MISSING or INCOMPLETE DATA

A (fairly) complete review of basic practice

Don McLeish and Cyntha Struthers

University of Waterloo

Dec 5, 2015

Structure of the Workshop

Session 1	Common methods for dealing with missing data. Types of missing/incomplete data. When do we need to accommodate the missingness and how?
Session 2	Estimating functions and maximum likelihood estimation in the presence of missing data. Bayes' estimation, imputation and Rubin's multiple imputation (RMI).
Session 3	Missing/incomplete data in linear models. Regression with missing response or missing covariates. Considerations of efficiency. Longitudinal Data and example: CD4 counts.
Session 4a	The bias in the distribution of survivors. Models to accommodate survivorship. Gibbs Sampling, Data Augmentation and conditional models.
Session 4b	Sensitivity and Software.

Questions to be addressed

- When do we need to construct a model for missingness?
- Does it matter how we impute missing values? (For example, can we fill in the missing values with the mean of the available cases?)
- To what extent does our analysis depend on the assumptions made about the missing data mechanism? How sensitive are the inferences to the model assumptions?
- How does standard software deal with missing data, and when are the corresponding inferences valid?

QUESTIONS

- Have you analysed a data set with missing values?
- How many of you live in a rose coloured world of no missing data?

- Give a brief description of the data set.

QUESTION WWW.SOCRATIVE.COM

STUDENT LOGIN -> ROOM: CALGARY

1. Have you analysed a data set with missing values?

A: YES

B: NO

(finish)

2. IF YES, I

A: Ignored the cases with missing observations

B: Replaced the missing observations with a mean

C: Let my software[1] make the decision for me

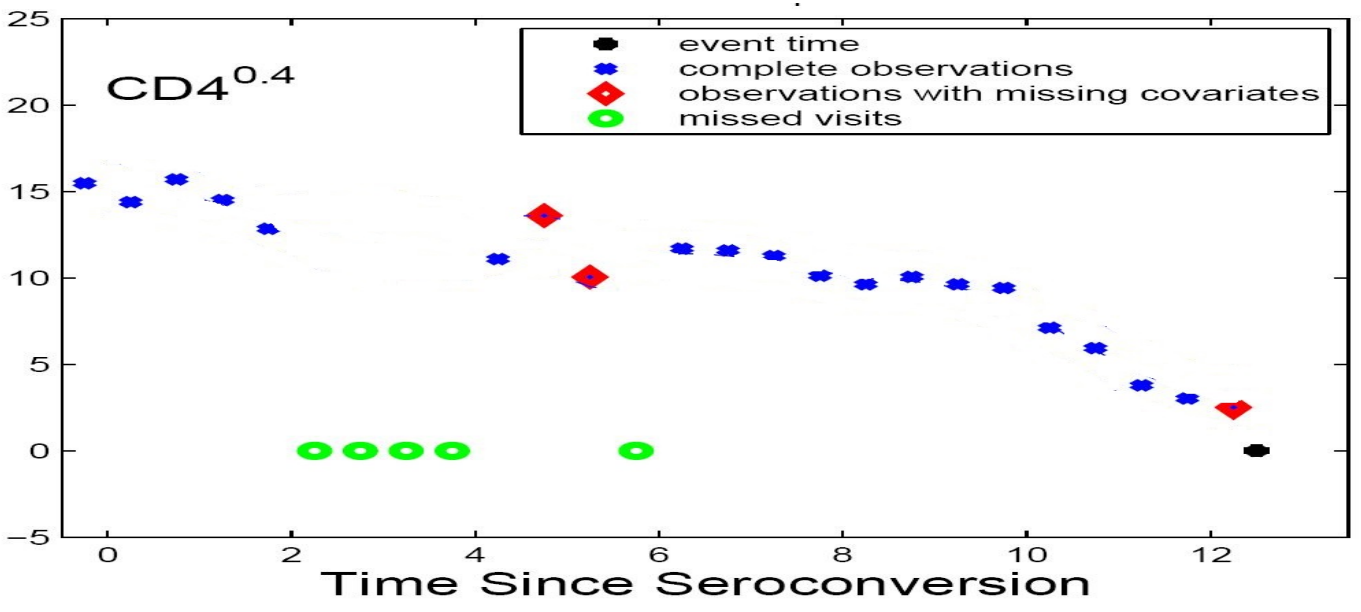
D: Berated my research assistant for not getting those values until he/she made something up

(finish)

[1] It's smarter than I am

Example 1

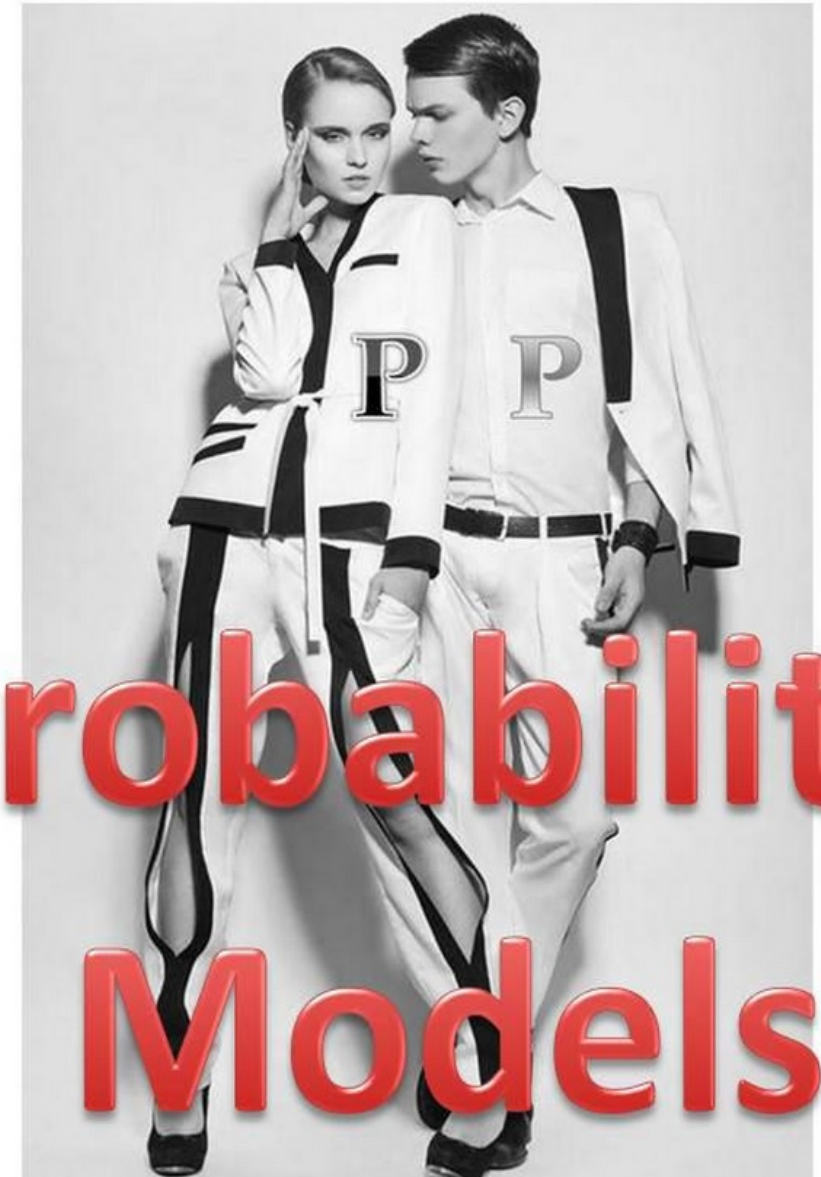
Multicenter AIDS Cohort Study (AIDS). Data collected on HIV+ men at semi-annual visits. Response variate was CD4 count. Covariates were time since seroconversion, smoking habits, number of sexual partners, etc. Response or some covariates missing at a visit. Some participants dropped out.



Example 2

R.J. Gladstone (1905). "A Study of the Relations of the Brain to the Size of the Head", Data: Brain weight (gms) and head size (cm^3) for 237 adults classified by sex.
Compare $Y_1 = \log(\text{head size})$ and $Y_2 = \log(\text{brain weight})$ for males (M) and females (F). (Y_1, Y_2) are available for 140. Y_2 is missing for the other 97.

A few things off the chest....



Probability Models

Missing (Incomplete) Data

- **What constitutes a missing value?**
  - Anything relevant that is not observed: coded as "undecided, don't know, NA, non-response, duh,...."
  - A value that, if observed, would simplify analysis (e.g. TSX closed on July 1)
  - For the "missing" values, does a "true value" even exist?
- **Does missing data matter?**
  - Are the data missing completely at random?
  - Why are the data missing? Is the missingness related to the response?
  - How do we deal with the missing data?
- Example: Canadian National Household Survey.
  - Adjusting for nonresponse was required when nonrespondents differed from respondents.

Always ask: ”Why is it missing?”

How common are missing data problems?

Reviews on reporting practices for missing data in various fields: (Van Buuren, 2012)

- clinical trials (Wood et al., 2004)
- cancer research (Burton and Altman, 2004)
- educational research (Peugh and Enders, 2004)
- epidemiology (Klebanoff and Cole, 2008)
- developmental psychology (Jeliĉić et al., 2009)
- general medicine (Mackinnon, 2010)
- developmental pediatrics (Aylward et al., 2010)
- finance (Kofman and Sharpe, 2000)

How common are missing data problems?

Overall picture from these reviews:

- The presence of missing data is often not explicitly stated in the text.
- Default methods like listwise deletion (complete case analysis) are used without mentioning them.
- Different tables are based on different sample sizes.
- Model-based missing data methods, such as maximum likelihood and multiple imputations are underutilized.

Missing data in clinical trials

Google "missing data" and "clinical": approximately 4 million results.
Missing data problems are common but are inadequately handled in clinical trials (Wood et al., 2004).

- In 71 papers in medicine, 89% had partially missing outcome data.
- In 37 trials with repeated outcome measures, 46% performed complete case analysis.
- Only 21% analysed sensitivity to the missing mechanism

Missing data in finance

Kofman and Sharpe.

Table 2 Treatment of incomplete data in finance: 1995–1999.

