# Lec 11: Missing Data
*MATH 456 - Spring 2016*

## Reading & References

- Little, R. and Rubin, D. Statistical Analysis with Missing Data, 2nd Ed., Wiley, 2002
  - Standard reference
  - Requires some math

- Allison, P. Missing Data, Sage, 2001
  - Small and cheap
  - Requires very little math

- Multiple Imputation.com http://www.stefvanbuuren.nl/mi/
- A. Afifi, et al. "Methods for improving Regression Analysis for Skewed Continuous or Counted Responses" Annual Review of Public Health, 2007, 28:95–111 http://arjournals.annualreviews.org/doi/pdf/10.1146/annurev.publhealth.28.082206.094100
- http://www.analyticsvidhya.com/blog/2016/03/tutorial-powerful-packages-imputing-missing-values/

## Terminology

- General: Item non-response. Individual pieces of data are missing.

- Unit non-response: Records have some background data on all units, but some units don't respond to any question.
- Monotonone missing data: Variables can be ordered such that one block of variables more observed than the next.

## Recap: Identifying Item non-response

- Missing data in `R` is denoted as `NA`
- Arithmetic functions on missing data will return missing

```
library(MASS)
head(survey$Pulse)
```

```
## [1]  92 104  87  NA  35  64
```

```
mean(survey$Pulse)
```

```
## [1] NA
```

The `summary()` function will always show missing.

```
summary(survey$Pulse)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   35.00   66.00   72.50   74.15   80.00  104.00      45
```

The `is.na()` function is helpful to identify rows with missing data

```
table(is.na(survey$Pulse))
```

```
##
## FALSE   TRUE
##   192     45
```

The function `table()` will not show NA by default.

```
table(survey$M.I)
```

```
##
## Imperial   Metric
##       68      141
```

```
table(survey$M.I, useNA="always")
```

```
##
## Imperial   Metric     <NA>
##       68      141       28
```

### Effects of Nonresponse

Textbook example: Example reported in W.G. Cochran, Sampling Techniques, 3rd edition, 1977, ch. 13

> Consider data that come form an experimental sampling of fruit orcharts in North Carolina in
> 1946. Three successive mailings of the same questionnaire were sent to growers. For one of the
> questions the number of fruit trees, complete data were available for the population. . .

| Ave. # trees | # of growers | % of pop'n | Ave # trees/grower |
|---|---|---|---|
| 1st mailing responders | 300 | 10 | 456 |
| 2nd mailing responders | 543 | 17 | 382 |
| 3rd mailing responders | 434 | 14 | 340 |
| Nonresponders | 1839 | 59 | 290 |
| Total population | 3116 | 100 | 329 |

- The overall response rate was very low.
- The rate of non response is clearly related to the average number of trees per grower.
- The estimate of the average trees per grower can be calculated as a weighted average from responders $\bar{Y}_1$ and non responders $\bar{Y}_2$.

**Bias**: The difference between the observed estimate $\bar{y}_1$ and the true parameter $\mu$.

$$E(\bar{y}_1) - \mu = \bar{Y}_1 - \bar{Y}$$
$$= \bar{Y}_1 - \left[(1-w)\bar{Y}_1 - w\bar{Y}_2\right]$$
$$= w(\bar{Y}_1 - \bar{Y}_2)$$

Where $w$ is the proportion of non-response.

- The amount of bias is the product of the proportion of non-response and the difference in the means between the responders and the non-responders.
- The sample provides no information about $\bar{Y}_2$, the size of the bias is generally unknown without information gained from external data.

# Missing Data Mechanisms

Process by which some units observed, some units not observed

- Missing Completely at Random (MCAR): The probability that a data point is missing is completely unrelated (independent) of any observed and unobserved data or parameters.
  - P(Y missing| X, Y) = P(Y missing)
  - Ex: Miscoding or forgetting to log in answer
- Missing at Random (MAR): The probability that a data point is missing is independent can be explained or modeled by other observed variables.
  - P(Y missing|x, Y) = P(Y missing | X)
  - Ex: Y = age, X = sex - Pr (Y miss| X = male) = 0.2 - Pr (Y miss| X = female) = 0.3 - Older people are less likely to fill out an income survey, so the missing data on income is related to age. After accounting for age the missing data is unrelated to income.
- Not missing at Random (NMAR): The probability that a data point is missing depends on the value of the variable in question.
  - P(Y missing | X, Y) = P (Y missing|X, Y)
  - Ex: Y = income, X = immigration status
    * Richer person may be less willing to disclose income
    * Illegal immigrant may be less willing to disclose income

