data point
- Several approaches proposed and/or in use
- Simplest approaches are generally statistically most problematic
- Any approach involving “made-up” data is problematic to some degree

SINGLE IMPUTATION APPROACHES

- Last Observation Carried Forward (LOCF)
  - Use last measurement available in patients with missing data after a certain point
- Baseline Observation Carried Forward (BOCF)
  - Use the measurement at baseline as final data point
- Group means
  - Assign average value of outcome variable among those in that treatment group (or in total study population) with complete data

Historically popular, Statistically not recommended

LAST OBSERVATION CARRIED FORWARD (LOCF)

- Used in trials with repeated measures but where primary comparison is final value to baseline value
- Concept: whatever the last measurement was prior to dropout, use that as the final value
- Commonly used in trials evaluating symptom relief for new products

EXAMPLE

- Trials of new antidepressants generally use the Hamilton Depression (Ham-D) scale as the primary outcome
- Trials typically last 4-8 weeks; subjects are evaluated weekly
- Primary comparison is value at final week to baseline value
- If subject drops out after week 3 evaluation, week 3 score is used to assess effect in that subject

APPEAL OF LOCF

- It's simple
- Allows all randomized subjects receiving at least one evaluation to be included in final analysis: looks sort of like ITT
- Assumption is that dropout is related to lack of effect; scores after dropout may improve if subject initiates active medication
- In certain settings, some (including regulators) have argued this approach is conservative (unlikely to inflate Type 1 error) but. . . .

PROBLEMS WITH LOCF

- The last observation before dropout is not a reliable estimate of the effect of the treatment at the desired time point
- Variance underestimated: get same credit for full sample size as if all data were available
- Will not always be conservative!
- Generally regarded as a suboptimal method

WHY IS LOCF NOT ALWAYS CONSERVATIVE?

- If active treatment has side effects that cause subjects to discontinue, efficacy data may look OK at time of treatment stop but would be expected to worsen once subject is off treatment
- If tolerability is a major cause of treatment stop, LOCF analysis could overstate treatment benefit
- Allowing use of entire sample size increases power for detecting differences (that may not be real)

BASELINE OBSERVATION CARRIED FORWARD

- Same idea as LOCF, but use baseline as final value
- Often used in pain studies; assumption is that subjects abandoning treatment are getting no relief
- Idea again is to be conservative; assume no improvement from baseline
- Big problem: people might have worsened on assigned treatment, so could be anticonservative
- As with LOCF, get credit for more data than actually have
- Generally regarded as a suboptimal method

GROUP MEANS

- Assumption: best estimate of effect for individual is the average effect of that individual's treatment group
- This is equivalent to simply excluding dropouts from analysis, except worse—we give ourselves credit for a larger sample size and so reduce variance
- Generally regarded as a suboptimal method
- Perhaps more conservative: use average effect of both groups combined or average effect from the control group –but still not accounting for added uncertainty from missing data

MORE SOPHISTICATED: MODEL-BASED APPROACHES (MAR)

- Multiple Imputation: Predict missing outcome on basis of outcomes for other patients with similar characteristics
- Inverse probability weighting – Predict probability of being complete and perform weighted regression, inversely weighting records according to missingness probability
- These approaches will yield comparisons with little to no bias if the data are “missing at random” and the models correct

MULTIPLE IMPUTATION

- LOCF, BOCF, group means are all examples of single imputation methods
  - Choose an estimate for the missing data point, and insert that, so that the subject can be included in the analysis
  - This essentially gives you credit for an observation you don't have; results in an underestimate of variability, artificially increases precision of estimate
- Multiple imputation
  - A way to account for the extra variability that is inherent in estimating the missing data point

MI: Basic Idea

- Create a model to predict missing value on basis of covariates and outcomes
  - Develop this model on those with complete data
- Apply this model to obtain multiple versions of the “completed data” – impute multiple estimates of possible data points for each missing observation
- Data analysis proceeds by using each of these completed datasets to get the target estimate, then aggregates results into an overall results
- Variability among outcomes for subjects with missing data incorporated into overall variability

Simple Example: Regression based imputation

Molenberghs et al. (2014)

Suppose only a continuous outcome is missing

Step 1: Use complete cases to build model and predict outcome Y based on covariates observed on everyone

- That prediction model will have regression coefficients (α) that have uncertainty and residual unexplained variance (σ^2)

Step 2: Draw $\hat{\alpha}^{(m)}$ and $\hat{\sigma}^{(m)}$ from “posterior” distributions

- Note, a “proper” imputation would take into account the uncertainty of the model parameters and prediction error

Step 3: For each missing y_i, draw $\varepsilon_i^{(m)}\ N(0, \hat{\sigma}^{(m)})$ and impute (predict) the missing outcome: $y_i^{(m)} = X_i^t \hat{\alpha}^{(m)} + \varepsilon^{(m)}$

Step 4: Do outcome regression on completed data to obtain parameter estimate $\hat{\beta}^{(m)}$ of interest

Step 5: Repeat Step 2-4 m times (10,25,50...)

Step 6 : Average β and get Rubin’s variance/df

Rubin's Rules

- Each imputation yields $\hat{\beta}^m$ and $\mathrm{var}(\hat{\beta}^m)$
- Pool the estimates across the M imputations:

  $\hat{\beta}_{MI} = \text{Avg}(\hat{\beta}^m)$

  $\mathrm{var}(\hat{\beta}_{MI}) = \text{Avg}(\mathrm{var}(\hat{\beta}^m))) + \mathrm{var}(\hat{\beta}_{MI})$ (Rubin's rules)
- Note the typical formula for Rubin's rules was for a small number of imputations (M≈5) and had an adjustment to the 2nd variance term that isn't needed for large M
- For reproducibility of results, today M≈25 or larger is much more typical and better for reproducibility
- A few references for practical considerations are White et al. (2011) and Von Hippel (2020)

More advanced multiple imputation techniques

- This is a deep topic that needs its own short course
- Multivariate Imputation by Chained Equations (MICE) (mice package in R, proc MI in SAS) - this involves specifying a set of equations for how each variable can be predicted. (e.g. may assume multivariate normality if all variables continuous.)
  - Don't assume the package defaults are okay for your problem
  - Generally a Gibbs sampler type algorithm and needs a "burn-in"
  - There are pitfalls to be aware of with common algorithms: in that you could specify conditional models that are not compatible with a joint distribution, which can lead to inflated variance estimates.
- *Multiple Imputation of Covariates by Substantive Model Compatible Fully Conditional Specification* is one alternative in R and Stata (smcfcs package) that addresses common pitfalls of MICE (Bartlett et al., 2015)

Handling MAR in clinical trials

- In the simple case of missing outcome only, adjustment of the outcome model for covariates will address bias from missingness under MAR
    - Groenwold et al. (2012) show you get similarly consistent estimates with MI and adjusted complete case analysis.
    - Some argue (Sullivan et al., 2018) that non-MI approaches should be considered more often under MAR (e.g. mixed modeling or adjustment method), with caveat MI still offers chance to be more precise
- MI approaches allow you to handle MAR and still do the original statistical test (e.g. two group comparison).
- MI approaches directly satisfy ITT

INVERSE PROBABILITY WEIGHTING

- If the probability of having an observed value (i.e., not missing) is closely related to the values of other measured variables, this method can be useful
- Weights observed values inversely by the probability of being observed
- Can reduce bias if the assumption that the probability of being observed is a function of other measured variables is reasonable
  - In other words, if we are in a “missing at random” situation

Nitty Gritty: IPW

Step 1: Create a binary variable indicating completeness outcome

- 1=complete case
- 0=non-complete case (missing at least one of: outcome or key covariate)

Step 2: Model this completeness outcome (e.g. with logistic regression

- Include characteristics thought to be related to missingness (need to be observed on most people)

Step 3: Create IPW weights for each person= 1/predicted probability from logistic model

Step 4: Perform a weighted regression using IPW

Step 5: Final analysis should consider variability in the weights: either using a bootstrap or sandwich approach ((Carpenter and Kenward, 2007))

You can examine the success

- Should evaluate the predictive accuracy of the IPW model.
  - Could consider AUC for the ROC for phat to predict the binary outcome complete (yes/no)
  - AUC close to 0.50 means your model could not predict which data were complete.
  - Poor AUC means wont have a good adjustment for missingness
- Analytical concern: sometimes large weights can create instability
  - Look for influential points in regression
  - More sophisticated algorithms are available that stabilize weights
- IPW approach can be more variable than MI
  - Classic Bias/variance tradeoff

Take aways

- Validity of analytical approaches under MAR depend on assumptions that we can model the systematic mechanisms driving the missing data
- Build models to predict missing outcome or to predict probability of being observed
- IPW seen as more “robust” (relying on fewer assumptions) than MI
- MI seen as more “efficient” (less variable) than IPW
- MI approach more common, more intuitive way to preserve ITT and efficiency is attractive

What to do if data are MNAR?

SENSITIVITY ANALYSES

- Analyze data under variety of different assumptions regarding missing data—see how much the inference changes
- Sensitivity analyses could involve any of approaches already described, and others
- When multiple sensitivity analyses are done without changing conclusions, we can be more comfortable that missing data are not obscuring important information about treatment effect

SIMPLE SENSITIVITY ANALYSIS: Impute the most extreme scenarios

- Provide bounds for the “true” results if all planned data points had been observed
- Provides a sense of how far off any of the other analyses could be
- Does not provide a single “answer” but aids in interpretation of other “answers”

EXAMPLE

- Suppose outcome for each subject is “response” or “nonresponse”
- In worst case analysis, assume each missing subject in treatment group has “nonresponse” and each missing subject in control group has “response”
- In best case analysis, reverse
- Less straightforward but still feasible when outcome is a continuous rather than a binary variable
  - In weight loss trial, there are biological limits to how much weight could change in 6 months.