Journal	Articles acknowledging missing values	Missing value treatment			
		Listwise deletion	Regression imputation	Ad hoc imputation	Proxy imputation
JBF	67	56	5	5	3
JF	98	77	6	9	7
JFE	53	44	2	3	5
JFQA	20	18	–	1	2
RFS	19	10	3	7	1
All	257	205	16	25	18

This table presents information regarding the treatment of an incomplete data problem in those articles that acknowledge the presence of incomplete data. The journals include the *Journal of Banking and Finance* (JBF), vols. 19–23, the *Journal of Finance* (JF), vols. 50–54, the *Journal of Financial Economics* (JFE), vols. 37–52, the *Journal of Financial and Quantitative Analysis* (JFQA), vols. 30–34, and the *Review of Financial Studies* (RFS),

Common methods for dealing with missing data

A. Mean imputation
B. Listwise deletion (LD) or complete-case analysis
C. Pairwise deletion (PD) or available-case analysis
D. Regression imputation
E. Stochastic regression imputation
F. Inverse probability weighting (IPW)
G. Imputation (hot deck imputation, predictive mean matching, multiple imputation)

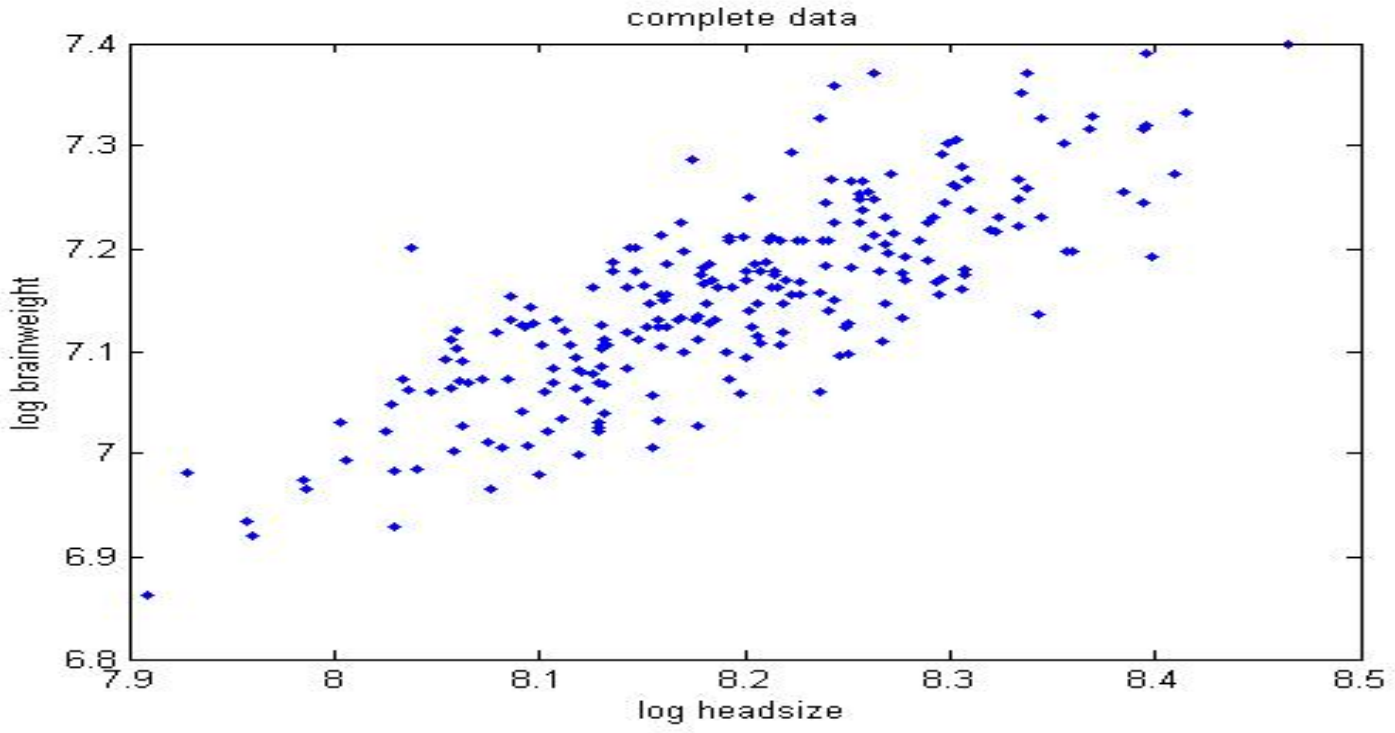
QUESTION: WWW.SOCRATIVE.COM (CALGARY)

Which of the above methods have you heard of or seen? What about

H. Fudge the data.

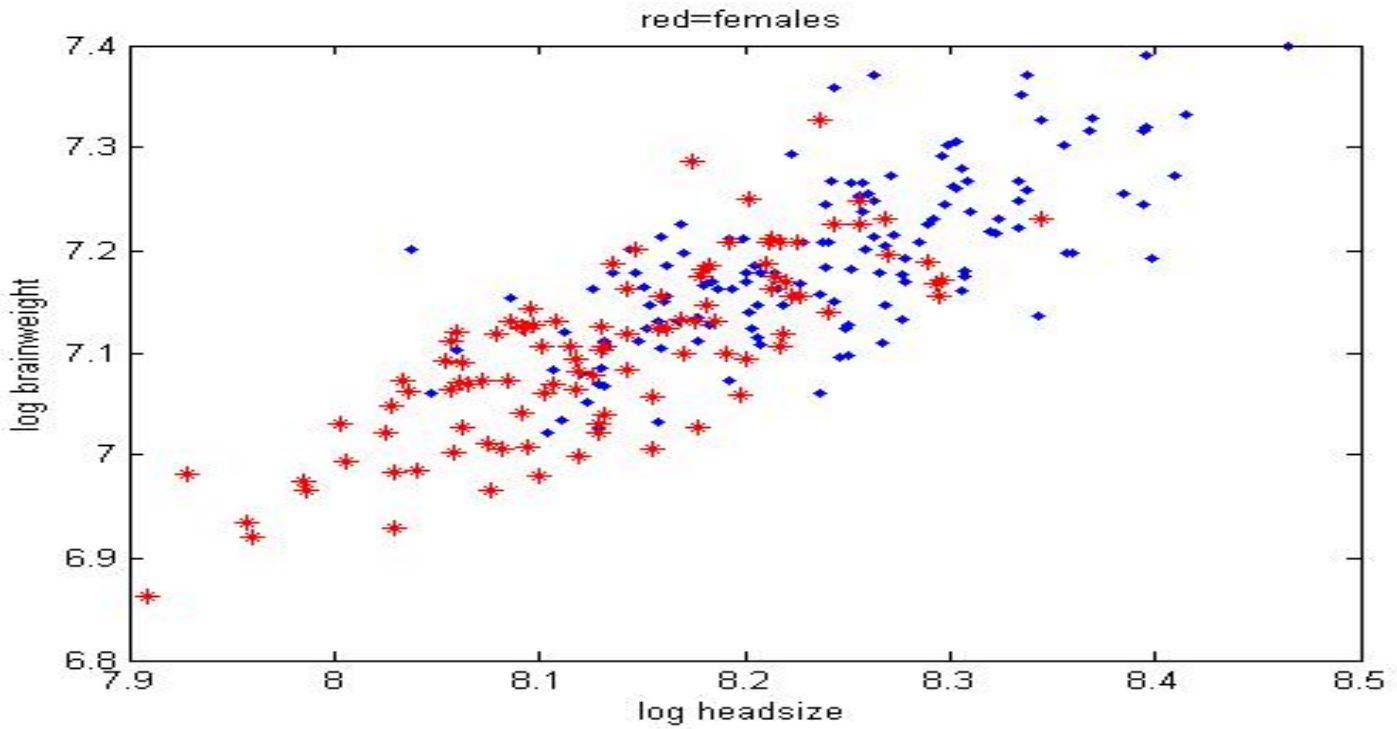
Example: Head size and brain weight

Compare $Y_1 = \log(\text{head size})$ and $Y_2 = \log(\text{brain weight})$ for males (M).



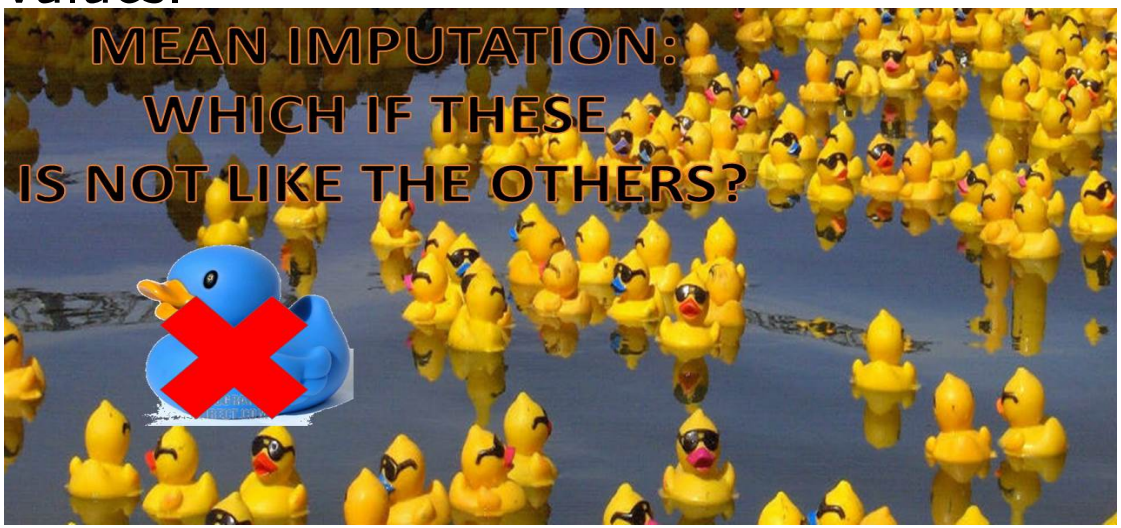
Example: Head size and brain weight

Compare $Y_1 = \log(\text{head size})$ and $Y_2 = \log(\text{brain weight})$ for males (blue) and females (red).
Create data set with Y_1 always observed and Y_2 missing for $\frac{97}{237} = 41\%$ of the cases. We use these data to illustrate common methods for dealing with missing data.