Does it matter to inferences? **Yes!**

## Demonstration via Simulation

### MCAR

1. Draw a random sample of size 100 from a standard Normal distribution (Z) and calculate the mean.

```
z <- rnorm(100)
mean.z <- mean(z)
mean.z
```

```
## [1] 0.04184733
```

2. Delete data at a rate of $p$ and calculate the complete case (available) mean.

- Sample 100 random Bernoulli (0/1) variables with probability $p$.

```
x <- rbinom(100, 1, p=.5)
```

- Find out which elements are are 1's

```
delete.these <- which(x==1)
```

- Set those elements in z to NA.

```
z[delete.these] <- NA
```

- Calculate the complete case mean

```
mean(z, na.rm=TRUE)
```

```
## [1] 0.1869944
```

3. Calculate the bias
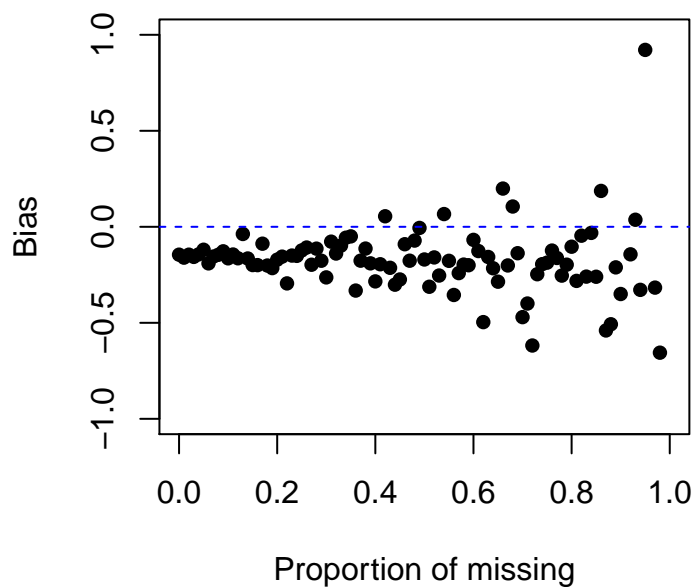
```
mean.z - mean(z, na.rm=TRUE)
```

```
## [1] -0.145147
```

How does the bias change as a function of the proportion of missing? Let $p$ range from 0% to 99% and plot the bias as a function of $p$.

```
calc.bias <- function(p){
  mean.z - mean(ifelse(rbinom(100, 1, p)==1, NA, z), na.rm=TRUE)
}

p <- seq(0,.99,by=.01)

plot(c(0,1), c(-1, 1), type="n", ylab="Bias", xlab="Proportion of missing")
  points(p, sapply(p, calc.bias), pch=16)
  abline(h=0, lty=2, col="blue")
```

**NMAR: Missing related to data**

What if the rate of missing is related to the value of the outcome? 1. Randomly draw 100 random normal samples.

```
Z <- rnorm(100)
```

