

Missing Data in Health Research: The Good, the Bad and the Ugly

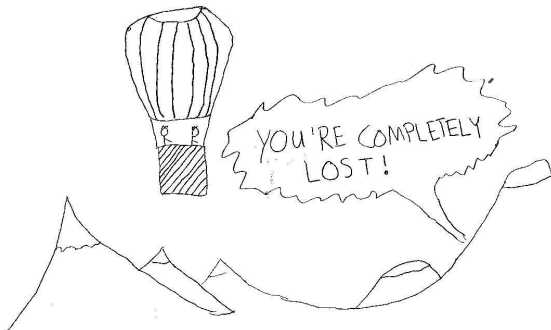
Brendan Klick

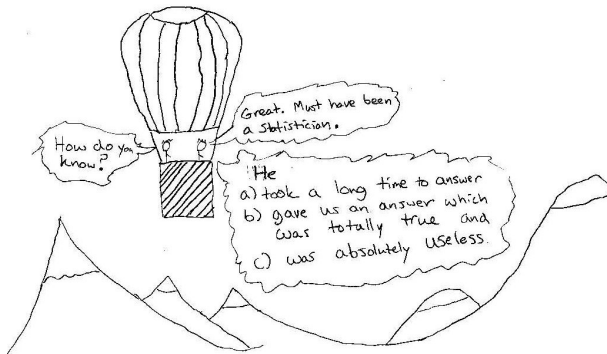
March 20, 2007

Two Friends Get lost in a
Hot Airballoon



Fifteen Minutes later they hear
a distant voice





Types of Missing Data

- ▶ Missing Completely at Random (MCAR)
- ▶ Missing at Random (MAR)
- ▶ Not Missing at Random (Informative) (NMAR)

Specifically, assume that a complete data set \mathbf{Y} , is comprised of two parts, \mathbf{Y}_{obs} , the data observed, and \mathbf{Y}_{mis} , the data which is missing. Let Y_i be the i th value in \mathbf{Y} and \mathbf{X} and be some additional completely observed data. For all data points Y_i in \mathbf{Y} , let $M_i = 1$ if Y in \mathbf{Y}_{mis} and $M_i = 0$ if Y_i in \mathbf{Y}_{obs} . Then data is MCAR if

$$Pr(M_i = 1 | \mathbf{Y}_{obs}, \mathbf{Y}_{mis}, \mathbf{X}) = c$$

for all i , data is MAR if

$$Pr(M_i = 1 | \mathbf{Y}_{obs}, \mathbf{Y}_{mis}, \mathbf{X}) = f(\mathbf{Y}_{obs}, \mathbf{X})$$

for all i and data is MNAR if

$$Pr(M_i = 1 | \mathbf{Y}_{obs}, \mathbf{Y}_{mis}, \mathbf{X}) = f(\mathbf{Y}_{obs}, \mathbf{Y}_{mis}, \mathbf{X})$$

for some i .

Example I of Missing Completely at Random

An investigator is conducting a study examining the effect of mind-body interventions on retinitis pigmentosa. Three subjects drop out of the control arm and one from the treatment arm. When asked why, subjects report that they are relocating for career reasons.

Example II of Missing Completely at Random

An investigator is studying the effects of eating on Ghrelin, a hormone that stimulates appetite. A sample sent to the lab from subject y is contaminated during analysis and therefore no data is recorded.

Example I for Missing at Random

An investigator is studying ethnic disparities in income. It's found that a proportional higher number of Hispanics refuse to answer questions concerning their income.

Example II for Missing at Random

An investigator is studying the effect of a mind-body intervention on subjects with anorexia. Subjects are followed longitudinally for two years and their weight is measured. However, if at any point a subject's weight, compared to their baseline rate, has decreased by more than 15 lbs the subject is removed from the study and referred to an in-patient weight center.

Example for Not Missing at Random (Informative)

An investigator is examining the effect of sleep on pain. Subjects are called daily and asked questions about last night's sleep and their pain today. Patients who are experiencing severe pain are more likely to not come to the phone leaving the data missing for that particular day.

Two Forms of Bias

Consider a study where we want to estimate the income of diabetic patients attending a clinic in Baltimore. Suppose that older women tend to refuse to disclose their income more in the the study.

If we calculate the mean income based on the data that we have observed, our estimate of the population mean will be biased due to **differential non-response**.