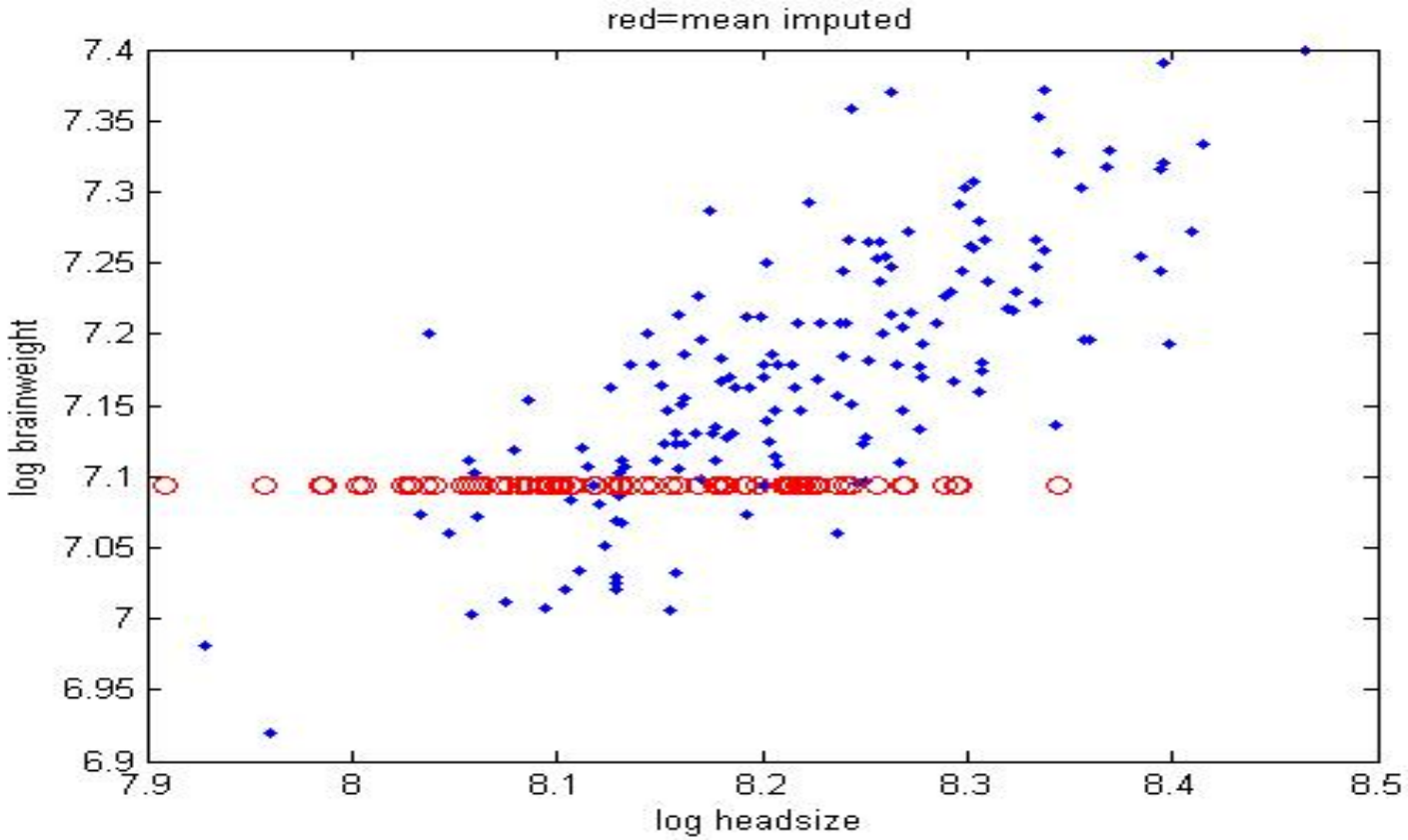
Methods for missing data: Mean imputation

Replace missing values by average of (relevant) observed values.



- Implicitly assumes the missing observations are a simple random sample from the same population.
- Underestimates the variance by a factor of $\frac{r-1}{n-1} (= \frac{139}{236} = 59\%)$ where
  $r =$ number of observed, $n - r =$ number missing.
- **Drawback:** coverage of confidence intervals is less than

Example: Mean imputation



	$E(Y_1)$	$SD(Y_1)$	$E(Y_2)$	$SD(Y_2)$	*corr*	No.
True Value	8.193	0.100	7.152	0.094	0.806	237
Estimate	8.193	0.100	7.143	0.085	0.701	140

Y_1 fully observed, 41% of Y_2 values missing.

QUESTION: WWW.SOCRATIVE.COM (CALGARY)

When would you be comfortable filling in missing values with the mean of the observed values?

A. When estimating a population mean for independent draws from a homogeneous population
B. When estimating a population variance for independent draws from a homogeneous population
C. When 32% of the data is missing on the National Household Survey.

(jeopardy)

Methods for missing data: Listwise deletion (LD) or complete-case analysis

Most statistical software simply excludes observations with any missing variable values from the analysis.

In R, this is done automatically for classical regression (data points with any missingness in the explanatory variates or response variate are deleted from the analysis).

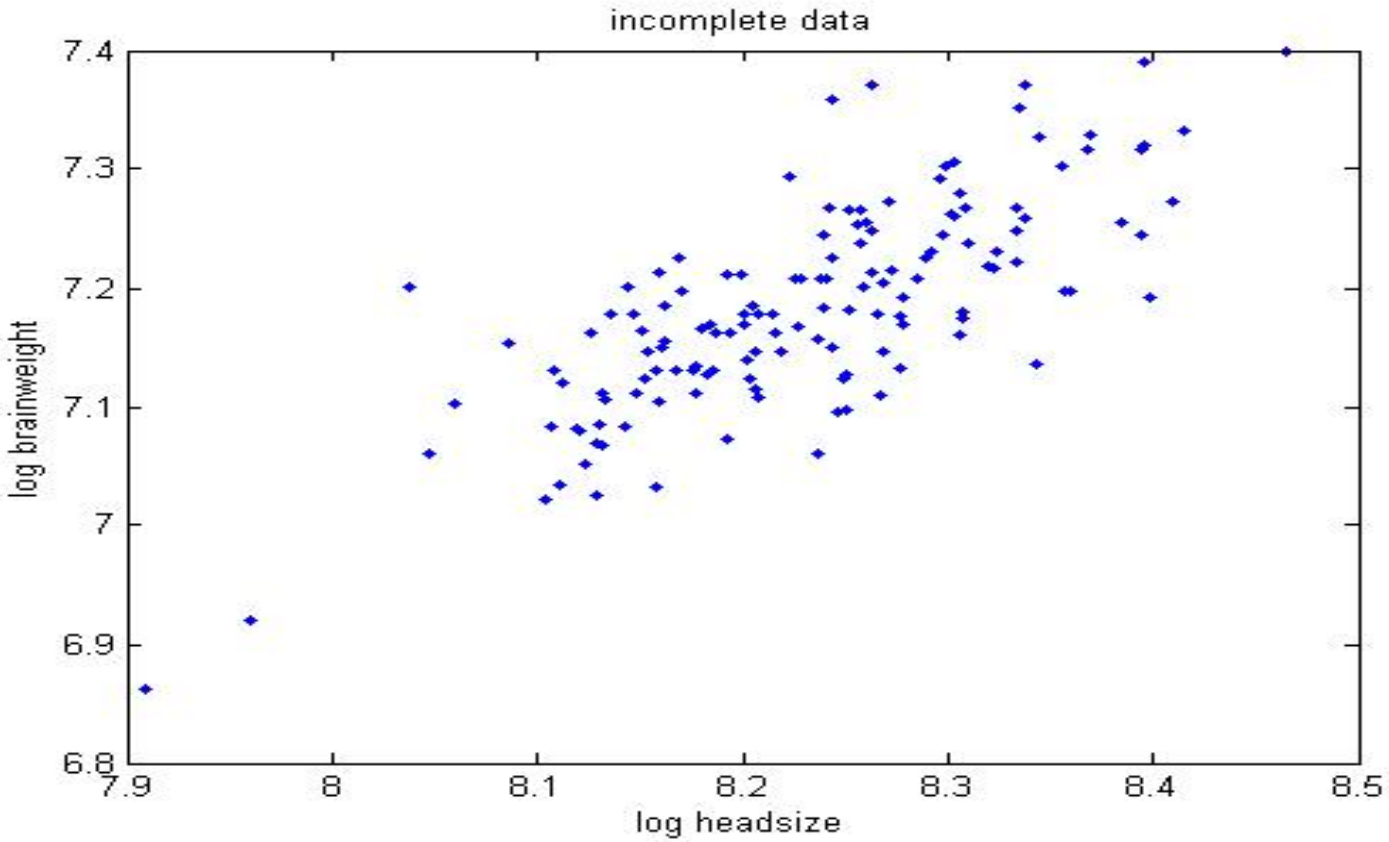
- **Is this a good idea?**
  1. If the units with missing values differ systematically from the completely observed cases, then the complete-case analysis will be biased.
  2. If there are many variates, then there may be very few complete cases.

Example:

? = missing. There are only 3 complete cases.

y_1	y_2	y_3	y_4	
?		**?**		
				$\longleftarrow$
	?			
				$\longleftarrow$
	?			
				$\longleftarrow$
			?	
?				

Example: Listwise deletion (LD) or complete case analysis



	$E(Y_1)$	$SD(Y_1)$	$E(Y_2)$	$SD(Y_2)$	*corr*	No.
True Value	8.193	0.100	7.152	0.094	0.806	237
Estimate	8.2314	0.092	7.185	0.087	0.764	140

Methods for missing data: Pairwise deletion (PD) or available-case analysis

Use all cases available for a given calculation.

No.	$\mathbf{y}_1$	$\mathbf{y}_2$	$\mathbf{y}_3$	$\mathbf{y}_4$
1	?		?	
2				
3		?		
4				
5		?		
6				
7				?
8	?			

In the above example the estimation of $Cov(Y_1, Y_2)$ would be based on observations $2, 3, 6, 7$ and estimation of $Cov(Y_3, Y_4)$ would be based on observations $2, 3, 4, 5, 6, 8$.

Methods for missing data: Pairwise deletion (PD) or available-case analysis

- Different aspects of a problem are studied with different subsets of the data.
- Pairwise deletion uses more data than listwise deletion but may it provide inconsistent conclusions (e.g. non-positive definite covariance matrix? degrees of freedom?)
- Estimates are still biased if the nonrespondents differ systematically from the respondents.

