The Ugly, the Bad, and the Good of Missing and Dropout Data in Analysis and Sample Size Selection

Keith E. Muller
Professor and Director of the Division of Biostatistics, Epidemiology and Health Policy Research, COM
University of Florida, Gainesville
Keith.Muller@biostat.ufl.edu; web site http://ehpr.ufl.edu/muller

UF UNIVERSITY of FLORIDA

Motivation and Introduction

Missing and dropout data cause mischief in data analysis and sample size selection, from minor annoyances to catastrophes.

Module 1: How to classify missing data.
Module 2: How to avoid missing data.
Module 3: How to treat missing data in analysis.
Module 4: How to allow for missing data in study planning.

(Brief rest break when appealing)
Please silence your cell phones and pagers.
Invite brief questions of clarification.

UF UNIVERSITY of FLORIDA

Study planning requires a four step iterative process:
1) define the purpose
2) choose a design
3) define the analysis, in terms of variables
4) select a sample size and observation pattern (sampling plan)

UF UNIVERSITY of FLORIDA

The four step iterative process aligns with the presentation:
1) define the purpose; (Module 1: identify missing data)
2) choose a design; (Module 2: avoid missing data)
3) define the analysis; (Module 3: treat missing data in analysis)
4) select a sample size (Module 4: allow for missing data in N, p)

Handout has a bibliography.

UF UNIVERSITY of FLORIDA

Module 1: How to Classify Missing Data Types.

Context: a model expressing observed responses in terms of predictors, either occurring naturally or presented by the scientist.

Example: Loudness of distress call upon sighting a predator as a function of distance and type of predator.

Terms and Definitions

Observation unit: description of one instance of the response.

Independent sampling unit (ISU): an observation from one ISU is statistically independent of an observation from a distinct ISU.

Isolated primates can generate independent data.

Observations from members of a group (family) are often correlated, which can make the group the ISU.

Repeated measures: two or more observations measured on a common scale (hence *commensurate*).

Often repeated across time or space.

Multivariate measures: two or more observations in distinct units, different variables, example: {height, weight}

Doubly multivariate: {height, weight} repeated over time

How Many Variables?

Number of Responses	Number of Predictors	Model Description
1	1	Univariate
1	Many	Multivariable
Many	1 or many	Multivariate
Many	Many	Multivariate
Repeated	1 or many	Repeated measures

Types of Missing Data

missing *predictors*, X, missing between subjects (distance)

missing *responses*, Y, missing within subject (loudness of call)

Dropout: implicitly for time ordered (repeated measures), implies *monotone*, missing at time t implies missing at time t+1

Missing predictors much less common, much less studied.

Missing responses usually the problem.

Uninformatively missing avoids many concerns about bias.
missing completely at random,
missing at random,
drop out if loss to follow-up for unrelated reasons

Mis-timed data actually correspond to missing values in a finer grid of times, such as days rather than months.
Informatively missing important to detect and treated differently:
some drop outs, if related to response,
always censored data, including survival time,
below or above detection limits, high or low clipping

Purposefully Missing Data *Can be Created*

Incomplete designs (Latin squares, etc) (WHY?) omit some treatment combinations when have a jillion

Challenge protocols
exercise stress test
methacholine challenge
Time to event or trials to event may be the outcome.

Summary: The Ugly and the Bad

Statistical problems with missing data arise mostly from repeated and multivariate measures responses.
Missing data makes estimation hard, and inference harder, especially in small samples.
More trouble with analysis, computing, thinking, interpreting.
If statistical methods exist may not have software, or works poorly.
Missing data can bias estimates and bias tests (inflate alpha).
Typically reduces power (less chance to discover real difference).

There's Ugly, Flaming Ugly, and Iridescent Flaming Ugly.

Missing data can reflect or create logical or scientific problems in study design and interpretation.

However, if recognized the problem can be turned to our advantage.