Two Forms of Bias (Cont'd)

Now suppose the demographics of people who refuse to disclose their income appears similar to the overall population. However, in actuality people with lower incomes regardless of age, race, gender, etc. tend to refuse to disclose this information.

If we calculate the mean income based on the data that we have observed, our estimate of the population mean will be biased due to an **informative missing data mechanism**.

Since Missing Not at Random data is dependent on the value of unobserved variables this missing data condition is fundamentally untestable statistically. However, there are methods which can provide clues.

Method I: Departure from Normality

If data is MCAR,

$$f(Y|M, \theta) = f(Y|\theta).$$

If data is NMAR,

$$f(Y|M, \theta) \neq f(Y|\theta).$$

Suppose we know \mathbf{Y} is normally distributed. If data is MCAR, $f(Y_{obs})$ is still normally distributed. If on the other hand data is NMAR, $f(Y_{obs})$ will in most cases not be normally distributed. We can use normality tests such Kolmogorov-Smirnov, Shapiro-Wilks, or Andersen-Darling to see whether Y_{obs} is normally distributed.

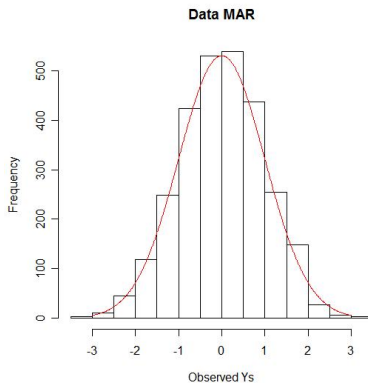
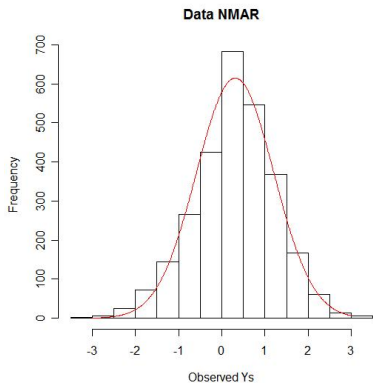
Example

We simulate two datasets both consisting of 200 data points from a normal distribution of which an average of 30% will be missing. In one dataset, data is MCAR. In other dataset, data is NMAR, specifically,

$$Pr(M = 1|Y = y) = \frac{4e^{-4y}}{5(e^{-4y} + 1)}.$$

We use a Kolmogorov-Smirnov test to assess departure from normality for the non-missing data points in each dataset. We repeat 300 times.

Comparison of NMAR and MCAR



	NMAR	MCAR
Median of the 300 p-values from K-S test	.0001	.499

Method II: Longitudinal Data

In longitudinal studies where we follow subjects over a period of time, missing data often occurs due to dropout. When dropout occurs, for a subject we have data up to time t and no data after this. In most longitudinal settings

$$\text{corr}(Y_{i,j}, Y_{i,k}) \neq 0$$

for subject i . If this is the case, we can use $Y_{i,t}, Y_{i,t-1}, \dots$ to predict $Y_{i,t+1}$. If data is MCAR,

$$Pr(M_{i,t+1} = 1 | Y_{i,t}, Y_{i,t-1}, \dots) = Pr(M_{i,t+1} = 1).$$

Take Home Message

- ▶ It is in an investigators best interest to attempt, if possible, to understand why and how missingness has occurred.
- ▶ Deciding whether missing data is MCAR, MAR, or NMAR is important for the best analysis of data. Making this determination is easiest when the reasons for missingness are known.
- ▶ In longitudinal settings it is important to investigate how treatment outcome predicts dropout.

Methods of Analysis

- ▶ Complete case analysis.
- ▶ Weighting procedures.
- ▶ Imputation procedures.
- ▶ Likelihood procedures.

Complete Case Analysis

The simple solution

Advantages

- ▶ Simple to compute and easy to explain.
- ▶ If data is MCAR or MAR, results are unbiased.
- ▶ P-value, standard error estimates and hypothesis tests are correct.

Disadvantages

- ▶ Can give biased estimates if data are NMAR.
- ▶ Does not use all available information (does not base inference of sufficient statistics). Statistical power may be significantly smaller than other procedures.