More Complex Sensitivity Analyses

- “Pattern mixture” methods consider how patterns in the missing data may be different than the observed patterns
- Can show how departures from the MAR assumption can influence conclusions
- Tipping point analysis: e.g. conjecture differential outcome models until results are reversed. Judge to what extent that scenario would be possible/believable.
- Using different models and showing minimal impact on conclusions will support primary findings

MISCONCEPTIONS ABOUT HANDLING MISSING DATA

MISCONCEPTIONS ABOUT HANDLING OF MISSING DATA

1. If a dropout rate of x% is expected, the sample size should be adjusted upward by x% to compensate.

MISCONCEPTIONS ABOUT HANDLING OF MISSING DATA

1. If a dropout rate of x% is expected, the sample size should be adjusted upward by x% to compensate.

   Problem: This provides the desired number of data points for precision, but the results may be biased

   Example: Suppose dropouts are those who are not benefiting from treatment. Then we are hiding treatment failures by replacing such patients.

   Not wrong to enlarge sample size; it just doesn't make the problem go away

Key consideration: to safeguard power, likely need to assume a certain % on tx arm wont respond /has a worse outcome... reduces power by bringing arms closer

MISCONCEPTIONS ABOUT HANDLING OF MISSING DATA

2. If the amount of missing data is the same in each study arm, everything is OK

MISCONCEPTIONS ABOUT HANDLING OF MISSING DATA

2. If the amount of missing data is the same in each study arm, everything is OK

   Problem: The reasons for data being missing may not be the same for each arm. On one arm, drop outs may be those doing well; on the other, those doing poorly

MISCONCEPTIONS ABOUT HANDLING OF MISSING DATA

3. If the baseline characteristics of subjects with missing data are similar to those of subjects with complete data, it is probably safe to limit analysis to “completers”

MISCONCEPTIONS ABOUT HANDLING OF MISSING DATA

3. If the baseline characteristics of subjects with missing data are similar to those of subjects with complete data, it is probably safe to limit analysis to “completers”

> Problem: Many aspects of prognosis are not understood and therefore not measured. The balance on unknown prognostic factors provided by randomization is no longer assured when randomized subjects are excluded.
>
> Note: this is the big concern about methods assuming MAR.

Final note on Reporting

- CONSORT Guidelines for Clinical Trial Reporting (Altman et al., 2001) have detailed recommendations on how to report missing data
- Must account for all randomized subjects/all individuals in original cohort
- Describe pre-specified approach to handling missing data in analysis
- Address extent to which results may be biased due to missing data
- Report results of sensitivity analyses

SUMMARY

- The less missing data, and the lower the rate of noncompliance, the less concern about
    - bias
    - unreliable conclusions
    - inappropriate methods of analysis
- Methods to replace/account for missing data are all problematic in important ways – but put your best foot forward! (In analysis, try to address bias, incorporate uncertainty)
- Sensitivity analyses are essential to evaluating reliability of conclusions

KEY REFERENCES

- Missing data in randomised controlled trials— a practical guide by John Carpenter (`https://researchonline.lshtm.ac.uk/id/eprint/4018500/1/rm04_jh17_mk.pdf`)
- National Research Council report: “The Prevention and Treatment of Missing Data in Clinical Trials” (National Academy Press, 2010)
  - Detailed discussion of different methods for handling missing data, with examples
- Little RJ, D’Agostino R, Cohen ML, Dickersin K, Emerson SS, Farrar JT, Frangakis C, Hogan JW, Molenberghs G, Murphy SA, Neaton JD. The prevention and treatment of missing data in clinical trials. NEJM 2012;367(14):1355-60.
  - Overview discussion of different methods for handling missing data, with examples
- Schulz KF, Altman DG, Moher D. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. Trials. 2010 Dec;11(1):1-8.

References I

Altman, D. G., Schulz, K. F., Moher, D., Egger, M., Davidoff, F., Elbourne, D., Gøtzsche, P. C., Lang, T., and Group, C. (2001). The revised consort statement for reporting randomized trials: explanation and elaboration. Annals of internal medicine **134,** 663–694.

Bartlett, J. W., Seaman, S. R., White, I. R., Carpenter, J. R., and Initiative*, A. D. N. (2015). Multiple imputation of covariates by fully conditional specification: accommodating the substantive model. Statistical methods in medical research **24,** 462–487.

Carpenter, J. R. and Kenward, M. G. (2007). Missing data in randomised controlled trials: a practical guide.

Coronary Drug Project Research Group (1975). Clofibrate and niacin in coronary heart disease. J Am Med Assoc **231,** 360–381.

Groenwold, R. H., Donders, A. R. T., Roes, K. C., Harrell Jr, F. E., and Moons, K. G. (2012). Dealing with missing outcome data in randomized trials and observational studies. American journal of epidemiology **175,** 210–217.

Little, R., D'Agostino, R., Cohen, M., Dickersin, K., Emerson, S., Farrar, J., Frangakis, C., Hogan, J., Molenberghs, G., Murphy, S., and Neaton, J. (2012). The prevention and treatment of missing data in clinical trials. NEJM **367,** 1355–60.

Little, R. J. and Lewis, R. J. (2021). Estimands, estimators, and estimates. JAMA **326,** 967–968.

Molenberghs, G., Fitzmaurice, G., Kenward, M. G., Tsiatis, A., and Verbeke, G. (2014). Handbook of missing data methodology. CRC Press.

References II

Olarte Parra, C., Daniel, R. M., and Bartlett, J. W. (2021). Hypothetical estimands in clinical trials: a unification of causal inference and missing data methods. arXiv e-prints pages arXiv–2107.

Sullivan, T. R., White, I. R., Salter, A. B., Ryan, P., and Lee, K. J. (2018). Should multiple imputation be the method of choice for handling missing data in randomized trials? Statistical methods in medical research **27,** 2610–2626.

Temple, R. and Pledger, G. W. (1980). The fda's critique of the anturane reinfarction trial. New England Journal of Medicine **303,** 1488–1492.

Von Hippel, P. T. (2020). How many imputations do you need? a two-stage calculation using a quadratic rule. Sociological Methods & Research **49,** 699–718.

White, I. R., Royston, P., and Wood, A. M. (2011). Multiple imputation using chained equations: issues and guidance for practice. Statistics in medicine **30,** 377–399.

Lecture 7: Multiple Comparisons

Introduction

Strong/Weak Control of FWER

Methods for Strong Control of FWER
- Bonferroni
- Holm’s Sequentially Rejective Bonferroni
- Graphical Alpha Transfer
- The Closure Principle

Multiple Arms
- Dunnett and Variants

Multiple Subgroups
- Forest Plots
- Permutation Methods
- Hochberg

False Discovery Rate (FDR)

Outline

Introduction

Strong/Weak Control of FWER

Methods for Strong Control of FWER

- Bonferroni
- Holm’s Sequentially Rejective Bonferroni
- Graphical Alpha Transfer
- The Closure Principle

Multiple Arms

- Dunnett and Variants

Multiple Subgroups

- Forest Plots
- Permutation Methods
- Hochberg

False Discovery Rate (FDR)

Introduction

- See Hochberg and Tamhane (1987) for thorough treatment and Proschan and Brittain (2020) for a tutorial on multiple comparisons.

- Multiple comparisons arise in many ways in clinical trials: Arms, endpoints, subgroups, analyses, covariates, time points (monitoring).

- Problem: With enough comparisons, some will be significant even if nothing is going on.

  > *"If you torture the data long enough, it will confess to anything."*

  Ronald H. Coase, British economist.

- Sometimes the size of the multiplicity problem is not clear.

Introduction: Magnitude of Problem

- In 2003, VaxGen announced their AIDSVAX HIV vaccine may be effective in Blacks and Asian-Americans.
  - 5,108 men who have sex with men (MSM) and 309 high risk women in North America and the Netherlands.
  - Overall, only a 3.8% reduction in vaccine arm compared to placebo (p=0.76).
  - Blacks: 4/203 vaccine versus 9/111 placebo participants got HIV.
  - Asians: 2/53 vaccine versus 2/20 placebo participants got HIV.

    > *"If we announced to the world that we were abandoning the project because the study failed in whites, we'd be crucified."*

    VaxGen CEO Lance Gordon.
- Very controversial.
- How many subgroups did they examine?

Introduction: Magnitude of Problem

- ISIS-2 trial evaluated effect of aspirin on mortality in heart patients (ISIS-2 (1988)).
- Journal asked authors to report subgroup results, but they felt it would mislead. To prove that, they facetiously reported:

  "...for patients born under Gemini or Libra there was a slightly adverse effect of aspirin on mortality (9% SD 13 increase; NS), while for patients born under all other astrological signs there was a strikingly beneficial effect (28% SD 5 reduction; 2p<0.00001)."

- They combined 2 astrological signs and compared to others; number of potential comparisons huge!

Introduction: Magnitude of Problem

- Similar examples can be found in other fields.
- Bible code controversy (Drosnin (1997); History-Channel (2003)).
  - Believers (including Sir Isaac Newton) say there is secret code in first 5 books of the Old Testament.
  - Must skip letters (e.g., read every 50th letter) to uncover code words like “Bin Laden”, “twin towers”, etc.
  - Code supposedly proven by Witztum et al. (1994) by permutation test: Code revealed that rabbis’ names were too close to dates of birth to be explained by chance.
- Debunked by Bar-Hillel et al. (1999).

Outline

- Introduction
- Strong/Weak Control of FWER
- Methods for Strong Control of FWER
  - Bonferroni
  - Holm’s Sequentially Rejective Bonferroni
  - Graphical Alpha Transfer
  - The Closure Principle
- Multiple Arms
  - Dunnett and Variants
- Multiple Subgroups
  - Forest Plots
  - Permutation Methods
  - Hochberg
- False Discovery Rate (FDR)

Strong/Weak Control of FWER

- Consider testing hypotheses $H_1, \ldots, H_k$.
- ***Global null hypothesis*** is $\cap_i H_i$.
- ***Familywise error rate (FWER)*** is
  P(at least 1 false rejection of a null hypothesis).
- Can calculate FWER under global null or under other nulls.
- ***Weak control of FWER*** means FWER is controlled if global null is true.
- ***Strong control of FWER*** means FWER is controlled no matter which nulls are true.

Strong control $\Rightarrow$ weak control.

Strong/Weak Control of FWER: Fisher's LSD

- Example: Comparing all pairs of means of arms 1,2,3,4.

- ***Fisher's least significant difference (LSD) procedure***: If F-test comparing all arms is significant at level α, compare each pair with level α t-test.

- Fisher's LSD controls FWER weakly, but not strongly.