A Real Example from Automobile Racing

Porsche has always emphasized the importance of reliability testing for success in endurance racing.
The company sent a team to a race track to test modifications to their current model, and had the team run an entire practice race.
Upon returning to the factory and happily reporting nothing had broken, the team was ordered back to the track for more testing.
Why? They had not learned anything,
they had not found the weakest links.
Missing information made the expensive and carefully done work of little value.

UF UNIVERSITY of FLORIDA

Summary: The Bad

Statistical problem come from partial fixes that can mislead by hiding problems.
Filling in missing values in unprincipled ways,
especially in small samples.
Example: replace a missing value with the mean.
Bad effects: falsely inflates the sample size and degrees of freedom so uses wrong distribution for confidence intervals and tests and
underestimates the variance.
May bias the means estimates if not done wisely.

UF UNIVERSITY of FLORIDA

The Possible Good in Missing Data

Missing data can define new responses (such as time to event) and uncover lurking variables.
Missingness may be the primary outcome.
Stress by drug dose example (a real study in rats)

	0mg/kg	10mg/kg	20mg/kg
stress	n=10	5	0
no stress	10	10	10

Overall, a good design, analysis, and planning process includes accommodations for missing and dropout data

UF UNIVERSITY of FLORIDA

A Real Example from World War II

Task: collect and analyze data on planes returning from combat missions to define improvements for crew survival.
Obvious response: tabulate bullet holes and other damage by location on the airplane, by type of plane and mission features.

Should we strengthen the planes around the bullet holes are?
No!
Anywhere else. All of the useful information arises from the missing data.

UF UNIVERSITY of FLORIDA

Module 2: Strategies for Avoiding Missing Data

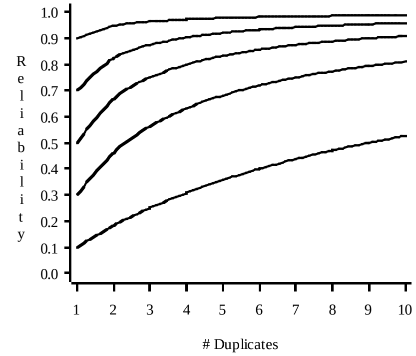
1) Measure fewer thing less often;
do less, better. (good career advice)

2) record data with a reliable machine (analog, digital)

3) redundant recording, backup (videotape, 2 of them, 2 observers)

4) measure the same variable more often, aliquots, replicates.

For exchangeable data, Spearman-Brown prophecy formula gives



5) Use point of observation data recording
and *data validation* (range, possible values,...)

Example: Epi Info (www.cdc.gov, freeware) on laptop for chart abstraction, surveys, etc. HIGHLY RECOMMENDED.

Obviously can use special purpose software.
Please avoid EXCEL. Pretty please.

6) use a truly practical design, pilot study, rehearsal and training.
Can you really run a video tape, take notes and ...?

Module 3: Treat Missing Data in Analysis

Always start by analyzing missingness patterns as a binary outcome. Listing tables can be good.
Look for patterns.
Look for biasing drop out such as censoring or below detection.
Are looking for informatively missing data.

Specific Missing Data Tools

Missing predictors not common, so less studied
  Missing responses usually the problem
Statisticians have created good but sometimes difficult to use methods for estimation with missing data
Much less has been done for hypothesis testing and other forms of inference, especially in small samples

Software Tips

Use software that is general (ANOVA, balance between vs within).
Avoid sums of squares formulas (and some software) that require balance.

Parametric Approaches

Most used tools are parametric approaches for MAR and MCAR, likelihood based such as EM algorithm.
Likelihood based methods such as EM algorithm, assume a particular distribution, such as a Gaussian distribution.
One of the big advantages of a mixed model formulation.
  Opens Pandora's box without careful use.
Problems from underfitting the covariance model,
  collinearity and automatic coding leading to nonconvergence.
Repeated measures typically do not have a simple covariance model.

Special Tools for Informatively Missing

Usually decent software for survival analysis, time to event.

Otherwise hope for special purpose free software for informatively missing data in particular special cases.