An Example

Suppose we conduct a simple linear regression of the form

$$Y_i = \beta_0 + \beta_1 x_i + e, e \sim N(0, \sigma^2), i = 1, \dots, n.$$

Suppose that x_i is recorded for all n observations but some of Y_i 's are missing MCAR or MAR. Since the mean of Y_i 's are conditional on design variables x_i 's, no information can be provided by the partially missing data. Furthermore, the MCAR or MAR assumption guarantees that inference based on the complete cases will be unbiased.

Weighting Procedures

Advantages

- ▶ Provides sample representative inference.
- ▶ Can remove some nonresponse bias.
- ▶ Simple.

Disadvantages

- ▶ Can give biased estimates if data are NMAR.
- ▶ Can ignore data. Not efficient.

An Example

Consider a survey to measure average health care expenditures in the US population. We survey 1000 people by telephone of which 850 are caucasians, 50 are Hispanic and 100 are African American. The refusal rate is 5 percent in caucasians but 20 percent in Hispanics and African Americans. An unbiased estimator of population mean log medical expenditures is

$$\bar{Y}_{pop} = \frac{850 \times \bar{Y}_{cauc} + 50 \times \bar{Y}_{Hisp} + 100 \times \bar{Y}_{AfAm}}{1000}$$

where \bar{Y}_{cauc} , \bar{Y}_{Hisp} and \bar{Y}_{AfAm} are the observed sample mean log medical expenditures for caucasians, Hispanics and African Americans respectively. But note a good estimate for the standard error of the mean is NOT the usual

An Example (Cont'd)

$$\widehat{sem} = \frac{s}{\sqrt{n}}$$

where s is the sample standard deviation. There is a better estimate but it's quite messy.

Imputation Methods

“It often amuses me to hear men impute all their misfortunes to fate, luck, or destiny, whilst their successes or good fortune they ascribe to their own sagacity, cleverness or penetration.” Samuel Taylor Coleridge.



Advantages

- ▶ (Generally) valid under MCAR and MAR assumptions.
- ▶ Use available data in a complete way.
- ▶ Can preserve balanced designs necessary for certain statistical procedures such as repeated measures ANOVA, MANOVA and Tukey HSD post-hoc comparisons.
- ▶ Using assumptions about the missing data pattern can protect against NMAR.

Disadvantages

- ▶ If performed incorrectly it may lead to biased results.
- ▶ Tempting to make assumptions based on standard errors, p-values, confidence intervals from the method of analysis of imputed dataset which do not account for the uncertainty in the imputation procedure.
- ▶ Currently available software is limited and technical.
- ▶ Can be computational resource intensive.

“The idea of imputation is both seductive and dangerous. It is seductive because it can lull the user into the pleasurable state of believing that the data are complete after all, and it is dangerous because it lumps together situations where the problem is sufficiently minor that it can be legitimately handled in this way and situations where standard estimators applied to the real and imputed data have substantial biases.” Dempster and Rubin, 1983.

Single Imputation Techniques

- ▶ Mean/Median Imputation.
- ▶ Regression Imputation.
- ▶ Stochastic error imputation.
- ▶ Hot deck imputation-substitute missing data with similar responding observation.
- ▶ Substitution-a nonresponse causes another similar unit to be surveyed.
- ▶ Cold deck-filling missing data using previous data or an external study.
- ▶ EM algorithm imputation/likelihood algorithm.

A Note about Last Observation Carried Forward (LOCF)

LOCF for dropout is a form of Hot Deck imputation which is most valid under to MAR assumptions. When a randomized trial of a drug or intervention is being conducted under an intent-to-treat framework, it can be argued that since subjects are expected to be improving, imputation by LOCF produces conservative estimates of treatment effect. In certain contexts this may be a reasonable approach but there are often better ways of handling dropout.

Warning Message!!

After performing imputation complete data analysis procedures can be used for estimation and perform hypothesis testing. However, standard errors of estimates formed by complete data procedures do NOT take into account the uncertainty involved in the imputation. Correct standard error estimates can be formed using the bootstrap or jackknife methods. Or better yet perform....

Multiple Imputation

- ▶ Proposed by D. Rubin in 1978.
- ▶ Becoming a standard method of handling missing data.
- ▶ 373 citations on PubMed.