- Weak control: To declare any differences, F must be significant, and $P(F \text{ significant} \mid \text{global null}) = \alpha$.

- Therefore, Fisher's LSD controls FWER weakly.

Strong/Weak Control of FWER: Fisher's LSD

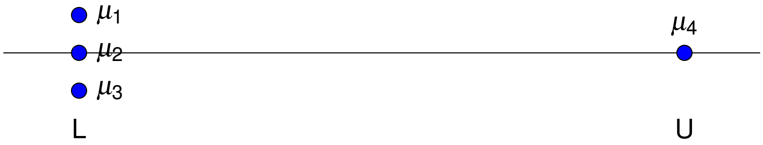


Figure: Configuration of 4 population means that inflates the FWER for Fisher's LSD: $\mu_1 = \mu_2 = \mu_3 = L$ and $\mu_4 = U$, where $U - L$ is huge.

- Lack of strong control of Fisher's LSD with 4 arms: Suppose $\mu_1 = \mu_2 = \mu_3 = L$, $\mu_4 = U$, where $U - L$ is huge.
  - F is almost guaranteed significant.
  - Fisher's LSD almost same as doing unadjusted t-tests, which inflates FWER among μ_1, μ_2, μ_3.
  - So Fisher's LSD does not strongly control FWER for ≥ 4 arms.

Strong/Weak Control of FWER: Fisher's LSD

- Sometimes easy to show a procedure strongly controls FWER by enumerating all possibilities.

- Can use method to show Fisher's LSD **with 3 arms** does strongly control FWER.

- Why?
  - If global null is true, P(F significant)$\leq \alpha$, so FWER protected.
  - If global null is false, at most one pairwise null is true and its t-test protects type 1 error rate if that pairwise null is true.

- This argument shows that Fisher's LSD strongly controls FWER when $k = 3$ arms.

Outline

Introduction

Strong/Weak Control of FWER

Methods for Strong Control of FWER

- Bonferroni
- Holm's Sequentially Rejective Bonferroni
- Graphical Alpha Transfer
- The Closure Principle

Multiple Arms

- Dunnett and Variants

Multiple Subgroups

- Forest Plots
- Permutation Methods
- Hochberg

False Discovery Rate (FDR)

Strong Control Methods: Bonferroni

- ***Bonferroni method*** always strongly controls FWER.
- Bonferroni method requires $P_i \leq \alpha/k$ to reject H_i, where P_i is p-value for ith comparison and $k =$number comparisons.
- Why does Bonferroni strongly control FWER?
  - Suppose t is number of true nulls, & without loss of generality, assume they are the first t.
  - By Bonferroni inequality,

$$\begin{aligned} P(\cup_{i=1}^{t} \text{reject } H_i) &\leq \sum_{i=1}^{t} P(\text{reject } H_i) \\ &\leq t(\alpha/k) \leq k(\alpha/k) \\ &= \alpha. \end{aligned} \quad (1)$$

Strong Control Methods: Bonferroni

- Bonferroni too conservative if statistics highly correlated.
- If statistics had correlation 1, then could use level α for each.
- If statistics independent, Bonferroni only slightly conservative:

$$\begin{aligned} FWER &= P(\cup_{i=1}^{k} \text{reject } H_i) \\ &\geq \sum_{i=1}^{k} P(\text{reject } H_i) - \sum\sum_{1 \leq i < j \leq k} P(\text{reject } H_i \cap \text{reject } H_j) \\ &= k(\alpha/k) - \binom{k}{2}(\alpha/k)^2 \\ &\geq \alpha - \alpha^2/2. \text{ Thus,} \end{aligned}$$

$$\alpha - \alpha^2/2 \leq \text{FWER} \leq \alpha. \tag{2}$$

- If $\alpha = 0.05$, then $0.04875 \leq \text{FWER} \leq 0.05$.

Strong Control Methods: Holm's Sequentially Rejective Bonferroni

- ***Holm's sequentially rejective Bonferroni method (Holm (1979)):***
  - Order p-values $P_{(1)} < P_{(2)} < \ldots < P_{(k)}$.
  - Compare $P_{(1)}$ to α/k.
    - If $P_{(1)} > \alpha/k$, stop and declare nothing significant.
    - if $P_{(1)} \leq \alpha/k$, reject null associated with $P_{(1)}$ and proceed to next step.
  - Compare $P_{(2)}$ to $\alpha/(k-1)$.
    - If $P_{(2)} > \alpha/(k-1)$, stop and declare nothing else significant.
    - If $P_{(2)} \leq \alpha/(k-1)$, reject null associated with $P_{(2)}$ and proceed to next step.
  - Compare $P_{(3)}$ to $\alpha/(k-2)$.
    - If $P_{(3)} > \alpha/(k-2)$, stop and declare nothing else significant.
    - If $P_{(3)} \leq \alpha/(k-2)$, reject null associated with $P_{(3)}$ and proceed to next step.

  $\vdots$

Strong Control Methods: Holm's Sequentially Rejective Bonferroni

- ***Holm's sequentially rejective Bonferroni method (Holm (1979)):***
  - Order p-values $P_{(1)} < P_{(2)} < \ldots < P_{(k)}$.
  - Compare $P_{(1)}$ to α/k.
    - If $P_{(1)} > \alpha/k$, stop and declare nothing significant.
    - if $P_{(1)} \leq \alpha/k$, reject null associated with $P_{(1)}$ and proceed to next step.
  - Compare $P_{(2)}$ to $\alpha/(k-1)$.
    - If $P_{(2)} > \alpha/(k-1)$, stop and declare nothing else significant.
    - If $P_{(2)} \leq \alpha/(k-1)$, reject null associated with $P_{(2)}$ and proceed to next step.
  - Compare $P_{(3)}$ to $\alpha/(k-2)$.
    - If $P_{(3)} > \alpha/(k-2)$, stop and declare nothing else significant.
    - If $P_{(3)} \leq \alpha/(k-2)$, reject null associated with $P_{(3)}$ and proceed to next step.

  $\vdots$

Strong Control Methods: Holm's Sequentially Rejective Bonferroni

- ***Holm's sequentially rejective Bonferroni method (Holm (1979)):***
  - Order p-values $P_{(1)} < P_{(2)} < \ldots < P_{(k)}$.
  - Compare $P_{(1)}$ to α/k.
    - If $P_{(1)} > \alpha/k$, stop and declare nothing significant.
    - if $P_{(1)} \leq \alpha/k$, reject null associated with $P_{(1)}$ and proceed to next step.
  - Compare $P_{(2)}$ to $\alpha/(k-1)$.
    - If $P_{(2)} > \alpha/(k-1)$, stop and declare nothing else significant.
    - If $P_{(2)} \leq \alpha/(k-1)$, reject null associated with $P_{(2)}$ and proceed to next step.
  - Compare $P_{(3)}$ to $\alpha/(k-2)$.
    - If $P_{(3)} > \alpha/(k-2)$, stop and declare nothing else significant.
    - If $P_{(3)} \leq \alpha/(k-2)$, reject null associated with $P_{(3)}$ and proceed to next step.

$\vdots$

Strong Control Methods: Holm's Sequentially Rejective Bonferroni

- ***Holm's sequentially rejective Bonferroni method (Holm (1979)):***
  - Order p-values $P_{(1)} < P_{(2)} < \ldots < P_{(k)}$.
  - Compare $P_{(1)}$ to α/k.
    - If $P_{(1)} > \alpha/k$, stop and declare nothing significant.
    - if $P_{(1)} \leq \alpha/k$, reject null associated with $P_{(1)}$ and proceed to next step.
  - Compare $P_{(2)}$ to $\alpha/(k-1)$.
    - If $P_{(2)} > \alpha/(k-1)$, stop and declare nothing else significant.
    - If $P_{(2)} \leq \alpha/(k-1)$, reject null associated with $P_{(2)}$ and proceed to next step.
  - Compare $P_{(3)}$ to $\alpha/(k-2)$.
    - If $P_{(3)} > \alpha/(k-2)$, stop and declare nothing else significant.
    - If $P_{(3)} \leq \alpha/(k-2)$, reject null associated with $P_{(3)}$ and proceed to next step.

  $\vdots$

Strong Control Methods: Holm's Sequentially Rejective Bonferroni

- In a trial of hospitalized COVID-19 patients, suppose 3 endpoints were (1) death, (2) mechanical ventilation/death, and (3) WHO 8-point ordinal score.
- Suppose all considered primary and p-values are 0.042, 0.020, and 0.013, respectively.
- $p_{(1)} = 0.013 \leq 0.05/3$, so WHO 8-point scale declared significant.
- $p_{(2)} = 0.020 \leq 0.05/2$, so mechanical ventilation/death declared significant.
- $p_{(3)} = 0.042 \leq 0.05/1$, so mortality declared significant.
- With ordinary Bonferroni, only WHO 8-point scale declared significant.

Strong Control Methods: Holm's Sequentially Rejective Bonferroni

- Holm strongly controls FWER & more powerful than Bonferroni.

- Informal rationale for Holm: Once we reject null associated with $P_{(1)}$, we have either made a type 1 error or not.
  - If so, it doesn't matter what happens next because we already made ≥ 1 type 1 error.
  - If not, then null associated with $P_{(1)}$ was false, so there were at most $k-1$ true nulls. Can use Bonferroni with $\alpha/(k-1)$.

- Once we reject null associated with $P_{(2)}$ we have either made ≥ 1 type 1 error or not.
  - If so, it doesn't matter what happens next because we already made ≥ 1 type 1 error.
  - If not, then nulls associated with $P_{(1)}$ and $P_{(2)}$ were false, so there were at most $k-2$ true nulls. Can use Bonferroni with $\alpha/(k-2)$.

$\vdots$

Strong Control Methods: Holm's Sequentially Rejective Bonferroni

- Holm strongly controls FWER & more powerful than Bonferroni.

- Informal rationale for Holm: Once we reject null associated with $P_{(1)}$, we have either made a type 1 error or not.
  - If so, it doesn't matter what happens next because we already made ≥ 1 type 1 error.
  - If not, then null associated with $P_{(1)}$ was false, so there were at most $k-1$ true nulls. Can use Bonferroni with $\alpha/(k-1)$.