2. Randomly draw 100 uniform random values between 0 and 1 to serve as probabilities.

```
p <- runif(100, 0, 1)
```

3. Sort both the value of Z and the probability of missingness $p$ in ascending order

```
dta <- data.frame(Z=sort(Z), p=sort(p))
head(dta)
```

```
##           Z             p
## 1 -2.827216 0.0007012342
## 2 -2.240347 0.0141965412
## 3 -2.180211 0.0231836869
## 4 -1.773113 0.0365393106
## 5 -1.761387 0.0365585333
## 6 -1.559321 0.0499052980
```

```r
ggplot(dta, aes(x=p, y=Z)) + geom_point() + xlab("P(missing)") + ylab("Z~Normal(0,1)")
```
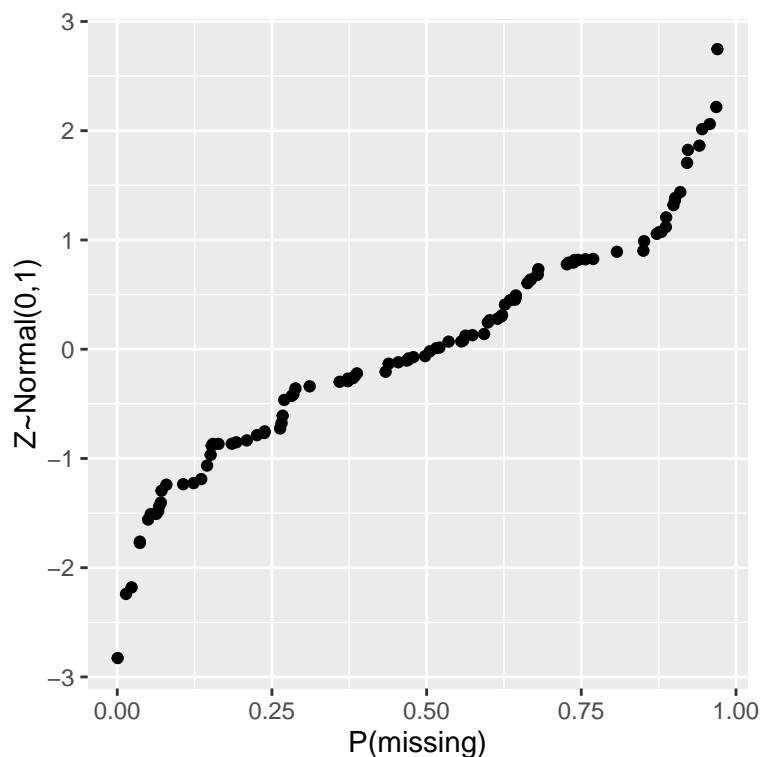


4. Set $Z$ missing with probability equal to the $p$ for that row. _Create a new vector `dta$z.miss` that is either 0, or the value of `dta$Z` with probability `1-dta$p`. Then change all the 0's to `NA`.

```r
dta$Z.miss <- dta$Z * (1-rbinom(NROW(dta), 1, dta$p))
head(dta)
```

```
##          Z           p     Z.miss
## 1 -2.827216 0.0007012342 -2.827216
## 2 -2.240347 0.0141965412 -2.240347
## 3 -2.180211 0.0231836869 -2.180211
## 4 -1.773113 0.0365393106 -1.773113
## 5 -1.761387 0.0365585333 -1.761387
## 6 -1.559321 0.0499052980 -1.559321
```

```r
dta$Z.miss[dta$Z.miss==0] <- NA
```

5. Calculate the complete case mean and the bias

```r
mean(dta$Z.miss, na.rm=TRUE)
```

```
## [1] -0.6462471
```

```r
mean(dta$Z) - mean(dta$Z.miss, na.rm=TRUE)
```

```
## [1] 0.6491546
```

**NMAR: Pure Censoring**

Consider a hypothetical blood test to measure a hormone that is normally distributed in the blood with mean $10\mu g$ and variance 1. However the test to detect the compound only can detect levels above 10.

```r
z <- rnorm(100, 10, 1)
y <- z
y[y>10] <- NA
mean(y, na.rm=TRUE)
```

```
## [1] 9.287761
```

When the data is not missing at random, the bias can be much greater.

*Problem: Usually you don't know the missing data mechanism.*

## Degrees of difficulty

- MCAR is easiest to deal with.
- MAR: we can live with it.
- NMAR: most difficult to handle.

## Evidence?

What can we learn from evidence in the data set at hand?

- May be evidence in the data rule out MCAR - test responders vs. nonresponders.
    - Example: Responders tend to have higher/lower average education than nonresponders by t-test
    - Example: Response more likely in one geographic area than another by chi-square test
- No evidence in data set to rule out MAR (although there may be evidence from an external data source that bears on this

## What is plausible?

- Cochran example: when human behavior is involved, MCAR must be viewed as an extremely special case that would often be violated in practice
- Missing data may be introduced by design (e.g., measure some variables, don't measure others for reasons of cost, response burden), in which case MCAR would apply
- MAR is much more common than MCAR
- We cannot be too cavalier about assuming MAR, but anecdotal evidence shows that it often is plausible when conditioning on enough information

### Ignorable nonresponse

- If missing-data mechanism is MCAR or MAR then nonresponse is said to be "ignorable"
- Origin of name: in likelihood-based inference, both the data model and missing-data mechanism are important but with MCAR or MAR, inference can be based solely on the data model, thus making inference much simpler