QUESTION: WWW.SOCRATIVE.COM (CALGARY)

How many of you have used

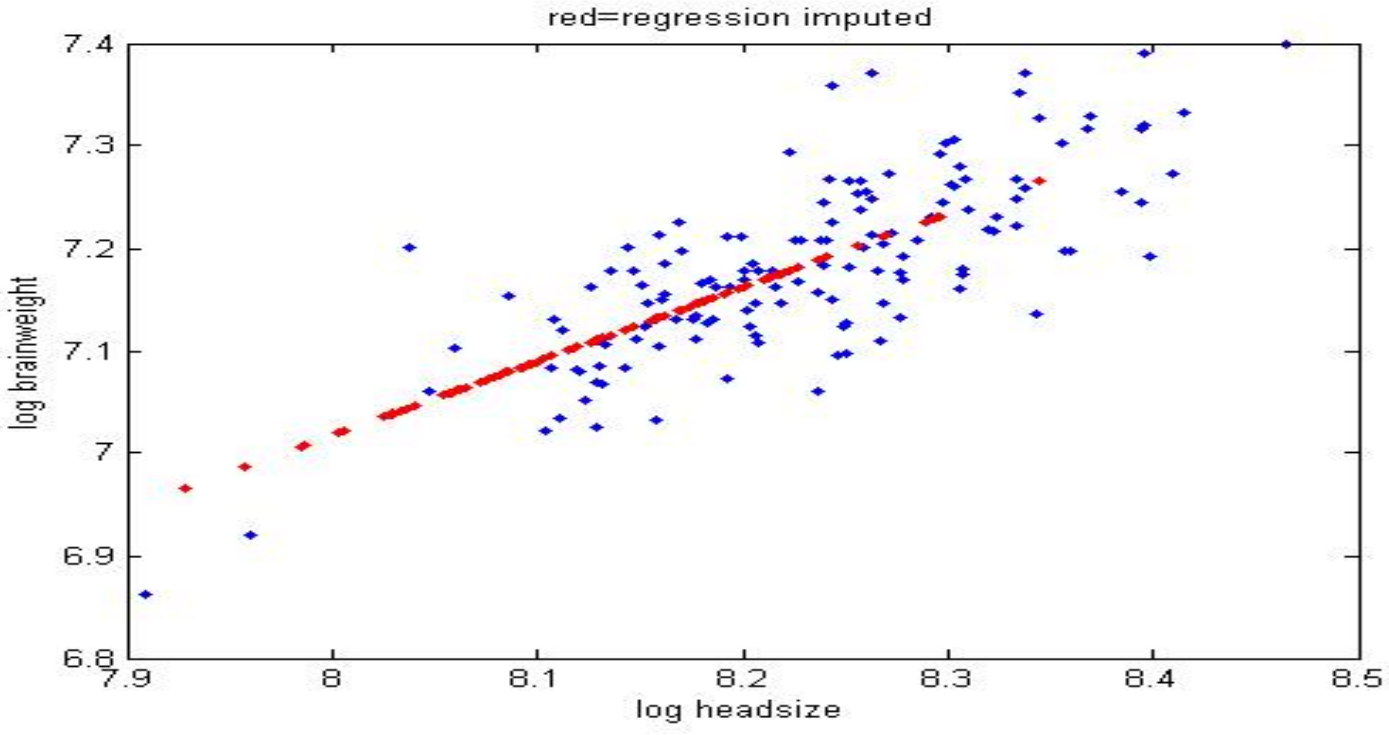
A. Listwise Deletion (Complete-cases)
B. Pairwise Deletion (Available-cases)?

- (Note that the names Listwise **Deletion** and Pairwise **Deletion** remind us that data have been deleted!)
- Did you feel bad?

Treatment of missing data: Regression imputation

- If values of Y_2 are missing, fit the regression for Y_2 given Y_1 to the complete data.
- Replace any unobserved value by its fitted value on the regression line.
- **Drawbacks:**
  - Any bias in the complete data parameters may influence results.
  - Coverage of confidence intervals is less than the nominal value.

Example: Regression imputation

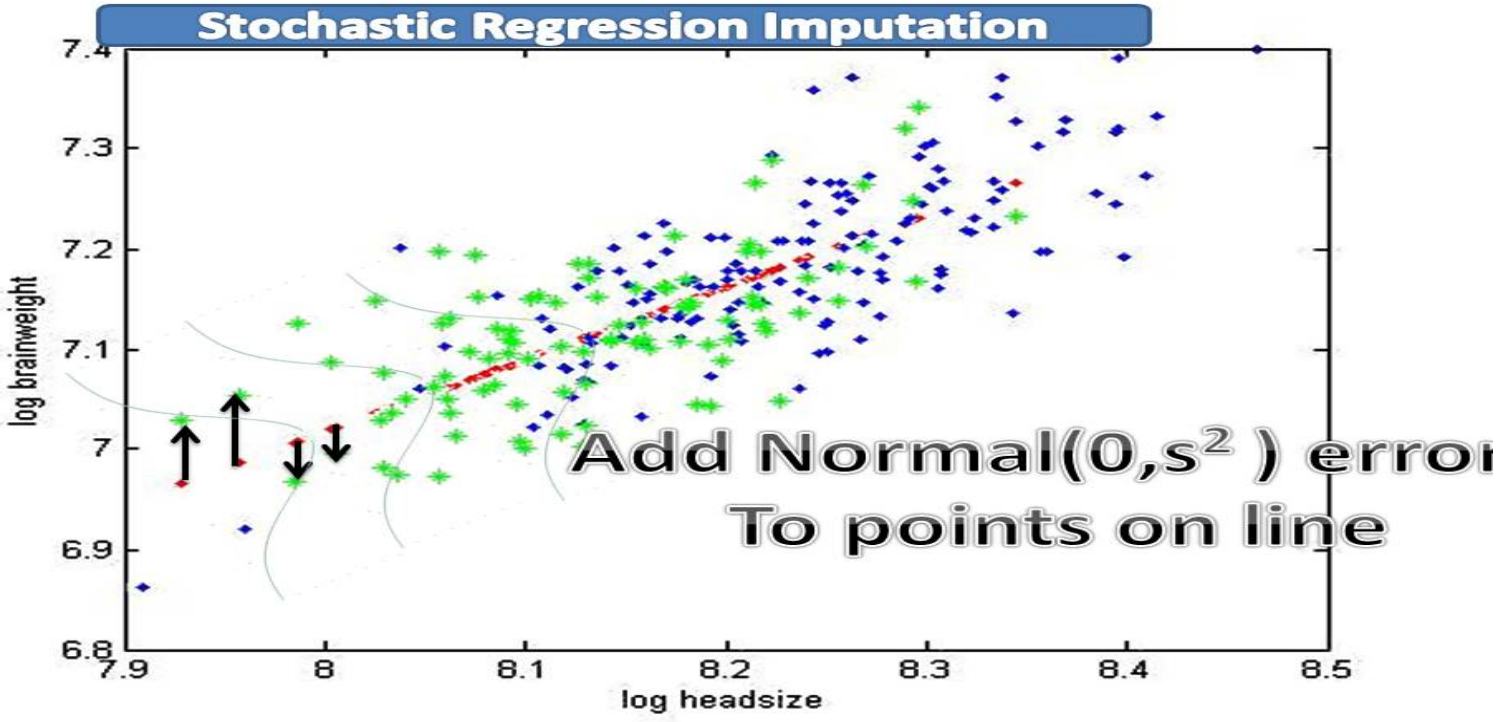


	$E(Y_1)$	$SD(Y_1)$	$E(Y_2)$	$SD(Y_2)$	*corr*	No.
True Value	8.193	0.100	7.152	0.094	0.806	237
Estimate	8.193	0.100	7.157	0.085	0.858	140

Methods for missing data: Stochastic regression imputation

- If values of Y_2 are missing, fit the regression for Y_2 given Y_1 to the complete data.
- Replace an unobserved value by an imputed value which is obtained by adding a $N\left(0, s^2\right)$ observation to the fitted value from the regression.
- Close to ML estimate based on the observed data.
- Assumes the regression model for Y_2 given Y_1 is correct.
- Does not incorporate the uncertainty in the regression parameters.

Example: Stochastic regression imputation



	$E(Y_1)$	$SD(Y_1)$	$E(Y_2)$	$SD(Y_2)$	*corr*	No.
True Value	8.193	0.100	7.152	0.094	0.806	237
Estimate	8.193	0.100	7.158	0.089	0.770	140

Methods for missing data: Inverse probability weighting (IPW)

First used in sampling. See: Horvitz and Thompson (1952).

- IPW addresses the bias in complete case or available case analysis due to non-response by re-weighting the *observed* responses to restore representativeness.
- It uses a model to predict the nonresponse in that variable using all the observed variables.
- The inverse of predicted probabilities of response from this model are used as weights to make the complete case sample representative.

Example: Inverse probability weighting (IPW)