- Once we reject null associated with $P_{(2)}$ we have either made ≥ 1 type 1 error or not.
  - If so, it doesn't matter what happens next because we already made ≥ 1 type 1 error.
  - If not, then nulls associated with $P_{(1)}$ and $P_{(2)}$ were false, so there were at most $k-2$ true nulls. Can use Bonferroni with $\alpha/(k-2)$.

$\vdots$

Strong Control Methods: Holm's Sequentially Rejective Bonferroni

- Holm strongly controls FWER & more powerful than Bonferroni.

- Informal rationale for Holm: Once we reject null associated with $P_{(1)}$, we have either made a type 1 error or not.
  - If so, it doesn't matter what happens next because we already made ≥ 1 type 1 error.
  - If not, then null associated with $P_{(1)}$ was false, so there were at most $k-1$ true nulls. Can use Bonferroni with $\alpha/(k-1)$.

- Once we reject null associated with $P_{(2)}$ we have either made ≥ 1 type 1 error or not.
  - If so, it doesn't matter what happens next because we already made ≥ 1 type 1 error.
  - If not, then nulls associated with $P_{(1)}$ and $P_{(2)}$ were false, so there were at most $k-2$ true nulls. Can use Bonferroni with $\alpha/(k-2)$.

$\vdots$

Strong Control Methods: Holm's Sequentially Rejective Bonferroni

- Holm strongly controls FWER & more powerful than Bonferroni.

- Informal rationale for Holm: Once we reject null associated with $P_{(1)}$, we have either made a type 1 error or not.
  - If so, it doesn't matter what happens next because we already made ≥ 1 type 1 error.
  - If not, then null associated with $P_{(1)}$ was false, so there were at most $k-1$ true nulls. Can use Bonferroni with $\alpha/(k-1)$.

- Once we reject null associated with $P_{(2)}$ we have either made ≥ 1 type 1 error or not.
  - If so, it doesn't matter what happens next because we already made ≥ 1 type 1 error.
  - If not, then nulls associated with $P_{(1)}$ and $P_{(2)}$ were false, so there were at most $k-2$ true nulls. Can use Bonferroni with $\alpha/(k-2)$.

$\vdots$

Strong Control Methods: Holm's Sequentially Rejective Bonferroni

- Holm strongly controls FWER & more powerful than Bonferroni.

- Informal rationale for Holm: Once we reject null associated with $P_{(1)}$, we have either made a type 1 error or not.
  - If so, it doesn't matter what happens next because we already made ≥ 1 type 1 error.
  - If not, then null associated with $P_{(1)}$ was false, so there were at most $k - 1$ true nulls. Can use Bonferroni with $\alpha/(k - 1)$.

- Once we reject null associated with $P_{(2)}$ we have either made ≥ 1 type 1 error or not.
  - If so, it doesn't matter what happens next because we already made ≥ 1 type 1 error.
  - If not, then nulls associated with $P_{(1)}$ and $P_{(2)}$ were false, so there were at most $k - 2$ true nulls. Can use Bonferroni with $\alpha/(k - 2)$.

$\vdots$

Strong Control Methods: Graphical Alpha Transfer

- Same reasoning to tranfer alpha from significant comparisons to other comparisons (Bretz et al. (2011)).
  - Pre-specify initial alpha levels $\alpha_1, \ldots, \alpha_k$ with $\sum_i \alpha_i \leq \alpha$.
  - Pre-specify plan for transferring alpha from significant comparisons to other comparisons.
- Example: 3 endpoints: (1) Cardiovascular disease (CVD), (2) Coronary heart disease (CHD), (3) stroke.
  - CVD primary, so initial allocation of alpha: $(\alpha, 0, 0)$ to (CVD, CHD, stroke).
  - If CVD is significant, transfer all of its alpha to secondary endpoint CHD.
  - If reach CHD & it is significant, transfer all of its alpha to secondary endpoint stroke.

Strong Control Methods: Graphical Alpha Transfer

- Called a ***gatekeeping procedure***. Must pass through each gate to test subsequent hypothesis.

- Great if you are confident of passing through gates. E.g., comparing doses of drug to placebo.
  - Compare high dose to placebo at level α. Stop if not significant.
  - If significant, compare middle dose to placebo at same level α. Stop if not significant.
  - If significant, compare low dose to placebo at same level α.

- Downside: Must stop testing if fail to pass through gate.

- Another option in 3-endpoint example: If CVD significant, transfer half of alpha to each of CHD & stroke.

- If reach CVD/stroke & one is significant, transfer all of its alpha to other one.

Strong Control Methods: Graphical Alpha Transfer

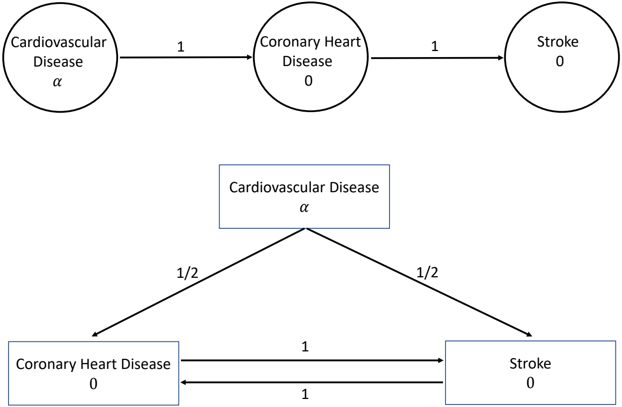


Figure: Two alpha transfer options for primary endpoint (cardiovascular disease) and two secondary endpoints (coronary heart disease, stroke).

Strong Control Methods: Graphical Alpha Transfer

- Can represent with graph: Initial alpha allocation inside nodes, proportion of alpha transferred shown on arrows.
- Top option: Gatekeeping procedure.
- Bottom option with $\alpha = 0.05$: If CVD not significant at $\alpha = 0.05$, stop. If CVD significant at $\alpha = 0.05$, test both secondary endpoints at $\alpha = 0.05/2 = 0.025$.
- If either secondary is significant at 0.025, transfer all of 0.025 to other secondary.
- I.e., if primary and one secondary significant, last endpoint is tested at $0.025 + 0.025 = 0.05$.

Strong Control Methods: Graphical Alpha Transfer

- Bretz et al. (2011) show that graphical alpha transfer method strongly controls FWER.
- Another advantage of method: Graph succinctly summarizes what would take pages of text to explain, and the text would be harder to understand!
- Order doesn't matter provided you continue until no longer able to transfer any alpha.
- Bretz et al. (2011) give efficient way to carry out procedure.

Strong Control Methods: The Closure Principle

- Marcus et al. (1976).
- Testing finite family $\mathscr{F}$ of all intersections of $\{H_{0i},\ i = 1, \ldots, n\}$.
- Suppose that for each subset $P \subset \{1, \ldots, n\}$, there is an α level test of $H_P = \cap_{i \in P} H_i$.
- ***Closure principle:*** Reject H_P if and only if the test of H_R is significant at $\alpha = 0.05$ for each H_R that implies H_P (i.e., for each R such that $R \supset P$).

Theorem

Following the closure principle strongly controls the FWER among the family $\mathscr{F}$.

Strong Control Methods: The Closure Principle

- Very powerful, but sometimes misunderstood.
- E.g., suppose compare 4 means, $H_{ij} : \mu_i = \mu_j$.
- ***Newman-Keuls procedure:*** Reject H_{12} if and only if T_{12}, F_{123}, F_{124}, and F_{1234} are all significant at level α, where T and F denote t and F statistics. Similarly for other pairwise comparisons.
- Does Newman-Keuls follow closure principle?
- No! Closure principle also requires an α level test of $H_{12} \cap H_{34}$.
- In fact, Newman-Keuls does not strongly control FWER for $k = 4$.

Strong Control Methods: The Closure Principle

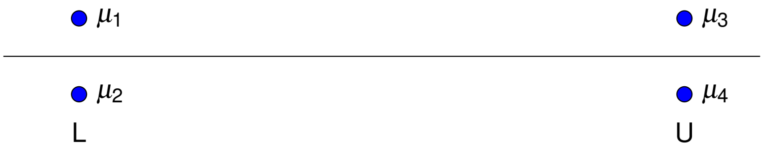


Figure: Configuration of 4 population means that inflates the FWER for Newman-Keuls: $\mu_1 = \mu_2 = L$ and $\mu_3 = \mu_4 = U$, where $U - L$ is huge.

- F_{123}, F_{124}, F_{134}, F_{234}, and F_{1234} nearly guaranteed to be statistically significant, so FWER of Newman Keuls $\approx$ FWER of rejecting if either T_{12} or T_{34} is significant at level α.

$$\begin{aligned} FWER &\approx P(T_{12} \text{ signif. } \cup T_{34} \text{ signif.}) \\ &= 1 - P(\text{neither signif.}) = 1 - (1 - 0.05)^2 \\ &= 0.0975. \end{aligned} \tag{3}$$

- Doesn't violate closure principle: To truly follow closure, must have α-level test of $H_{12\cap 34}$.

Outline

Introduction

Strong/Weak Control of FWER

Methods for Strong Control of FWER

- Bonferroni
- Holm's Sequentially Rejective Bonferroni
- Graphical Alpha Transfer
- The Closure Principle

Multiple Arms

- Dunnett and Variants

Multiple Subgroups

- Forest Plots
- Permutation Methods
- Hochberg

False Discovery Rate (FDR)

Multiple Arms: Dunnett and Variants

- Consider 1-sided problem of comparing k arms to same control on continuous endpoint Y.

- $H_i : \mu_i = \mu_0$, μ_0 and μ_i are means in control and arm i.

- Global null is $H = \cap H_i : \ \mu_0 = \mu_1 = \ldots = \mu_k$.

- Poor option: Bonferroni.
  - Reject hypothesis associated with $p_{(i)}$ if $p_{(i)} \leq \alpha/k$.

- Better option: Holm sequentially rejective Bonferroni using p-value thresholds of $\alpha/k, \alpha/(k-1), \ldots, \alpha/1$ for $p_{(1)}, \ldots, p_{(k)}$.

- Better still: ***Dunnett procedure*** Dunnett (1955): Find actual joint distribution of test statistics using clever trick.