- "*Ignorability*" is a relative assumption: missingness on income may be NMAR given only gender, but may be MAR given gender, age, occupation, region of the country

## General strategies

For handling missing data

- Complete-case/available-case analysis: drop cases that make analysis inconvenient

- Imputation procedures: fill in missing values, then analyze completed data sets using complete-date methods

- Weighting procedures: modify "design weights" (i.e., inverse probabilities of selection from sampling plan) to account for probability of response

- Model-based approaches: develop model for partially missing data, base inferences on likelihood under that model

### Ignore it

**Complete cases analysis**

If not all variables observed, delete case from analysis

- Advantages:
  - Simplicity
  - Common sample for all estimates
- Disadvantages:
  - Loss of information, e.g., if there are 20 variables, each 10% missing MCAR, then expect 12% complete cases (0.92 -0.12)
  - Bias due to violation of MCAR

**Available-case analysis**

- Use all cases where the variable of interest is present
  - Potentially different sets of cases for means of X and Y
  - and complete pairs for $r_{XY}$

- Tempting to think that available-case analysis will be superior to complete-case analysis

- But it can distort relationships between variables by not using a common base of observations for all quantities being estimated.

**Imputation**

Fill in missing values, analyze completed data set

- Advantage:
    - Rectangular data set easier to analyze
- Disadvantage:
    - "Both seductive and dangerous"" (Little and Rubin)
    - Can understate uncertainty due to missing values.
    - Can induce bias if imputing under the wrong model.

**Methods**

- Unconditional mean substitution. **Never use**
    - Impute all missing data using the mean of observed cases
    - Highly biased
    - Artificially decreases the mean.
- Hot deck imputation
    - Impute values by randomly sampling values from observed data.

    - Good for categorical data
    - Reasonable for MCAR and MAR
- Model based imputation
    - Conditional Mean imputation: Use regression on observed variables to estimate missing values
    - Predictive Mean Matching: Fills in a value randomly by sampling observed values whose regression-predicted values are closest to the regression-predicted value for the missing point.
        * Cross between hot-deck and conditional mean
    - Categorical data can be imputed using classification models
    - Less biased than mean substitution
    - but SE's could be inflated
- Adding a residual
    - Impute regression value $\pm$ a randomly selected residual based on estimated residual variance
    - Over the long-term, we can reduce bias, on the average

. . . but we can do better.

# Multiple Imputation (MI)

## Goals

- Accurately reflect available information
- Avoid bias in estimates of quantities of interest
- Estimation could involve explicit or implicit model
- Accurately reflect uncertainty due to missingness

## Technique

1. For each missing value, impute $m$ estimates (usually $m = 5$)

   - Imputation method must include a random component

2. Create $m$ complete data sets
3. Perform desired analysis on each of the $m$ complete data sets
4. Combine final estimates in a manner that accounts for the between, and within imputation variance.

## MI as a paradigm

- Logic: "Average over" uncertainty, don't assume most likely scenario (single imputation) covers all plausible scenarios
- Principle: Want nominal 95% intervals to cover targets of estimation 95% of the time
- Simulation studies show that, when MAR assumption holds:

  - Proper imputations will yield close to nominal coverage (Rubin 87)
  - Improvement over single imputation is meaningful
  - Number of imputations can be modest - even 2 adequate for many purposes, so 5 is plenty

*Rubin 87: Multiple Imputation for Nonresponse in Surveys, Wiley, 1987).*

## Inference on MI

Consider $m$ imputed data sets. For some quantity of interest $Q$ with squared $SE = U$, calculate $Q_1, Q_2, \ldots, Q_m$ and $U_1, U_2, \ldots, U_m$ (e.g., carry out $m$ regression analyses, obtain point estimates and SE from each).

Then calculate the average estimate $\bar{Q}$, the average variance $\bar{U}$, and the variance of the averages $B$.

$$\bar{Q} = \sum_{i=1}^{m} Q_i/m$$

$$\bar{U} = \sum_{i=1}^{m} U_i/m$$

$$B = \frac{1}{m-1} \sum_{i=1}^{m} (Q_i - \bar{Q})^2$$

Then $T = \bar{U} + \frac{m+1}{m} B$ is the estimated total variance of $\bar{Q}$.