Basic Idea

Let $T(\mathbf{Y})$ with associated variance V be an estimator of parameter θ provided data \mathbf{Y} . Some portion, \mathbf{Y}_{mis} , is missing. Assume that $Y_{mis} \sim g(y_{mis})$. Fill in the missing data, to form “complete” dataset \mathbf{Y}_1 , by randomly drawing from $g(y_{mis})$. Repeat this procedure D times, to form $\mathbf{Y}_2, \dots, \mathbf{Y}_D$ datasets. Find estimators $T_1(\mathbf{Y}_1), \dots, T_D(\mathbf{Y}_D)$ and associated variances V_1, \dots, V_D . Define

$$\tilde{T} = \frac{\text{sum}[T_1(\mathbf{Y}_1), \dots, T_D(\mathbf{Y}_D)]}{D}$$

Basic Idea (Cont'd)

and

$$\tilde{V} = \frac{\text{sum}(V_1, \dots, V_D)}{D}.$$

Then

$$\text{var}(\tilde{T}) = \tilde{V} + \frac{D+1}{D^2 - D} \sum_{d=1}^D (T_d(\mathbf{Y}_d) - \tilde{T})^2.$$

Approximate Bayesian Bootstrap

A major difficulty with imputation is deciding on a suitable distribution from which to base the imputation. Under the MCAR mechanism, the approximate Bayesian Bootstrap is a reasonable solution. In this case, partially missing observations are filled in by randomly selecting from similar completely observed observations.

Another popular solution is to use the Gibbs Sampler/MCMC to draw missing data from a posterior distribution.

Very Simplified Example of the Bayesian Bootstrap

In a study of health care expenditure, we ask people participating in a survey to estimate last month's health care expenditures. Ellen can't remember the answer to this question. Therefore we chose to randomly substitute a response from Ruth, Sandra, Carolyn, Nancy, Barbara, or Lisa. All of whom are relatively healthy, married, women ages 50-60 having similar socio-economic backgrounds.

Cautionary Note

Multiple imputation is becoming widely used in the literature. However, some if not most scientists don't understand the dangers. Consider using regression model

$$E(Y) = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k$$

to make inference about β_1 . Suppose that the y 's are fully observed but some of the x_1 's are missing. Suppose the regression model is correct and we use a multiple imputation with the missing x_1 's chosen from the distribution

$$p(x_1 | x_2, \dots, x_k).$$

Cautionary Note (Cont'd)

Then

- ▶ the complete case estimator of β_1 will be unbiased.
- ▶ the estimated standard error of the mean based on MI will almost certainly be smaller than the estimated standard in the complete case analysis. However,
- ▶ if either the missing x_1 's are NMAR or $p(x_1|x_2, \dots, x_k)$ is incorrect the MI estimator of β_1 will be biased.
- ▶ in some cases using MI is making a bias/variance trade-off!

Practical Example—Data from Burn Victims

Burn victims are asked about insomnia at hospital discharge. These victims are followed longitudinally for two years and asked a variety of health questions. We are interested in whether self-reported insomnia at discharge predicts a greater improvement in pain scores. We use a linear mixed model to examine whether change in sf36 bodily pain is predicted by time since burn, discharge insomnia, discharge sf36 mental health, preburn sf36 general health and pain, gender, burn area, graft area, days in the ICU, age at burn and education.

Practical Example (Cont'd)

Unfortunately,

- ▶ only 286 people have complete predictor data,
- ▶ 350 have insomnia data at discharge but are missing some of the other predictor information, and
- ▶ 520 have some predictor information.

We perform two multiple imputation for

- ▶ the 350 people having discharge insomnia data, and
- ▶ the 520 have some discharge information.

Practical Example (Cont'd)–results

Analysis	Insomnia			Age		
	$\hat{\beta}$	\widehat{sem}	p-value	$\hat{\beta}$	\widehat{sem}	p-value
complete case	9.5	4.5	.036	0.26	0.11	.018
MI for n=350	11.6	3.9	.003	0.23	0.09	0.011
MI for n=520	7.5	3.3	0.023	0.23	0.07	0.001

Recommendation

When contemplating Multiple Imputation I recommend:

- ▶ paying close attention to model choice, then
- ▶ examine the standard error of model estimates.
- ▶ If these are found to be undesirable, then do MI. Finally,
- ▶ compare MI estimates to complete case estimates. If these are found to be similar there is probably little danger in making inference based on the MI standard errors.