Multiple Arms: Dunnett and Variants

- Assume Y normally distributed with same variance σ^2 in different arms, and for simplicity, assume n large enough to treat σ^2 as known.
- $Z_i = \frac{\bar{Y}_i - \bar{Y}_0}{\sqrt{2\sigma^2/n}}, \ i = 1, \ldots, k.$
- To compute FWER, assume without loss of generality that $Y_i \sim N(0,1)$, $i = 1, \ldots, k$.
- Z_i are dependent only because of shared control. Condition on control sample mean to remove dependence.

Multiple Arms: Dunnett and Variants

Given $\bar{Y}_0 = y_0$,

$$Z_1 = \frac{\bar{Y}_1 - y_0}{\sqrt{2/n}},\ Z_2 = \frac{\bar{Y}_2 - y_0}{\sqrt{2/n}}, \ldots,\ Z_k = \frac{\bar{Y}_k - y_0}{\sqrt{2/n}} \text{ are iid.}$$

Let $M = \max(Z_1, \ldots, Z_k)$. Then

$$\begin{aligned} P(M \le c_k \mid \bar{Y}_0 = y_0) &= P\left\{\max\left(\sqrt{n}\,\bar{Y}_1, \ldots, \sqrt{n}\,\bar{Y}_k\right) \le \sqrt{2}\,c_k + \sqrt{n}\,y_0\right\} \\ &= \left\{\Phi\left(\sqrt{2}\,c_k + \sqrt{n}\,y_0\right)\right\}^k. \end{aligned}$$

Multiple Arms: Dunnett and Variants

- Now integrate over density, $f_n(y_0)$, of $\bar{Y}_0$, which is normal with mean 0 and variance $1/n$.

$$P(M \le c_k) \quad = \quad \int_{-\infty}^{\infty} \left\{ \Phi\left(\sqrt{2}\, c_k + \sqrt{n}\, y_0 \right) \right\}^k f_n(y_0) dy_0. \qquad (4)$$

- Make substitution $z_0 = \sqrt{n}\, y_0$. Because $Z_0 = \sqrt{n}\, \bar{Y}_0 \sim N(0,1)$, get

$$P(M \le c_k) = \int_{-\infty}^{\infty} \left\{ \Phi\left(\sqrt{2}\, c_k + z_0 \right) \right\}^k \phi(z_0) dz_0,$$

  where $\phi(z_0)$ is standard normal density.

- Equate to $1 - \alpha$ and solve for c_k.

Multiple Arms: Dunnett and Variants

Table: Dunnett critical values c_k for 1-sided test at $\alpha = 0.025$ when σ^2 is known.

k	Z-score boundary
1	1.960
2	2.212
3	2.349
4	2.442
5	2.511
6	2.567

- Substitute pooled variance $\hat{\sigma}^2$ for σ^2 in Z_i.
- When n is small, must also integrate over distribution of $\hat{\sigma}^2$ to find c_k.

Multiple Arms: Dunnett and Variants

- Even better option: Use sequentially rejective version based on closure principle.
  - Compare $Z_{(k)}$ to c_k.
    - If $Z_{(k)} < c_k$, stop.
    - If $Z_{(k)} \geq c_k$, reject hypothesis associated with $Z_{(k)}$ and proceed to next step.
  - Compare $Z_{(k-1)}$ to c_{k-1}.
    - If $Z_{(k-1)} < c_{k-1}$, stop.
    - If $Z_{(k-1)} \geq c_{k-1}$, reject hypothesis associated with $Z_{(k-1)}$ and proceed to next step.
  - Compare $Z_{(k-2)}$ to c_{k-2}.
    - If $Z_{(k-2)} < c_{k-2}$, stop.
    - If $Z_{(k-2)} \geq c_{k-2}$, proceed to next step.

$\vdots$

Outline

- Introduction
- Strong/Weak Control of FWER
- Methods for Strong Control of FWER
    - Bonferroni
    - Holm’s Sequentially Rejective Bonferroni
    - Graphical Alpha Transfer
    - The Closure Principle
- Multiple Arms
    - Dunnett and Variants
- Multiple Subgroups
    - Forest Plots
    - Permutation Methods
    - Hochberg
- False Discovery Rate (FDR)

Multiple Subgroups

- Subgroups difficult to deal with because of their sheer number:
  - Demographics: race, sex, age, etc.
  - Prognostic characteristics: prior heart attack or family history in heart disease trial, risky behaviors in HIV vaccine trial, etc.

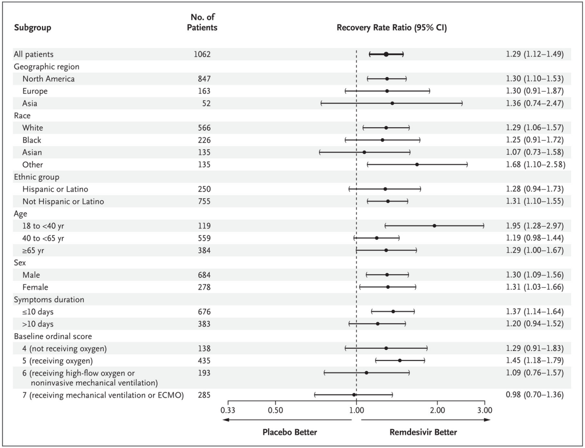
- Two opposing problems: (1) Multiplicity leads to false positives and (2) Low power to detect interactions, exacerbated by adjusting for multiple comparisons!

- Typical approaches either don't control, or weakly control, FWER.

- E.g., for each factor, first test (factor)$\times$(treatment) interaction. Only if interaction is significant, test treatment effect separately within subgroups.

- Accompanying plot: ***Forest plot***.

Multiple Subgroups: Forest Plots



Subgroup	No. of Patients	Recovery Rate Ratio (95% CI)
All patients	1062	1.29 (1.12–1.49)
Geographic region		
North America	847	1.30 (1.10–1.53)
Europe	163	1.30 (0.91–1.87)
Asia	52	1.36 (0.74–2.47)
Race		
White	566	1.29 (1.06–1.57)
Black	226	1.25 (0.91–1.72)
Asian	135	1.07 (0.73–1.58)
Other	135	1.68 (1.10–2.58)
Ethnic group		
Hispanic or Latino	250	1.28 (0.94–1.73)
Not Hispanic or Latino	755	1.31 (1.10–1.55)
Age		
18 to <40 yr	119	1.95 (1.28–2.97)
40 to <65 yr	559	1.19 (0.98–1.44)
≥65 yr	384	1.29 (1.00–1.67)
Sex		
Male	684	1.30 (1.09–1.56)
Female	278	1.31 (1.03–1.66)
Symptoms duration		
≤10 days	676	1.37 (1.14–1.64)
>10 days	383	1.20 (0.94–1.52)
Baseline ordinal score		
4 (not receiving oxygen)	138	1.29 (0.91–1.83)
5 (receiving oxygen)	435	1.45 (1.18–1.79)
6 (receiving high-flow oxygen or noninvasive mechanical ventilation)	193	1.09 (0.76–1.57)
7 (receiving mechanical ventilation or ECMO)	285	0.98 (0.70–1.36)

NEJM 2020;
383:1813-26

Figure: Forest plot from the ACTT-1 clinical trial of Remdesivir for treatment of hospitalized COVID-19.

Multiple Subgroups: Forest Plots

- Forest plots often show good agreement of treatment effect across subgroups.
- Problem: Subgroups based on prognostic factors can have big overlap.
  - E.g., in HIV, big overlap of patients in worst viral load and worst CD4 count subgroups.
- How do we account for overlap?
- There are bootstrap (Rosenkranz (2014)) and permutation (Dane et al. (2019); Lipkovich et al. (2011)) methods. We focus on one permutation method.

Multiple Subgroups: SEAMOS

- ***Standardized Effects Adjusted for Multiple Subgroups (SEAMOS)*** (Dane et al. (2019)).
  - Compute standardized difference between effect in category *j* of factor *i* and overall effect that ignores subgroups, then take max:

$$Z_{ij} = \frac{\hat{\delta}_{ij} - \hat{\delta}}{\mathrm{se}(\hat{\delta}_{ij})}, \quad M = \max_{ij}(Z_{ij}).$$

  - Fix outcome & treatment vectors, randomly interchange covariate vectors across people, compute *M* for covariate-interchanged data.
  - Repeat many times to get null reference distribution of *M* and compute 2.5 and 97.5 percentiles. Declare significant any subgroup effects outside these percentiles.

- SEAMOS preserves overall treatment effect and dependence of covariates.

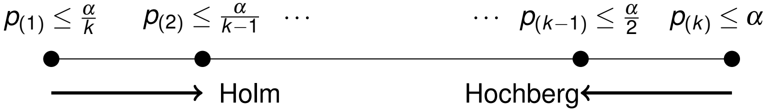
Multiple Endpoints: Hochberg*

- Hochberg (1988) is used to test whether effect in any subgroup.
- Holm's procedure started with most significant (smallest p-value) result and proceded to less significant results.
- Can also go in reverse direction.
- Start with largest p-value, $p_{(k)}$ and compare to α.
  - If $p_{(k)} \leq \alpha$, reject ALL hypotheses and stop.
  - If $p_{(k)} > \alpha$, do not reject hypothesis associated with $p_{(k)}$, but move to next step.
- Compare $p_{(k-1)}$ to $\alpha/2$.
  - If $p_{(k-1)} \leq \alpha/2$, reject hypotheses associated with $p_{(k-1)}, \ldots, p_{(1)}$.
  - If $p_{(k-1)} > \alpha/2$, do not reject hypothesis associated with $p_{(k-1)}$, but move to next step.
- Compare $p_{(k-2)}$ to $\alpha/3 \ldots$

*Not guaranteed to strongly control FWER without further conditions

Multiple Endpoints: Hochberg*

- Called ***Hochberg*** procedure.
- P-value thresholds for Holm and Hochberg are same, but Holm starts with most significant, while Hochberg starts with least significant.



$p_{(1)} \leq \frac{\alpha}{k}$ $\quad p_{(2)} \leq \frac{\alpha}{k-1}$ $\quad \cdots \quad \cdots$ $\quad p_{(k-1)} \leq \frac{\alpha}{2}$ $\quad p_{(k)} \leq \alpha$

Holm

Hochberg

Figure: Holm and Hochberg procedures.