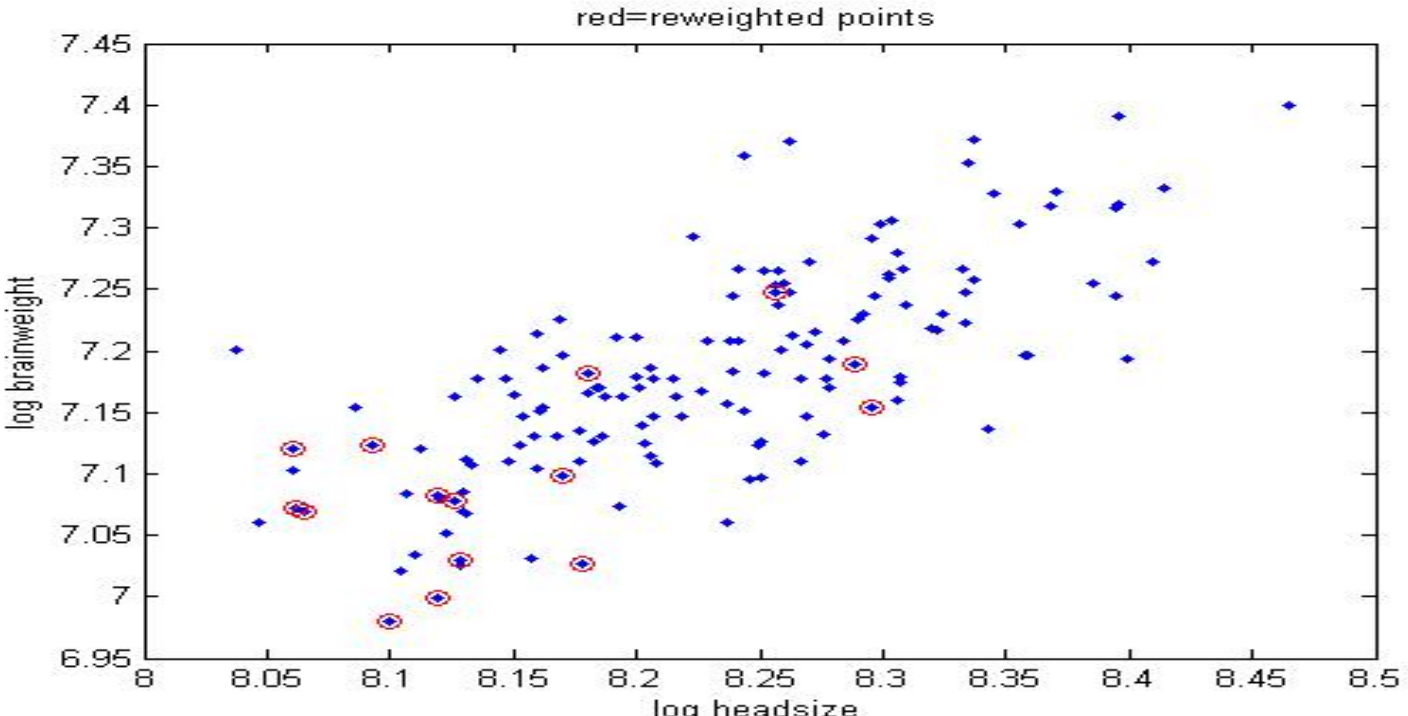
Suppose the probability of observing $Y_2 = \log(\text{brain weight})$ is a function $\pi(X)$ of the known covariate $X =$ sex (M or F). Define $R = 1$ if Y_2 observed, and $R = 0$ if Y_2 is not observed. Then

$$E\left[\frac{R}{\pi(X)}Y_2\right] = E\left\{E\left[\frac{R}{\pi(X)}Y_2|X\right]\right\} = E\left(Y_2|X\right).$$

- $\frac{R}{\pi(X)}Y_2$ is an unbiased estimator of $E(Y_2)$ so we can estimate it using the average of the observed values of $\frac{R}{\pi(X)}Y_2$.
- IPW is more complicated when more than one variate is missing for a case.
- If the predicted probabilities $\pi(X)$ are close to 0 then issues related to apparent bias and large variance arise. (More on this later.)

Example: Inverse probability weighting (IPW)

- Suppose the probability a female brain weight is observed is 0.2 so $\pi(X = \text{“}F\text{”}) = 0.2$ and $\pi(X = \text{“}M\text{”}) = 1$.
- Attach a weight of $w_i = \frac{1}{0.2} = 5$ to a female observation and attach a weight of $w_i = 1$ to a male observation. Estimator of mean log(brain weight) is: $\frac{\sum w_i Y_{2i}}{\sum w_i}$.



How do we know whether to adjust for the missingness?

- If we do not adjust, we risk introducing bias into the estimator.
- If we do adjust we may increase the variance of the estimator and therefore the mean squared error (MSE).

Example: The price of IPW

If inverse probability weighting (IPW) is used instead of listwise deletion (LD), what is the cost in terms of MSE? Suppose there are several strata and:

- P (observation is in stratum j) $= P(X_i = j) = q_j$.
- Responses Y_i, $i = 1, ..., n$ independent.
  P (Y observed $|X = j) = P(R = 1|X = j) = \pi_j$.
- Mean of stratum j: $E(Y_i|X = j) = \mu(j)$.
- Common variance for each stratum:
  $Var(Y_i|X = j) = \sigma^2$.
- Compare 2 estimators of the mean:

$$\hat{\theta}_{IPW} = \frac{1}{n} \sum_{i=1}^{n} \frac{R_i}{\pi_i} Y_i \quad \text{and} \quad \hat{\theta}_{LD} = \frac{\sum_{i=1}^{n} R_i Y_i}{\sum_{i=1}^{n} R_i}.$$

Example: To B[2] or not to B (through the eyes of MSE)

Compare the MSE for the estimators

$$\hat{\theta}_{IPW} = \frac{1}{n}\sum_{i=1}^{n}\frac{R_i}{\pi(X_i)}Y_i \quad \text{and} \quad \hat{\theta}_{LD} = \frac{\sum_{i=1}^{n} R_i Y_i}{\sum_{i=1}^{n} R_i}$$

Assume 2 strata of equal size ($P(X_i = 1) = P(X_i = 2) = \frac{1}{2}$).

Difference between stratum means $= 2\Delta\sigma$.

$P(\text{being observed in stratum } j) = \pi_j$, $j = 1, 2$.

$\hat{\theta}_{IPW}$ is unbiased so MSE = variance.

$\hat{\theta}_{LD}$ is a biased estimator if $\pi_1 \neq \pi_2$.

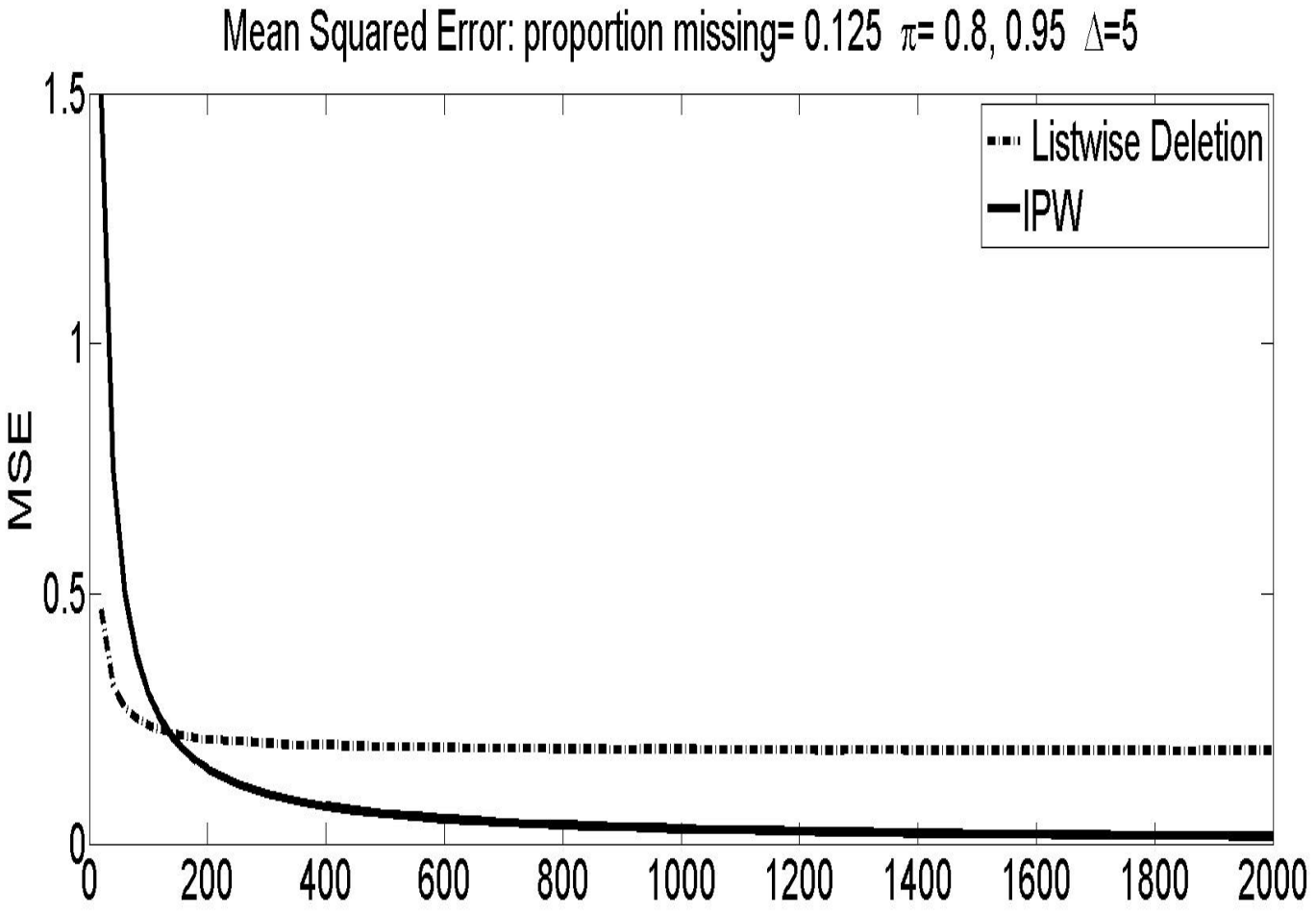
---

[2] Bias

Example: To Bias or not to Bias

$\hat{\theta}_{IPW} = \frac{1}{n}\sum \frac{R_i}{\pi(X_i)} Y_i$ and $\hat{\theta}_{LD} = \frac{\sum R_i Y_i}{\sum R_i}$