Multiple Imputation

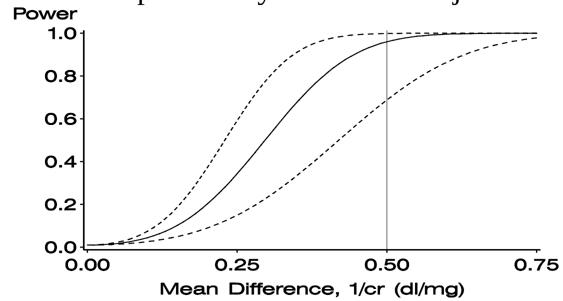
Multiple imputation uses a sub-sampling strategy (like jackknifing or boostrapping) to "fill in" the missing values.
Non-parametric tools, including multiple imputation.
  Multiple imputation great for LARGE studies.
  Very suspect in small studies.
  Easy to abuse.
Great resource at www.stat.psu.edu/~jls/mifaq.html

Analysis Tips And Strategies: Advice for All Seasons

1. Befriend or hire a statistician before you start.
2. Use powerful software. Don't use a screwdriver for a hammer.
3. Transformations for ratio scale data (Box-Cox).
   Assume errors Gaussian; check (jackknife studentized) residuals.
4. Center, scale, full rank code. Check and remove collinearity.
5. R^2, odds ratio, confidence intervals for any interpretation.
6. Fit models "backwards" from large to small, in fixed, planned order. (Use added-in-order SS and R squared, odds ratios)
7. Control multiple testing control.
8. Multistudy design strategies: Muller Barton Benignus (1984)

Module 4: Accounting for Missing in Sample Size

Power = probability of a correct rejection

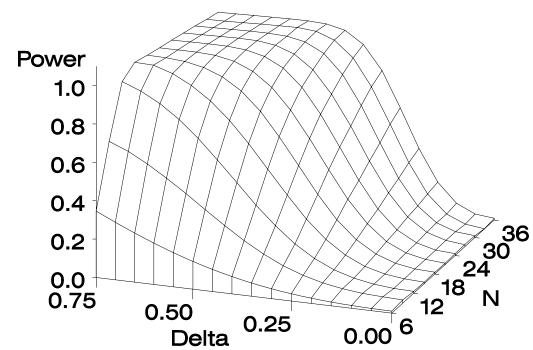


t-test power as a function of mean difference for 3 variances

Power Principles

For a fixed test, linear model power (ANOVA, regression) depends *only* on

1) mean differences
2) variance
3) sample size



t-Test Power as a Function of Mean Difference and N

Free Sample Size or Power

For linear models, 1) doubling N roughly equivalent to
2) doubling the mean difference or
3) cutting the variance in half.
Signaled avoidance with an auditory cue? Use a LOUD sound.
Measure repeatedly (blood aliquots), within animal.

Variability between animals >> variability within animals.

Use a Continuous Variable for More Power

Definition (a) *Nominal* scales only define categories or groups of observations.
(b) *Ordinal* scales provide numeric values sufficient only to rank observations.
(c) *Interval* scales provide numeric values with all differences of the same size being equivalent.
(d) *Ratio* scales give numeric values for which ratios of the same size are equivalent.

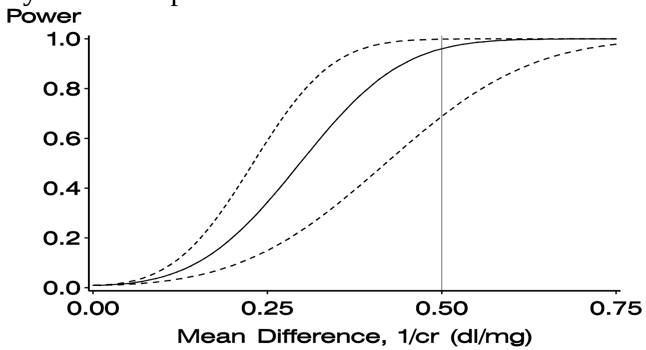
(e) *Continuous* data may include any sort of interval- and ratio-scale variables.
Differences due to scale on power essentially monotone,
higher scale gives higher power.
Why? Amount of information per value.
How important?
Sample sizes of 100's or 1000's versus 10's.