*Not guaranteed to strongly control FWER without further conditions

Multiple Endpoints: Hochberg*

- Hochberg always more powerful than Holm, which is more powerful than Bonferroni.
- Problem: Hochberg does not always strongly control FWER.
- However, Hochberg DOES strongly control FWER when test statistics independent or have certain types of positive dependence.

*Not guaranteed to strongly control FWER without further conditions

Multiple Endpoints: Hochberg*

- Hochberg is attractive for multiple positively correlated endpoints.

- Example: COVID-19 trial with primary endpoints 1. mechanical ventilation or death, 2. death.

- Declare benefit on at least one endpoint if
  - smaller p-value ≤ 0.25 for Bonferroni/Holm,
  - smaller p-value ≤ 0.025 **or** both p-values ≤ 0.05 for Hochberg.

- If the two endpoints are positively correlated, then FWER is strongly controlled.

*Not guaranteed to strongly control FWER without further conditions

Outline

Introduction

Strong/Weak Control of FWER

Methods for Strong Control of FWER

- Bonferroni
- Holm's Sequentially Rejective Bonferroni
- Graphical Alpha Transfer
- The Closure Principle

Multiple Arms

- Dunnett and Variants

Multiple Subgroups

- Forest Plots
- Permutation Methods
- Hochberg

False Discovery Rate (FDR)

False Discovery Rate (FDR)

- Another error rate is ***false discovery rate***.

Table:

	Declared not significant	Declared significant	
H_0 True	T	V	m_0
H_0 Untrue	U	W	$m-m_0$
	$m-X$	X	m

- $Q = V/X$ (defined as 0 if $X = 0$)
- FDR$=\mathrm{E}(Q)$. Expected proportion of true nulls among rejections.
- FDR$\leq$FWER. Why?

False Discovery Rate (FDR)

Table:

	Declared not significant	Declared significant	
H_0 True	T	V	m_0
H_0 Untrue	U	W	$m-m_0$
	$m-X$	X	m

- Claim: $I(V\geq 1)\geq Q=V/X$ because:
  - If $V=0$, then $I(V\geq 1)=0$ and $Q=V/X=0$ (even if $X=0$).
  - If $V\geq 1$, then $I(V\geq 1)=1$ and $Q=V/X\leq 1$.
- Because $I(V\geq 1)\geq Q$, FWER=$\mathrm{E}\{I(V\geq 1)\}\geq \mathrm{E}(Q)=\mathrm{FDR}$.

False Discovery Rate (FDR)

- Thus, FDR is less stringent than FWER.
- Can have FDR$\leq \alpha$ and FWER$> \alpha$ but not vice-versa.
- FDR control is reasonable when # comparisons very large.
- ***Benjamini-Hochberg procedure*** (Benjamini and Hochberg (1995)).
- Reject hypotheses associated with $p_{(1)}, \ldots, p_{(j)}$ at level q^* if $p_{(j)} \leq (j/k)q^*$.
- If test statistics are independent, FDR controls FDR at level q^*.

False Discovery Rate (FDR)

- Note that under global null, if test statistics independent and have continuous distribution function, $\mathrm{E}(P_{(j)}) = \mathrm{E}(U_{(j)})$ $= j/(k+1)$, where U_i are iid uniforms,

- In that case, Benjamini-Hochberg rejects hypotheses associated with $p_{(1)}, \ldots, p_{(j)}$ at level q^* if

$$\frac{p_{(j)}}{\mathrm{E}(P_{(j)})} \leq \left(\frac{k+1}{k}\right) q^* \approx q^* \text{ if } k \text{ large.}$$

- In terms of stringency,

$$\text{Benjamin-Hochberg} \leq \text{Hochberg} \leq \text{Holm} \leq \text{Bonferroni.}$$

- FDR usually reserved for large # comparisons (e.g., gene association studies).

Summary

- Multiple comparisons in clinical trials caused by multiple arms, endpoints, analyses, subgroups, time points (monitoring), etc.
- P(at least one type 1 error)=FWER. Can control it:
  - Weakly: Only under global null.
  - Strongly: Regardless of which nulls are true.
- Bretz et al. (2011) graphical method of transferring alpha from rejected hypotheses to other hypotheses is very powerful. Several techniques are special cases, including
  - Gatekeeping procedures.
  - Sequentially rejective Bonferroni.
- Closure principle also useful for proving strong control of FWER.
- With huge number of comparisons, may want to control FDR (less conservative) instead of FWER.

References I

Bar-Hillel, M., Bar-Natan, D., Kalai, G., and McKay, B. (1999). Solving the Bible Code Puzzle. Statistical Science **14,** 150 – 173.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal statistical society: series B (Methodological) **57,** 289–300.

Bretz, F., Posch, M., Glimm, E., Klinglmueller, F., Maurer, W., and Rohmeyer, K. (2011). Graphical approaches for multiple comparison procedures using weighted bonferroni, simes, or parametric tests. Biometrical Journal. Biometrische Zeitschrift **53,** 894 – 913.

Dane, A., Spencer, A., Rosenkranz, G., Lipkovich, I., and Parke, T. (2019). Subgroup analysis and interpretation for phase 3 confirmatory trials: White paper of the efspi/psi working group on subgroup analysis. Pharmaceutical Statistics **18,** 126–139.

Drosnin, M. (1997). The bible code. Touchstone (Simon & Schuster).

References II

Dunnett, C. W. (1955). A multiple comparison procedure for comparing several treatments with a control. Journal of the American Statistical Association **50,** 1096–1121.

History-Channel (2003). The Bible Code: Predicting Armageddon. A&E Television Networks.

Hochberg, Y. (1988). A sharper bonferroni procedure for multiple tests of significance. Biometrika **75,** 800–802.

Hochberg, Y. and Tamhane, A. C. (1987). Multiple comparison procedures. John Wiley & Sons, Inc.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. Scandinavian journal of statistics pages 65–70.

ISIS-2 (1988). Randomised trial of intraveneous streptokinase, oral aspirin, both, or neither among 17,187 cases of suspected acute myocardial infarction: Isis-2. The Lancet **332,** 349–360.

References III

Lipkovich, I., Dmitrienko, A., Denne, J., and Enas, G. (2011). Subgroup identification based on differential effect search—a recursive partitioning method for establishing response to treatment in patient subpopulations. Statistics in medicine **30,** 2601–2621.

Marcus, R., Eric, P., and Gabriel, K. R. (1976). On closed testing procedures with special reference to ordered analysis of variance. Biometrika **63,** 655–660.

Proschan, M. A. and Brittain, E. H. (2020). A primer on strong vs weak control of familywise error rate. Statistics in Medicine **39,** 1407–1413.

Rosenkranz, G. K. (2014). Bootstrap corrections of treatment effect estimates following selection. Computational Statistics & Data Analysis **69,** 220–227.

Witztum, D., Rips, E., and Rosenberg, Y. (1994). Equidistant letter sequences in the book of genesis. Statistical Science **9,** 429 – 438.

Lecture 8: Adaptive Procedures

Introduction

Blinded Sample Size Methods

- Binary Outcomes
- Continuous Outcomes

Unblinded Sample Size Methods

Unplanned Changes

Outline

Introduction

Blinded Sample Size Methods

- Binary Outcomes
- Continuous Outcomes

Unblinded Sample Size Methods

Unplanned Changes

Introduction

- WARNING: This is a brief overview of a vast topic! Only the main ideas are provided. See chapter 11 of Proschan (2021).
- Clinical trial planning requires information such as:
    - The control event rate (for binary outcome) or variance (for continuous outcome).
    - The treatment effect.
    - Whether data will be skewed, etc.
- This information is lacking in new disease.
- What do we do?
- Adaptive methods: **Pre-specified** procedures to use within-trial data to change design.

Introduction

- Examples:
  - Sample size re-estimation based on lumped event rate (binary outcome trial) or lumped variance (continuous endpoint trial).
  - Examine lumped data to detect outliers. If none, use t-test; if outliers, use Wilcoxon rank sum test.
  - Sample size re-estimation based on treatment effect.
  - Change of primary endpoint after examining treatment effect.
- First two are very different from last two: They don't require unblinding.
- Unblinding is undesirable because could lead to:
  - Background treatment bias, selection bias, etc.
  - Problems with regulators.
  - Inflation of type 1 error rate if not careful.

Outline

Introduction

Blinded Sample Size Methods

- Binary Outcomes
- Continuous Outcomes

Unblinded Sample Size Methods

Unplanned Changes

Blinded Sample Size Methods

- We begin with binary outcomes and safest adaptations–blinded ones (Gould (1992)).
- In trials with binary endpoint like 28-day mortality, we often over-estimate event rate. Why?
- Event rate estimates may be based on observational data. People volunteering for clinical trials:
  - May be more health-conscious.
  - May get better care in clinical trial.
  - Must satisfy entry criteria that could exclude sickest patients.
  - May undergo run-in to weed out non-adherers.
- If power is based on a given **relative effect** (e.g., 25% reduction in 28-day mortality), overestimating event rate means underpowering trial.

Blinded Sample Size Methods: Binary Outcomes

- Want specified power to detect **fixed** relative risk $R < 1$.
- Procedure:
  - Calculate initial sample size n_0/arm based on fixed relative risk $R < 1$ and pre-trial estimate of control probability p_C.
  - At planned halfway point, $n_1 = n_0/2$ per arm (or another fraction), calculate overall event rate $\hat{p}_1 = (\#$ with event$)/(\#$evaluated$)$.
  - Equate

$$\begin{aligned} (p_T + p_C)/2 &= \hat{p}_1 \\ p_T/p_C &= R, \end{aligned} \tag{1}$$

and solve for p_T and p_C.

$$p_C = \frac{2\hat{p}_1}{R+1}, \;\; p_T = \frac{2R\hat{p}_1}{R+1}. \tag{2}$$

Blinded Sample Size Methods: Binary Outcomes

- Example: COVID-19 trial of hospitalized patients, usual care+new drug versus usual care+placebo.

- Primary endpoint: ventilation/death by day 60.

- Initial estimate of control probability: $p_C = 0.20$.

- Want 85% power to detect 25% reduction ($R = 1 - 0.25 = 0.75$) using 2-tailed test at $\alpha = 0.05$.

- Initial sample size: $n_0 = 1,036$/arm.