Significance tests and interval estimates can be based on

$$\frac{\bar{Q} - Q}{\sqrt{T}} \sim t_{df}, \text{ where } df = (m-1)(1 + \frac{1}{m+1}\frac{\bar{U}}{B})^2$$

- df are similar to those for comparison of normal means with unequal variances, i.e., using Satterthwaite approximation.
- Ratio of (B = between-imputation variance) to (T = between + within-imputation variance) is known as the fraction of missing information.

## Practical Considerations

- Typically desirable to include many predictors, both to
    - improve precision of imputed values
    - make MAR assumption more plausible

- But the number of covariance parameters goes up as the square of the number of variables in the model,
- implying practical limits on the number of variables for which parameters can be estimated well

## Bottom Line for MI

Little & Rubin (Book, p. 218):

> "In our experience with real and artificial data..., the practical conclusion appears to be that multiple imputation, when carefully done, can be safely used with real problems even when the ultimate user may be applying models or analyses not contemplated by the imputer."

Read more about this here: https://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug_mi_sect032.htm

Imputation methods for complex survey data and data not missing at random is an open research topic.

# Imputation Methods in R

## Packages

We will go through a few new R packages that will help intelligently impute missing data. Install all of them sooner rather than later. We will use some of the functions (such as `prodNA`) to more easily generate missing values in a data frame.

Most of the following demonstrations come directly from the following tutorial: http://www.analyticsvidhya.com/blog/2016/03/tutorial-powerful-packages-imputing-missing-values/

- VIM `install.packages("VIM")`
- HMISC `install.packages("Hmisc")`
- Amelia `install.packages("Amelia")`
- MICE `install.packages("mice")`
- missForest `install.packages("missForest")`

```r
lib <- c("VIM", "Hmisc", "Amelia", "mice", "missForest")
invisible(lapply(lib, library, character.only=T))
```

### Data Example: Iris data

Let's use the `iris` training data to compare imputation methods. Set a seed and use the `prodNA()` function from the `missForest` package to create 10% missing values in this data set.

```r
iris <- iris
prop.table(table(is.na(iris)))
```

```
## 
## FALSE 
##     1
```

```r
set.seed(12345) # Note to self: Change the combo on my luggage
iris.mis <- prodNA(iris, noNA=0.1)
prop.table(table(is.na(iris.mis)))
```

```
## 
## FALSE   TRUE 
##   0.9    0.1
```

**Hmisc**

**missForest**

**MICE**

**AMELIA**

# Final thoughts

Take home messages:

1. Don't ignore missing data.
2. Impute sensibly and multiple times.

Here are some other reasons for missing data that we do not have time to cover. So I will leave you with some final thoughts.

- MI applies to subjects who have a general missingness pattern, i.e., they have measurements on some variables, but not on others.
- But, subjects can be lost to follow up due to death or other reasons (i.e., attrition).
- Here we have only baseline data, but not the outcome or other follow up data.
- If attrition subjects are eliminated from the sample, they can produce non-response or attrition bias.
- Use attrition weights.
  - Given a baseline profile, predict the probability that subject will stay and use the inverse probability as weight.
  - e.g., if for a given profile all subjects stay, then the predicted probability is 1 and the attrition weight is 1. Such a subject "counts once".
  - For another profile, the probability may be 0.5, attrition weight is $1/.5 = 2$ and that person "counts twice".
- For differential drop-out, or self-selected treatment, you can consider using Propensity Scores.

# On Your Own

**On Your Own**

1. For each distribution listed below, draw a random sample of $n = 100$ and delete 20% of the data completely at random and calculate the bias. State if the observed mean over- or under-estimates the true mean.

   a. $X \sim \mathcal{N}(10, 2)$
   b. $Y \sim \mathcal{Poisson}(10)$
   c. $B \sim \mathcal{Binomial}(10, .5)$
   d. $F \sim \mathcal{Binomial}(10, .9)$

   Does the effect of MCAR on the bias differ for the different distributions?

2. Repeat #1 but set the missing data mechanism to b NMAR, where $p$ is negatively correlated with the data.

Afifi 9.18. For the variables describing the average number of cigarettes smoked during the past 3 months (`SMOKEP3M`) and the variable describing the mother's education (`EDUMO`) in the Parental HIV data:

3. Determine the percent of missing data per variable, and combined (records with at least one of these variables missing)
4. For each variable describe a hypothetical scenario which might have led to these values being

   a. missing completely at random
   b. missing at random but not completely at random
   c. neither missing completely at random nor missing at random.