Arguably the most under used principles of power!

More Free Sample Size or Power

Ratio scale variables often (nearly always?) become more Gaussian with transformation which usually increases power as well as helping meet assumptions.
Examples: Cardiac ejection fraction,
CD34 counts, any concentration in blood or other liquid

Power uncertainty: nuisance parameters such as variance.



Confidence Limits for t-test Power with *Estimated* Variance)

The Silly Paradoxes of Power in Practice.

Power and sample size is hard work.
^Theory harder than for data analysis.
^Software very limited and worse than for data analysis.
^Defensible power analysis usually as much work as data analysis.
^Simplest forms and explanations often not clear.
^Power analysis seems to require more knowledge about the truth than seems fair to require.

We are often slackers.
-Scientists and statisticians want (demand) quick power.
-Scientists and statisticians want (demand) easy power analysis.
-Power often depends on unknown nuisance parameters. Yuck.
-Power often based on biased or very wobbly estimates.
-Scientists and statisticians want (demand) a sample size, a fixed answer, not a power analysis.

Threats to Good Power Analysis

The biggest threat to a defensible sample size comes from mis*alignment* of the power analysis with the data analysis, not from missing data.
Bad, common example: t test power for repeated measures.
Can either under or over estimate power.
Examples in Muller LaVange, Ramey and Ramey 1992.
Statisticians often just as guilty.
Solution: appropriate power analysis for the data analysis.

Next Biggest Threat

Failure to account for uncertainty

Solutions: library time,
 pilot study, external, or
 internal pilot design, or
 multistudy research.

Impatience and asking the data to do too much too quickly leads to sloppy low powered studies that frustrate, confuse and waste everyone's time and money.

If it were easy enough, someone else probably would have already done it.

Missing Data Adjustments for Power

Applying simple strategies to a fully aligned power analysis provide reasonable approximations with standard software.

Use adequate software (motorcycle helmet advice).

Account for substantial expected unbalance between groups.

It matters! Most power with balanced ANOVA.

Two Effects of Missing Data on Power

Missing responses has two possible effects:
 reduction in effective sample size
 unbalance.

Reducing sample size to adjust power only accurate and sufficient with roughly uniform missing pattern across conditions.

Dropout, with biggest effect at end, a common example of the possible bias in a power calculation.

Example power bias: 3 dose groups, with most missing at highest.

Just like unbalance across dose groups in concept, but theory different for repeated measures.

Adaptive Designs

pure sequential designs (testing widgets at a widget factory).

Classical group sequential design use interim data analysis with adjusted test levels to allow peeking at the data during the study.
 Mostly used in clinical trials.
 Almost all theory developed only in LARGE samples.
 Hence can be very biased in small samples.

Internal pilot designs conduct an interim power analysis,
but do not conduct an interim data analysis.
Useful when have uncertainty about nuisance variable, variance
For Gaussian data t test, use fixed scientifically important mean difference chosen a priori.
Use variance estimate after (typically) collected 1/2 of the data to do new power analysis and adjust sample size up or down.
Good methods for univariate linear models available for small samples, methods for some repeated measures settings coming in future years.

Summary Review

Module 1: How to classify missing data.
terminology
what to worry about a little, what to worry about a lot

Module 2: How to avoid missing data.
Plan,
simplify,
backups and duplicates,
automatic recording

Module 3: How to treat missing data in analysis.
Many possible answers, depending on the situation.
Must identify situation as first step.
Parametric methods still problematic in small samples,
special purpose methods in some cases.

Module 4: How to allow for missing data in study planning.
Conduct reasonable power analysis.
Consider both effective sample size reduction and
location of observations in the design.
Use multiple study approaches of various kinds.

I close by recommending the bibliography and online **scholarly** searches.

POWER TO THE PRIMATOLOGISTS!