- After $n_1 = 500$/arm, 150 have events (combined across arms), so proportion with events is $\hat{p}_1 = 150/1,000 = 0.15$.

Blinded Sample Size Methods: Binary Outcomes

- Compute

$$p_C = \frac{2\hat{p}_1}{R+1} = \frac{2(0.15)}{0.75+1} = 0.1714$$

$$p_T = \frac{2R\hat{p}_1}{R+1} = \frac{2(0.75)(0.15)}{0.75+1} = 0.1286.$$

- New sample size based on $p_C = 0.1714$ and $p_T = 0.1286$: $n = 1249$/arm.
- **Increase sample size and use z-test of proportions at end as if sample size were fixed in advance.**
- Could also stratify analysis by before/after sample size change (always good to compare results pre- and post-adaptation).

Blinded Sample Size Methods: Binary Outcomes

- One problem with this procedure: Overall event rate could be low for at least 2 reasons:
  1. Control event rate is lower than expected.
  2. Treatment effect is much higher than expected.
- If second reason is true, then don't need to increase sample size!
- But very large treatment effects are unusual in clinical trials.
- Why not look at control event rate instead of overall rate?
- Problem 1 with peeking at control event rate: Even blinded investigators may know overall event rate.
  - knowledge of control event rate plus overall rate reveals the observed treatment effect!

Blinded Sample Size Methods: Binary Outcomes

- Problem 2 with peeking at control event rate: Even if investigators don't know overall event rate, control event rate gives some information about treatment effect because control event rate is correlated with observed treatment effect:

$$\begin{aligned}
\mathrm{cov}(\hat{p}_{C1}, \hat{p}_{T_1} - \hat{p}_{C1}) &= \mathrm{cov}(\hat{p}_{C_1}, \hat{p}_{T1}) – \mathrm{cov}(\hat{p}_{C1}, \hat{p}_{C1}) \\
&= 0 - \mathrm{var}(\hat{p}_{C1}) \\
&= -p_C(1-p_C)/n_1 . \qquad (3)
\end{aligned}$$

- Dividing by product of standard deviations gives correlation $\rho \approx -0.71$ under the null hypothesis.

- Nonzero correlation implies that the type 1 error rate could be inflated by procedure that allows peeking at control event rate.

Blinded Sample Size Methods: Binary Outcomes

- Problem 2 with peeking at control event rate: Even if investigators don't know overall event rate, control event rate gives some information about treatment effect because control event rate is correlated with observed treatment effect:

$$\begin{aligned} \mathrm{cov}(\hat{p}_{C1}, \hat{p}_{T_1} - \hat{p}_{C1}) &= \mathrm{cov}(\hat{p}_{C_1}, \hat{p}_{T1}) – \mathrm{cov}(\hat{p}_{C1}, \hat{p}_{C1}) \\ &= 0 - \mathrm{var}(\hat{p}_{C1}) \\ &= -p_C(1-p_C)/n_1 . \end{aligned} \tag{3}$$

- Dividing by product of standard deviations gives correlation $\rho \approx -0.71$ under the null hypothesis.

- Nonzero correlation implies that the type 1 error rate could be inflated by procedure that allows peeking at control event rate.

Blinded Sample Size Methods: Binary Outcomes

- Problem 2 with peeking at control event rate: Even if investigators don't know overall event rate, control event rate gives some information about treatment effect because control event rate is correlated with observed treatment effect:

$$\begin{aligned} \mathrm{cov}(\hat{p}_{C1}, \hat{p}_{T_1} - \hat{p}_{C1}) &= \mathrm{cov}(\hat{p}_{C_1}, \hat{p}_{T1}) - \mathrm{cov}(\hat{p}_{C1}, \hat{p}_{C1}) \\ &= 0 - \mathrm{var}(\hat{p}_{C1}) \\ &= -p_C(1 - p_C)/n_1. \end{aligned} \tag{3}$$

- Dividing by product of standard deviations gives correlation $\rho \approx -0.71$ under the null hypothesis.

- Nonzero correlation implies that the type 1 error rate could be inflated by procedure that allows peeking at control event rate.

Blinded Sample Size Methods: Binary Outcomes

- Not a problem with blinded method because:

$$\begin{aligned}
&\operatorname{cov}\left\{(\hat{p}_{T1}+\hat{p}_{C1})/2, \hat{p}_{T_1}-\hat{p}_{C1}\right\} \\
&= (1/2)\left\{\operatorname{cov}(\hat{p}_{T_1}, \hat{p}_{T1}) - \operatorname{cov}(\hat{p}_{T1}, \hat{p}_{C1}) + \operatorname{cov}(\hat{p}_{C1}, \hat{p}_{T1}) - \operatorname{cov}(\hat{p}_{C1}, \hat{p}_{C1})\right\} \\
&= (1/2)\{\operatorname{var}(\hat{p}_{T1}) - 0 + 0 - \operatorname{var}(\hat{p}_{C1})\} \\
&= (1/2)\{p_T(1-p_T)/n_1 - p_C(1-p_C)/n_1\} \\
&= (1/2)\{p(1-p)/n_1 - p(1-p)/n_1\} \quad \text{(under } H_0\text{)} \\
&= 0. \qquad (4)
\end{aligned}$$

Blinded Sample Size Methods: Continuous Outcomes

- Same ideas apply to trials with continuous outcomes analyzed using unpaired t-test. Gould and Shih (1992).
- Sample size depends on treatment effect and variance σ^2.
- **Blinded method: Keep treatment effect fixed.**
  - **Use pre-trial estimate of σ^2 for initial sample size n_0/per arm.**
  - **At planned halfway point, $n_1 = n_0/2$/arm, re-estimate σ^2 using variance of combined data across arms.**
  - **Modify sample size and treat as fixed in final analysis (use ordinary t-test at end).**
  - **Alternatively, could use t-test stratified by stage.**
- Similar issue as with binary outcomes: Blinded variance could be large because treatment effect is huge.

Blinded Sample Size Methods: Continuous Outcomes

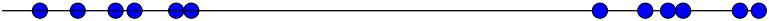


Figure: Lumped variance is inflated when there is large treatment effect.

- Still, unless treatment effect is huge, the problem is minor.
- E.g., if σ^2 is within-arm variance and $E = \delta/\sigma$ is treatment effect relative to standard deviation, then $\mathrm{var}(Y) = \sigma^2(1 + E^2/4)$.
- Even large treatment effect of $E = 1/2$ results in only 6% inflation of variance, 3% inflation of standard deviation.

Outline

Introduction

Blinded Sample Size Methods

- Binary Outcomes
- Continuous Outcomes

Unblinded Sample Size Methods

Unplanned Changes

Unblinded Sample Size Methods

- Now consider unblinded sample size methods: **Much riskier** and error rate can be inflated if not careful.

- Example: Suppose original sample size is $1,000$/arm and we peek at z-score after 10/arm.
  - If $Z_{10} > 1.96$, change final sample size to 10/arm & reject H_0.
  - If $Z_{10} < 1.96$, continue to $1,000$/arm and reject if $Z_{1,000} \geq 1.96$.

- Z_{10} and $Z_{1,000}$ are nearly independent, so type 1 error rate, $P\{Z_{10} \geq 1.645 \cup Z_{1,000} \geq 1.645\}$, is

$$\begin{aligned} &= \quad 1 - P\{Z_{10} < 1.645 \cap Z_{1,000} < 1.645\} \\ &\approx \quad 1 - (1 - 0.05)^2 = 0.0975. \end{aligned} \tag{5}$$

- Proschan and Hunsberger (1995) showed that the most nefarious method more than doubles the type 1 error rate.

Unblinded Sample Size Methods

- How can we protect the type 1 error rate?
- Assume n is large enough to treat nuisance parameters (like σ or p) as known.
- Again specify initial sample size n_0/arm and assume sample size reassessment is after $n_1 = n_0/2$/arm (fixed number).
- Let Z_1 be z-score at stage 1 with n_1/arm.
- Choose $n_2 = n_2(Z_1)$ (random, as is $n = n_1 + n_2$) based on Z_1. We give more details on choice of n_2 later.

Unblinded Sample Size Methods

- If make no change in sample size ($n_2 = n_0/2$), then usual, fixed sample size z-score using all data is

$$Z = \frac{Z_1 + Z_2}{\sqrt{2}}. \tag{6}$$

- Key to adaptive method: Even if we change n_2, still **equally weight** Z_1 and Z_2, as in (6), and reject if Z in (6) is $\geq z_\alpha$.

- Under H_0, $Z_1 \sim N(0,1)$ and $Z_2 \,|\, Z_1 \sim N(0,1)$.

- Because distribution of $Z_2 \,|\, Z_1$ does not depend on Z_1, (Z_1, Z_2) are iid $N(0,1)$ under H_0, same as in fixed sample size setting.

- **More generally, any level α rejection region $(Z_1, Z_2) \in R$ when n_2 is fixed remains level α in adaptive n_2 setting.**

Unblinded Sample Size Methods

- Weakness of adaptive method is clear: Equally weighting z-scores from stage 1 and 2 even though n_2 is much larger (or smaller) than n_1 is inefficient.
- Reasonable if n_2 doesn't change much, but not if it does!
- Can even reject H_0 when conventionally computed z-score is negative!
  - Proschan and Hunsberger (1995) showed this.
  - Burman and Sonesson (2006) "rediscovered" this fact.
- Other adaptive procedures have similar drawbacks.

Unblinded Sample Size Methods

- Could combine 1-sided p-values, $P_i = 1 - \Phi(Z_i)$, $i = 1, 2$ instead of z-scores.
- Called a ***p-value combination function***.
- **Any level α rejection region $(P_1, P_2) \in R$ when n_2 is fixed remains level α in adaptive n_2 setting.**
- **Z-score and p-value combination procedures are equivalent:**
  - **Any z-score combination procedure corresponds to some p-value combination function procedure and vice versa.**
- One such p-value combination function: Fisher's method of combining p-values:

$$-2\ln(P_1 P_2) \sim \chi^2_4 \text{ under } H_0. \qquad (7)$$

Unblinded Sample Size Methods

- Bauer and Köhne (1994) method:
    - After observing first stage p-value, change sample size but continue to use (7) at end.
    - Reject H_0 if $-2\ln(P_1P_2) \geq \chi^2_\alpha$, the $(1-\alpha)$th quantile of a chi-squared distribution with 4 degrees of freedom.
    - They also modified procedure to allow stopping at first stage for futility/benefit.