A Note about NMAR and Imputation

In cases where data is NMAR

$$Pr(Y|M, \theta) \neq Pr(Y|\theta).$$

We would like to know $Pr(Y|M = 1, \theta)$. Suppose we know or “guestimate” $Pr(Y|\theta)$, $Pr(M = 1|Y)$, and $Pr(M = 1)$ Bayes' rule gives us

$$Pr(Y|M = 1, \theta) = \frac{Pr(Y|\theta)Pr(M = 1|Y)}{Pr(M = 1)}.$$

Software

- ▶ One of the best easily available software for MI is SAS using Proc Mi.
- ▶ Stata's impute command does not perform true multiple imputation. However, user written programs ICE and WHOTDECK perform various forms of MI.
- ▶ In R and S-Plus, multiple imputation is available with the MICE and Hmisc packages.
- ▶ Not readily available in SPSS.

Likelihood Methods for Missing Data



Advantages

- ▶ Valid under MCAR and MAR assumptions.
- ▶ Uses available data in a completely way.
- ▶ Produces “exact” procedures for inference unlike imputation. Particularly appealing in small sample problems.

Disadvantages

- ▶ Difficult or impossible to compute in many situations.
- ▶ Can be computational resource intensive.

Example I

Consider three draws, A , B , C from a bivariate normal distribution

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N \left[\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \begin{pmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{pmatrix} \right].$$

For draw A , X and Y are both observed, for draw B only X is observed, and for draw C only Y is observed. The observed data likelihood is

$$L_{obs} = L_A \times L_B \times L_C$$

where L_A is a multivariate normal density involving $(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \sigma_{XY})$ and L_B and L_C are univariate normal densities involving (μ_X, σ_X^2) and (μ_Y, σ_Y^2) respectively. The observed data likelihood can be maximized with respect to $(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \sigma_{XY})$ using the Fisher scoring, Newton-Raphson or EM algorithms.

Example II

In longitudinal data analysis, we observe subjects at times t_1, t_2, \dots, t_m . General Estimating Equations and Linear Mixed Effects Models can be used on subjects who have partial data due to missing observations at times $t_i, i = 1, 2, \dots, m$. Inferences from these analyses are based on a complete data likelihood function which utilizes all observations made.

Final Simulation Example

Suppose we are interested in the effect of X_1 on Y . Assume, the true relationship of X_1 and Y is

$$Y = X_1 + X_2 + X_3 + \epsilon, \epsilon \sim N(0, 3^2)$$

where X_2 and X_3 are related confounding variables. Suppose X_1 , X_2 , and X_3 are multivariate normally distributed

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} \sim N \left[\begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}, \begin{pmatrix} 1 & .4 & .3 \\ .4 & 1 & .4 \\ .3 & .4 & 1 \end{pmatrix} \right].$$

Final Simulation Example (Cont'd)

We make roughly 27, 25, 20 percent of the X_1 's, X_2 's and X_3 's missing respectively with greater probability of lower values being missing. We then estimate the bias, and standard error for our estimate of $\beta_1 (= 1)$ using several methods of handling the missing data:

- ▶ complete case analysis,
- ▶ mean imputation,
- ▶ EM algorithm to impute data,
- ▶ Maximum likelihood estimate,
- ▶ MI by Bayesian Bootstrap, and
- ▶ MI by MCMC.

Final Simulation Example (Cont'd)

Simulation Results

method	bias	sem
all data	0	0.50
complete case analysis	0	0.84
mean imputation	0.08	0.58
EM algorithm	0	0.80
Maximum likelihood estimate	0.06	0.60
MI by Bayesian Bootstrap	0.04	0.51
MI by MCMC	0.06	0.72

Conclusions

- ▶ Investigators should, as much as possible, examine why missing data occur in their studies.
- ▶ It is worthwhile understanding the different missing data mechanisms and their implications for data analysis.
- ▶ The choice of the best statistical procedure to handle missing data is usually open-ended and depends on the type and conditions of a study.

References

- ▶ Little RJA, Rubin DB. 2002. *Statistical Analysis of Missing Data*. New York: Wiley.
- ▶ Raghunathan TE. 2004. What do we do with missing data? Some options for analysis of incomplete data. *Annu. Rev. Public Health* 25:99-117.
- ▶ Rubin DB. 1974. Characterizing the estimation of parameters in incomplete data problems. *J. Am. Statist. Assoc.* 69:467-74.
- ▶ Rubin DB. 1976. Inference and missing data. *Biometrika* 63:581-02.

Acknowledgments

My thanks to CMBR for the invitation to talk about this topic.