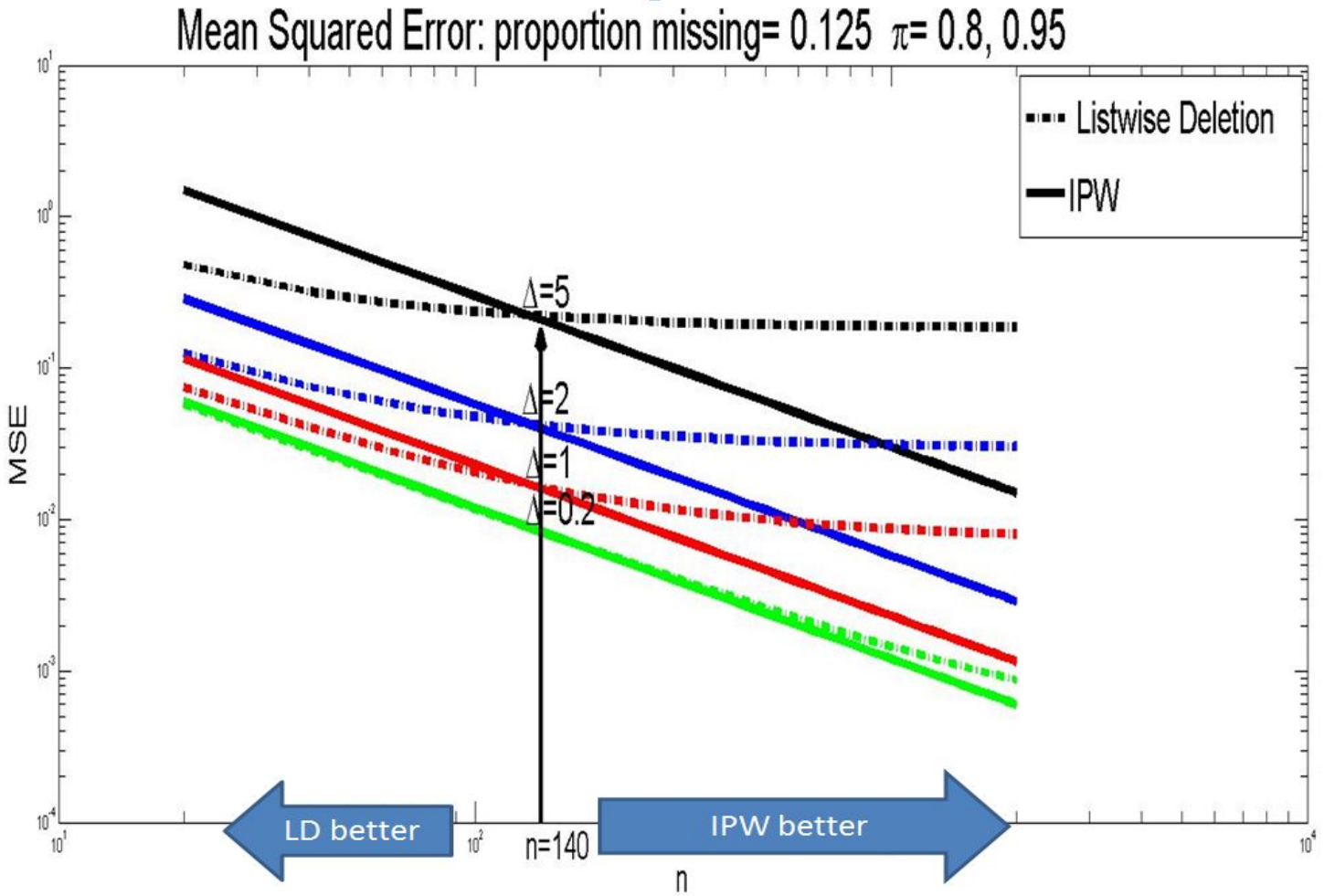
$\Delta = 5$, $\pi_1 = 0.8$ and $\pi_2 = 0.95$



Example:n versus MSE (log scales)

$\hat{\theta}_{IPW} = \frac{1}{n}\sum \frac{R_i}{\pi(X_i)} Y_i$ and $\hat{\theta}_{LD} = \frac{\sum R_i Y_i}{\sum R_i}$

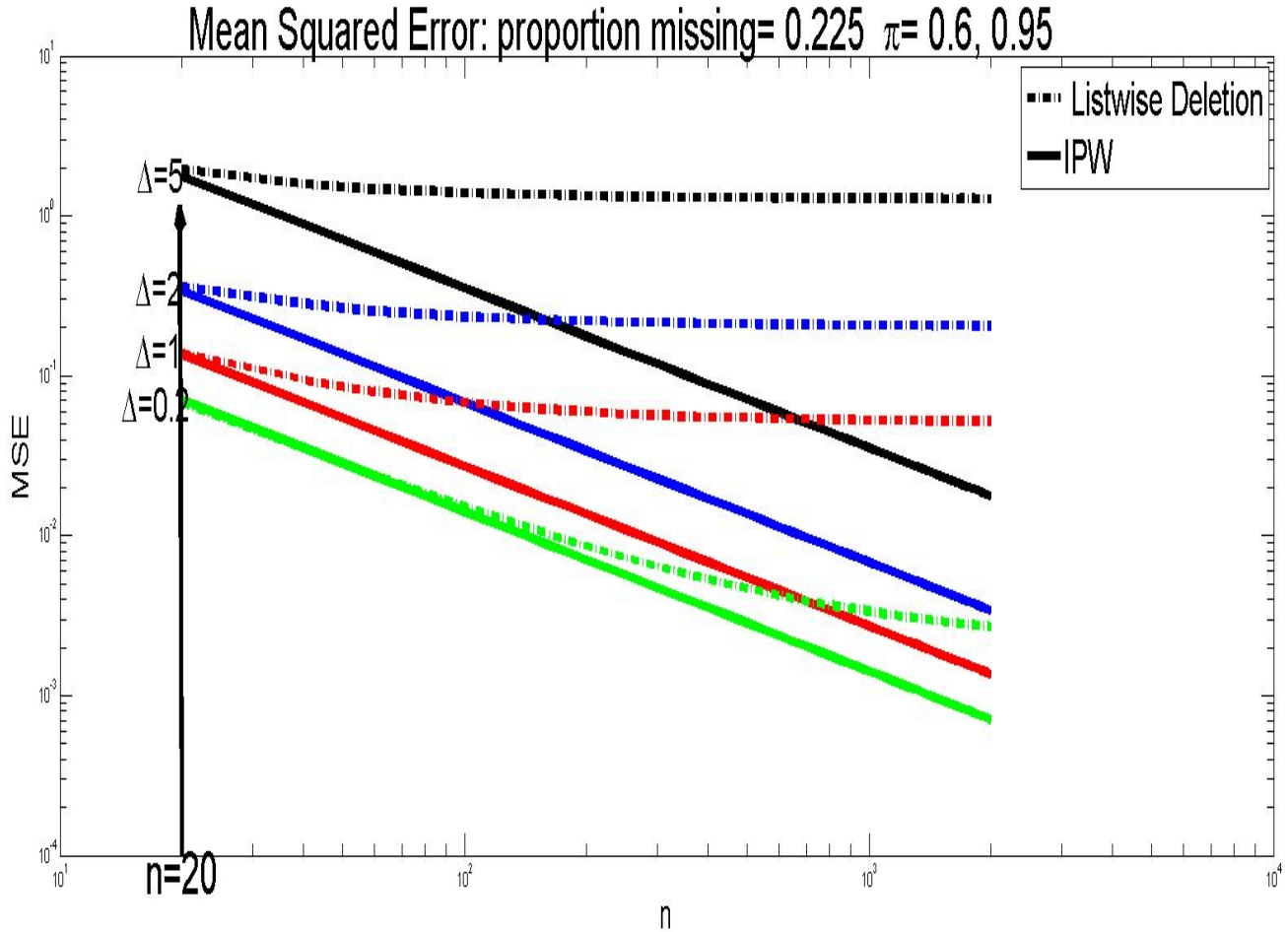
$\pi_1 = 0.8$ and $\pi_2 = 0.95$



Example: n versus MSE (log scales)

$\hat{\theta}_{IPW} = \frac{1}{n}\sum \frac{R_i}{\pi(X_i)} Y_i$ and $\hat{\theta}_{LD} = \frac{\sum R_i Y_i}{\sum R_i}$

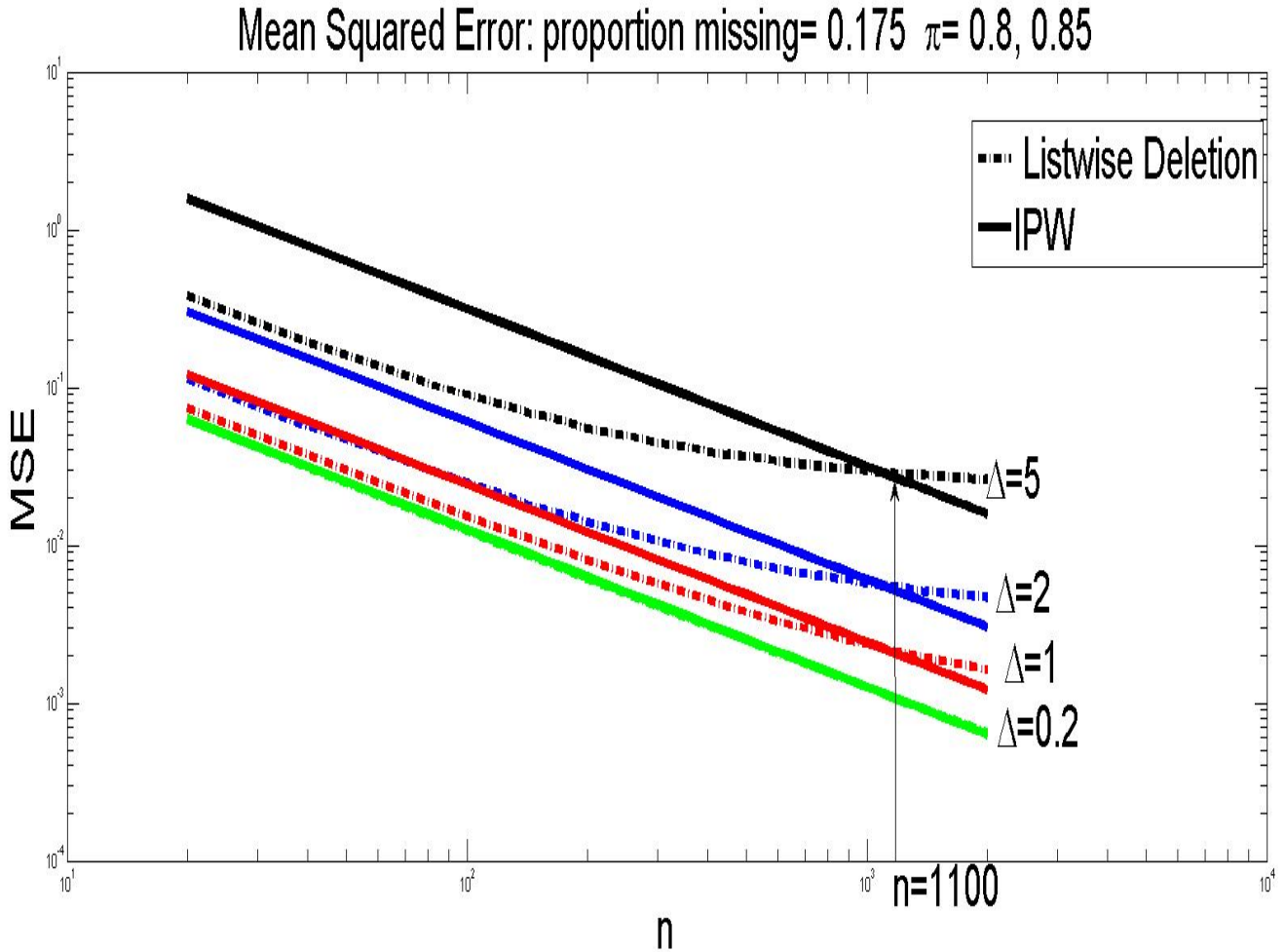
$\pi_1 = 0.6$ and $\pi_2 = 0.95$



Example: n versus MSE (log scales)

$\hat{\theta}_{IPW} = \frac{1}{n}\sum \frac{R_i}{\pi(X_i)} Y_i$ and $\hat{\theta}_{LD} = \frac{\sum R_i Y_i}{\sum R_i}$

$\pi_1 = 0.8$ and $\pi_2 = 0.85$



Example: To bias or not to bias

Conclusion:

- LD (Complete Cases Analysis) has smaller MSE than IPW for modest differences in the probability of response for different strata, and for small sample sizes (< 100).
- IPW should be used in situations in which a lack of bias is more important than MSE.