- Note: If $-2\ln(P_1) \geq \chi^2_4(\alpha)$, the $(1-\alpha)$th quantile of a χ^2_4 distribution, no need for second stage.

- Procedure has same drawback as equally weighting z-scores: Giving equal weight to p-values is inefficient if one p-value is based on much more information than the other.

Unblinded Sample Size Methods

- Another equivalent way to formulate adaptive methods, the ***conditional error function*** approach of Proschan and Hunsberger (1995), suggests how to modify sample size.

- We motivate the approach using z-score combination function $Z = (Z_1 + Z_2)/\sqrt{2}$. Given $Z_1 = z_1$, reject H_0 if $\frac{z_1 + Z_2}{\sqrt{2}} \geq z_\alpha$

$$\begin{aligned} &\Leftrightarrow \quad Z_2 \geq z_\alpha \sqrt{2} - z_1 \\ &\Leftrightarrow \quad 1 - \Phi(Z_2) \leq 1 - \Phi(z_\alpha \sqrt{2} - z_1) \\ &\Leftrightarrow \quad P_2 \leq A(z_1), \end{aligned} \tag{8}$$

where P_2 is second stage p-value, $1 - \Phi(Z_2)$, and

$$A(z_1) = 1 - \Phi\left(z_\alpha \sqrt{2} - z_1\right).$$

Unblinded Sample Size Methods

- **Can view conditional error function approach as follows:**
  - **After seeing $Z_1 = z_1$, do a test using second stage z-score Z_2 only, but use alpha level $A(z_1)$.**

- $A(z_1)$ is called a ***conditional error function*** because it is the conditional type 1 error rate given $Z_1 = z_1$ (conditional power under H_0).

- That is,

$$A(z_1) = P_0\left(\frac{Z_1 + Z_2}{\sqrt{2}} \geq z_\alpha \,\middle|\, Z_1 = z_1\right),$$

  where P_0 denotes probability computed under H_0.

- Of course, $0 \leq A(z_1) \leq 1$.

Unblinded Sample Size Methods

- $A(z_1)$ satisfies

$$\begin{aligned}\int A(z_1)\phi(z_1)dz_1 &= \int P_0\left(\frac{Z_1+Z_2}{\sqrt{2}} \geq z_\alpha \,\middle|\, Z_1 = z_1\right)\phi(z_1)dz_1 \\ &= P_0\left(\frac{Z_1+Z_2}{\sqrt{2}} \geq z_\alpha\right) = \alpha, \qquad (9)\end{aligned}$$

  where $\phi(z_1)$ is standard normal density.

- A different z-score combination results in a different $A(z_1)$ that satisfies (9), i.e., integrates to α.
  - E.g., Proschan and Hunsberger (1995) considered the “circular” conditional error function based on $Z_1^2 + Z_2^2 \geq c$.

Unblinded Sample Size Methods

- Approaches using conditional error functions, z-score combinations, and p-value combinations are equivalent, so why is conditional error formulation helpful?

- Because we can choose n by considering conditional power under an alternative hypothesis.

- At end of first stage, alpha level $A = A(z_1)$ to use in second stage is fixed.

- To achieve (conditional) power $1 - \beta$, use EZ principle: Equate

  $$\mathrm{E}(Z_2) = z_A + z_\beta,$$

  and solve for n_2.

Unblinded Sample Size Methods

- For example, in t-test setting, if $A(z_1) = 0.15$,

$$\mathrm{E}(Z_2) = \frac{\delta}{\sqrt{2\sigma^2/n_2}}.$$

- Estimate σ^2 and δ. Suppose estimates are $\hat{\sigma}^2 = 25$ and $\hat{\delta} = 1$. If we want 90% conditional power, set

$$\begin{aligned} \frac{\sqrt{n_2}\,\delta}{\sqrt{2\sigma^2}} &= z_{0.15} + z_{0.10} = 1.036 + 1.282 = 2.318 \\ n_2 &= \frac{2\sigma^2(2.318)^2}{\delta^2} = \frac{2(25)(2.318)^2}{1^2} \approx 269. \qquad (10) \end{aligned}$$

- Choose second stage sample size 269/arm.

Unblinded Sample Size Methods

- Adaptive methods based on treatment effect have been criticized on grounds of inefficiency and potentially strange behavior (Burman and Sonesson (2006); Jennison and Turnbull (2003); Jennison and Turnbull (2006)).

- Additionally, Tsiatis and Mehta (2003) showed that it is more powerful to set large initial sample size and use group-sequential monitoring to stop early.

- Therefore, in general, adaptive methods based on treatment effect should be avoided.

Outline

Introduction

Blinded Sample Size Methods

- Binary Outcomes
- Continuous Outcomes

Unblinded Sample Size Methods

Unplanned Changes

Unplanned Changes

- In rare cases, unplanned changes are needed (Posch and Proschan (2012)).
    - Investigators in TB trial examined blinded lung scans & discovered their primary endpoint could not be measured! They changed endpoint before breaking blind.

- **In blinded case like TB example, a *re-randomization test* still protects type 1 error rate under strong null hypothesis that treatment has no effect on any outcome examined.**
    - **Get null reference distribution by treating data as fixed constants, re-randomizing many times using whatever randomization method was used in trial, computing new value of test statistic for each re-randomization.**
    - **Compute p-value as proportion of re-randomized trials with result at least as extreme as observed result.**

Unplanned Changes

- What about unplanned sample size increase after examining data by arm?
- **One potentially useful method is due to Chen et al. (2004).**
  - **Compute conditional power (CP) under the current trend estimate, $B(t)/t = Z(t)/\sqrt{t}$ of drift parameter, θ (remember, $\theta = \mathrm{E}\{Z(1)\}$).**
  - **If CP is greater than** 0.50**, you can increase sample size with no penalty.**
- Justification: If CP$\geq$ 0.50, then increasing the sample size **decreases** the null conditional power.
- Proven to protect type 1 error rate with at most 1 interim analysis, and simulations suggest control of error rate with more interim analyses.

Unplanned Changes

- Müller and Schäfer (2004) proposed a method to make any design change:
  - Number or timing of interim analyses.
  - Sample size.
  - Primary endpoint.
  - Analysis method, etc.

- Let $\mathrm{CRP}_{\mathrm{orig}}$ be conditional rejection probability at interim analysis under original design:

  $$\mathrm{CRP}_{\mathrm{orig}} = P_0(\text{Cross future boundary with original design} \,|\, Z = z),$$

  where z is interim z-score.

- Make design change but make sure $\mathrm{CRP}_{\mathrm{new}} \leq \mathrm{CRP}_{\mathrm{orig}}$.

Unplanned Changes

- Controls type 1 error rate because

$$
\begin{aligned}
&P_0(\text{reject } H_0 \text{ with new design})\\
&= \int P_0(\text{reject } H_0 \text{ with new design} \,|\, Z = z)\phi(z)dz\\
&= \int \mathrm{CRP}_{\mathrm{new}}(z)\phi(z)dz \le \int \mathrm{CRP}_{\mathrm{orig}}(z)\phi(z)dz\\
&= \int P(\text{reject } H_0 \text{ with original design} \,|\, Z = z)\phi(z)dz\\
&= P(\text{reject } H_0 \text{ with original design}) = \alpha. \qquad (11)
\end{aligned}
$$

- **Still, use only in emergencies and for minor changes!**

Summary

- Adaptive methods use within-trial data to make design changes.
- Adaptive methods should be **preplanned**, although emergencies sometimes happen.
- Blinded methods are safer & less controversial than unblinded methods.
- Blinded sample size re-estimation is common.
  - Binary outcomes: Gould (1992): Use overall event rate and hypothesized treatment effect to compute p_T and p_C and re-compute sample size.
  - Continuous outcomes Gould and Shih (1992): Treat overall variance as within-arm variance and re-compute sample size.

Summary

- Unblinded sample size increase based on promising trend also safe if use Chen et al. (2004).
  - Can increase sample size if conditional power under current treatment effect estimate is ≥ 0.50.
  - Boundary at end is unchanged.
  - Logistical issues: Who makes sample size decision? Investigators should remain blinded.

- Müller and Schäfer (2004) can be used **in emergencies.**
  - Allows **any** design change after looking at data by arm.
  - Type 1 error rate is protected if conditional rejection probability (CRP) is $\leq$ CRP of original design.
  - Will generate controversy unless design change is minor.

References I

Bauer, P. and Köhne, K. (1994). Evaluation of experiments with adaptive interim analyses. Biometrics **50,** 1029–1041.

Burman, C.-F. and Sonesson, C. (2006). Are flexible designs sound? Biometrics **62,** 664–669.

Chen, Y. H., DeMets, D. L., and Lan, K. K. (2004). Increasing the sample size when the unblinded interim result is promising. Statistics in Medicine **23,** 1023–1038.

Gould, A. (1992). Interim analyses for monitoring clinical trials that do not materially affect the type i error rate. Statistics in Medicine **11,** 55–66.

Gould, A. L. and Shih, W. J. (1992). Sample size re-estimation without unblinding for normally distributed outcomes with unknown variance. Communications in Statistics-Theory and Methods **21,** 2833–2853.

Jennison, C. and Turnbull, B. (2003). Mid-course sample size modification in clinical trials based on the observed treatment effect. Statistics in Medicine **22,** 971–993.

References II

Jennison, C. and Turnbull, B. (2006). Adaptive and nonadaptive group sequential tests. Biometrika **93,** 1–21.

Müller, H. H. and Schäfer, H. (2004). A general statistical principle for changing a design any time during the course of a trial. Statistics in Medicine **23,** 2497–2508.

Posch, M. and Proschan, M. A. (2012). Unplanned adaptations before breaking the blind. Statistics in Medicine **31,** 4146–4153.

Proschan, M. A. (2021). Statistical Thinking in Clinical Trials. Chapman and Hall/CRC.

Proschan, M. A. and Hunsberger, S. A. (1995). Designed extension of studies based on conditional power. Biometrics pages 1315–1324.

Tsiatis, A. and Mehta, C. (2003). On the inefficiency of the adaptive design for monitoring clinical trials.