Methods for missing data: Imputation

- **Single imputation:**

Substitute a single value for each missing value. (Mean imputation and regression imputation are examples of single imputation.)

- **Hot deck imputation:**
  - For each missing response determine a set of respondents that have similar values on a set of matching variables (e.g. sex, age, marital status, etc.).
  - Fill in the missing response with the response from a randomly chosen unit from this set.

Methods for missing data: Predictive mean matching

- Predictive mean matching is an example of a hot deck method.
- It is similar to stochastic regression imputation except that missing values are imputed from a set of *nearest neighbour* observed values whose predicted values are closest to the predicted value for the missing value. See: Heitjan and Little (1991); Schenker and Taylor (1996).
- Predictive mean matching is a non-parametric alternative to stochastic regression imputation. It is more appropriate if the errors are not normality distributed. (See Horton and Lipsitz 2001, p. 246)

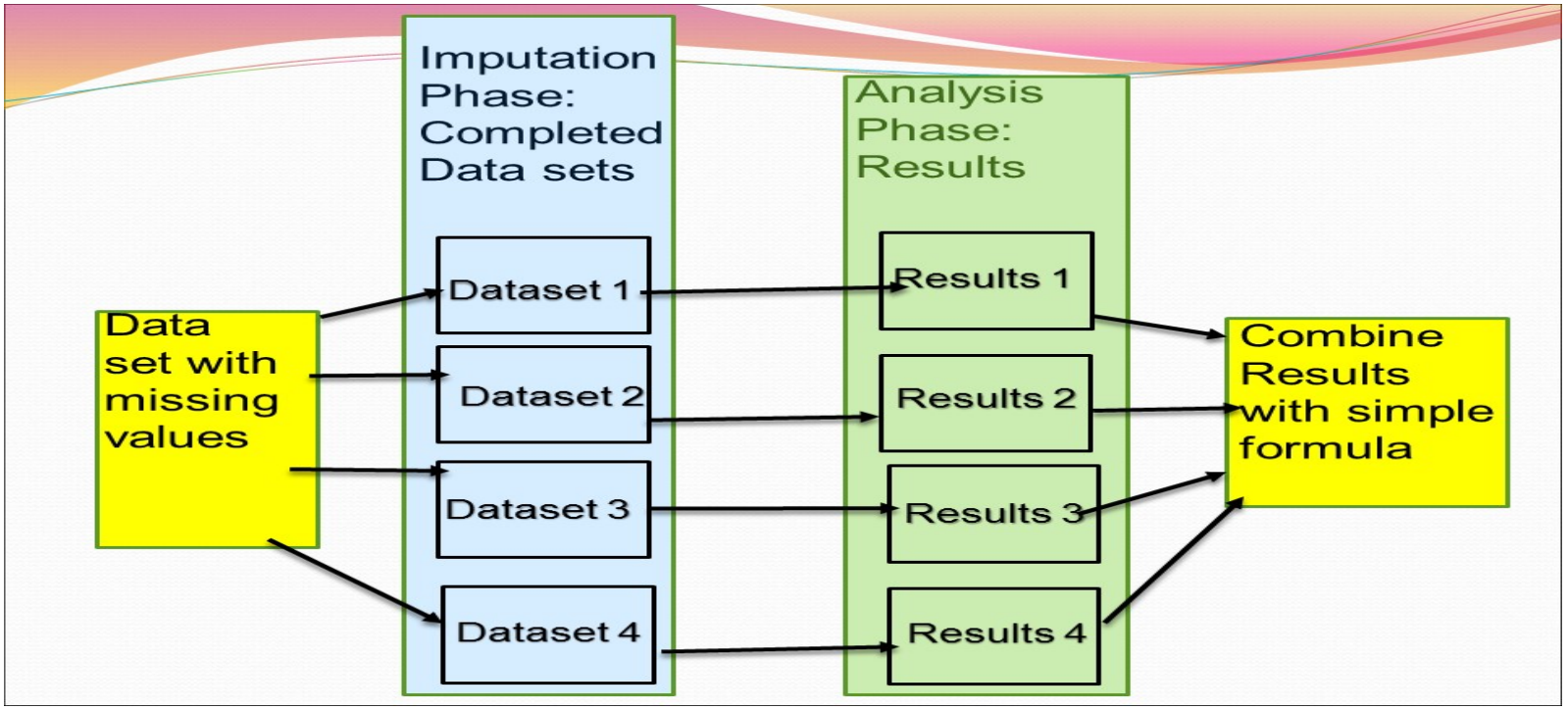
Methods for missing data: Predictive mean matching

1. Draw new parameters from the Bayes posterior distribution of the parameters (more on this later).
2. Find a predicted value for each missing value using these parameters.
3. Determine a set of k_0 (pre-specified) observations whose corresponding predicted values are closest to the one to be imputed.
4. Impute the missing value using a value drawn randomly from these k_0 observed values.
5. Small k_0 tends to increase the correlation among the multiple imputations for the missing observation and results in higher variability of point estimators in repeated sampling. A large k_0 may lead to biased estimators.

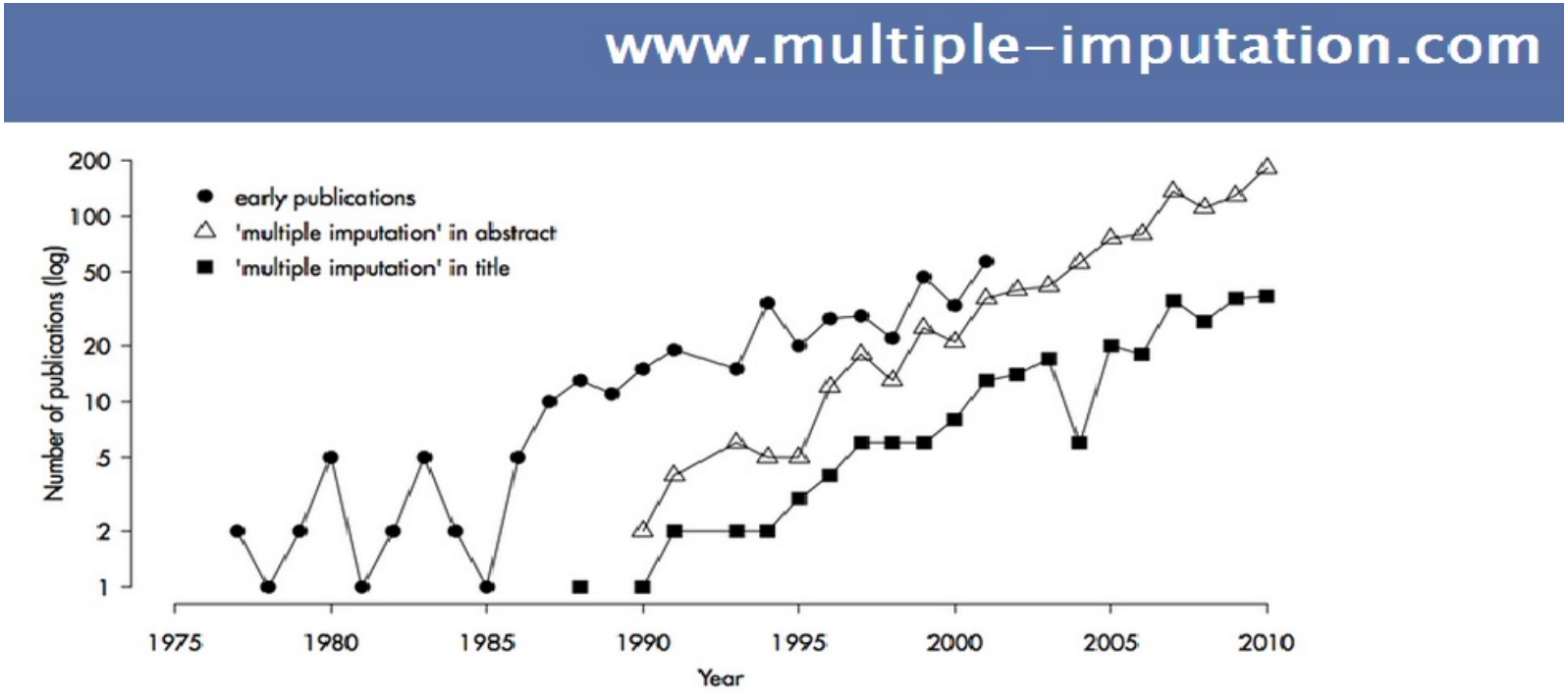
Multiple imputation (RMI)

- A random sample is drawn from *some* distribution to replace the missing values.
- This is done m times to form m completed datasets.
- Each of the m completed datasets are analysed and the results are combined using Rubin's Rules.
- Often a Bayesian approach is used in which the sample is drawn from a predictive distribution for the missing observations. (More on this later.)
- Not all multiple imputation software uses a Bayesian approach.

Multiple imputation (RMI)



Multiple imputation (RMI)



- Remaining issues - All important
  1. From what distribution do we draw the imputed values?
  2. How do we combine the analyses of the completed datasets?
  3. Can we separate (1) and (2)?

Imputation and black boxes

Statistics, a Black Box? MI... a grey box? (# shades?)

Imputation and black boxes

- Models used by the imputer and analyst need some degree of agreement.

Imputation and black boxes

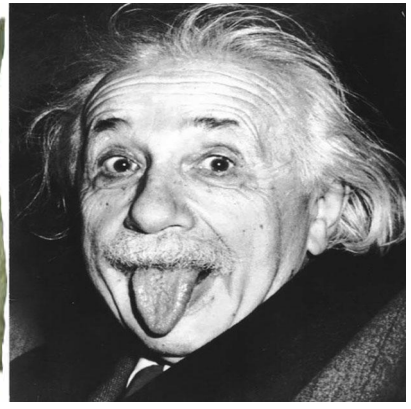
- Moreover, the water around RMI is somewhat muddied by an extensive and controversial literature.

THERE ARE NO BLACK BOXES![3]

Moral: Whenever possible use a maximum likelihood or Bayes approach. (More on this in next session.)

[3] foolproof ones

A method to my madness



...and assumptions in my methods.

Every statistical analysis has implicit assumptions. Make these explicit.

Are you making assumptions about

- sample size (large?)
- Independence?
- underlying parametric joint distribution (nonparametric, semiparametric)?

Assumptions about missingness

In addition: What assumptions are reasonable for

- why certain values are missing?
- any independence of missing data mechanism from other variables?
- relationship between missingness and response/covariates?
- Patterns of missingness... e.g. dropouts?

Types of missingness: Notation (standard)

To discuss types of missingness it is convenient to introduce the following (standard) notation of missing data:

- Y is the matrix (vector) of the complete data which is not fully observed
- R is the corresponding matrix (vector) of indicator variables whose elements indicate whether an observation in Y is missing ($R = 0$) or observed ($R = 1$).
- Y_{obs} is the observed part of Y
- Y_{mis} is the missing part of Y
- $Y = (Y_{obs}, Y_{mis})$

QUESTION

- How many of you recognize the term MCAR?

Types of missingness: Missing completely at random (MCAR)

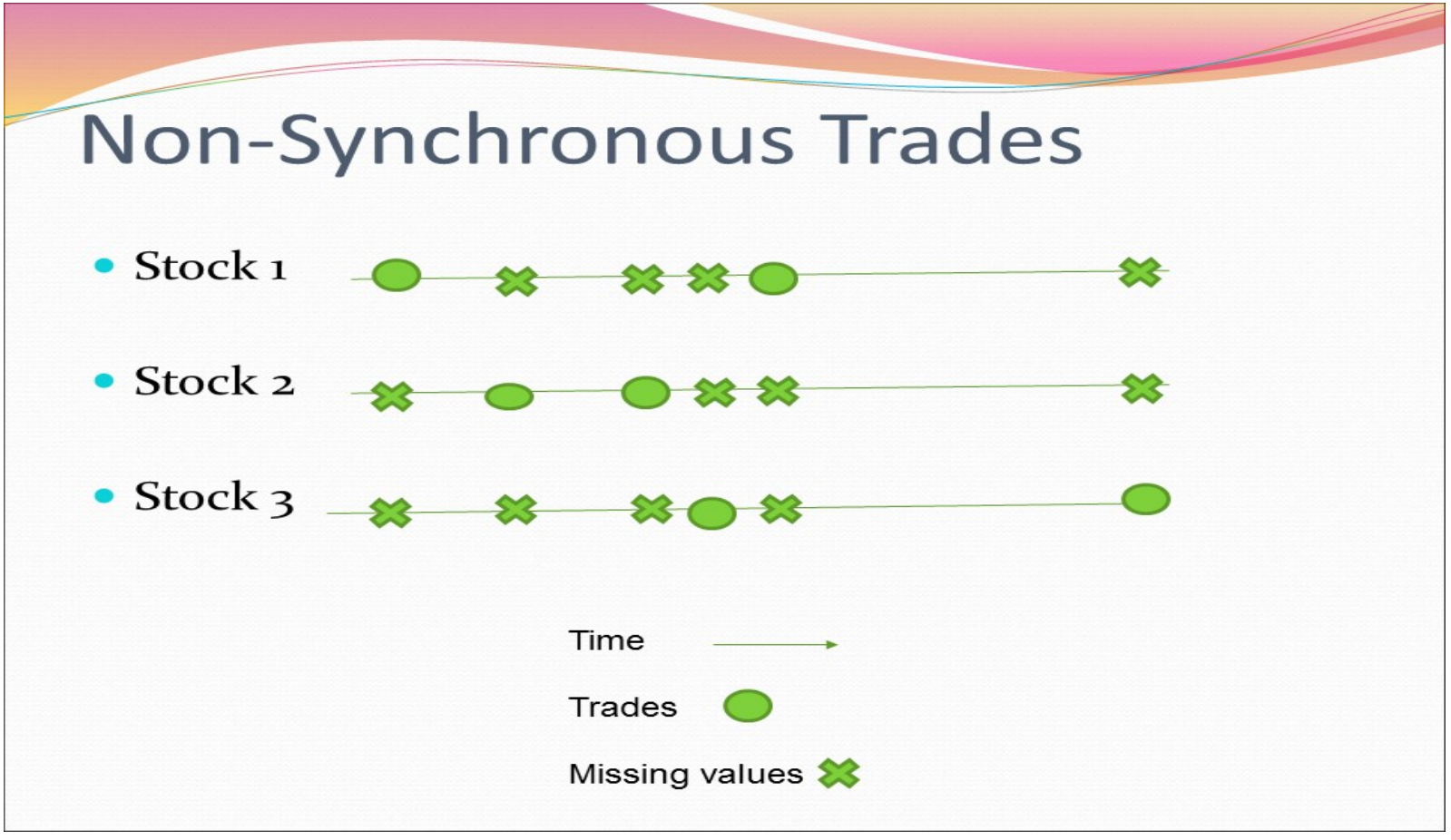
- Missing completely at random (MCAR):

$$P(R|Y_{obs}, Y_{mis}) = P(R)$$

  Missing values are randomly distributed across all observations.
  No systematic differences between the missing and observed values.
- **Example:** Measurements may be missing because of equipment failure.

Example: Missing completely at random (MCAR)

- **Example:** Non-synchronous trades. On July 1, TSE500 is MCAR, on July 4, DJIA is MCAR.



QUESTION

- How many of you recognize the term MAR?
- Is MAR a restaurant in Iceland?

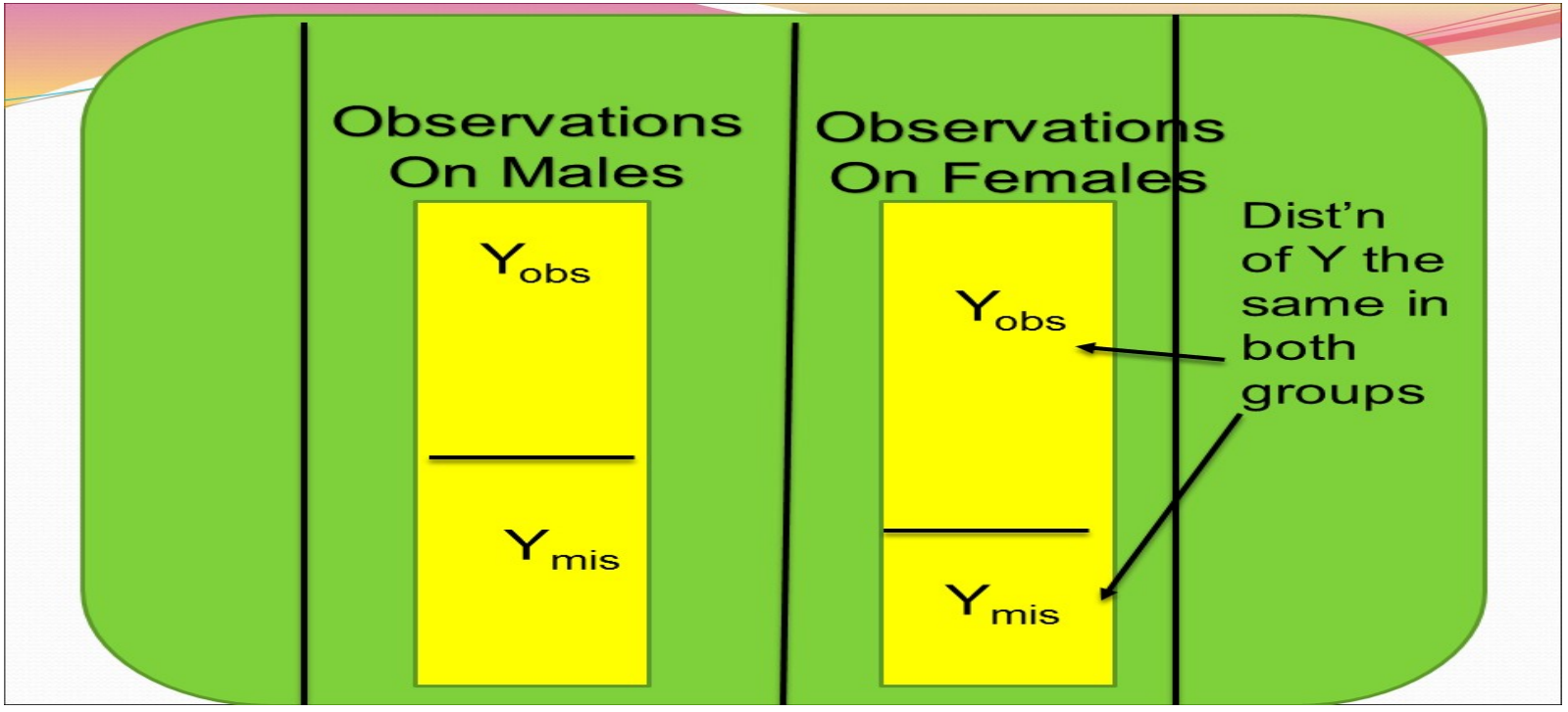
Types of missingness: Missing at random (MAR)

$$P(R|Y_{obs}, Y_{mis}) = P(R|Y_{obs})$$

- Distribution of R does not depend on the true value of the missing variable Y but it might depend on values of other observed variables (Y_{obs}).
- For example, missing values may be randomly distributed within observable strata (for example the probability of missing might depend on sex-an observed characteristic).
- MCAR is a special case of MAR.
- **Ignorable missingness:** MAR and parameters of $P(R|Y)$ distinct (independent) of parameters of $P(Y)$

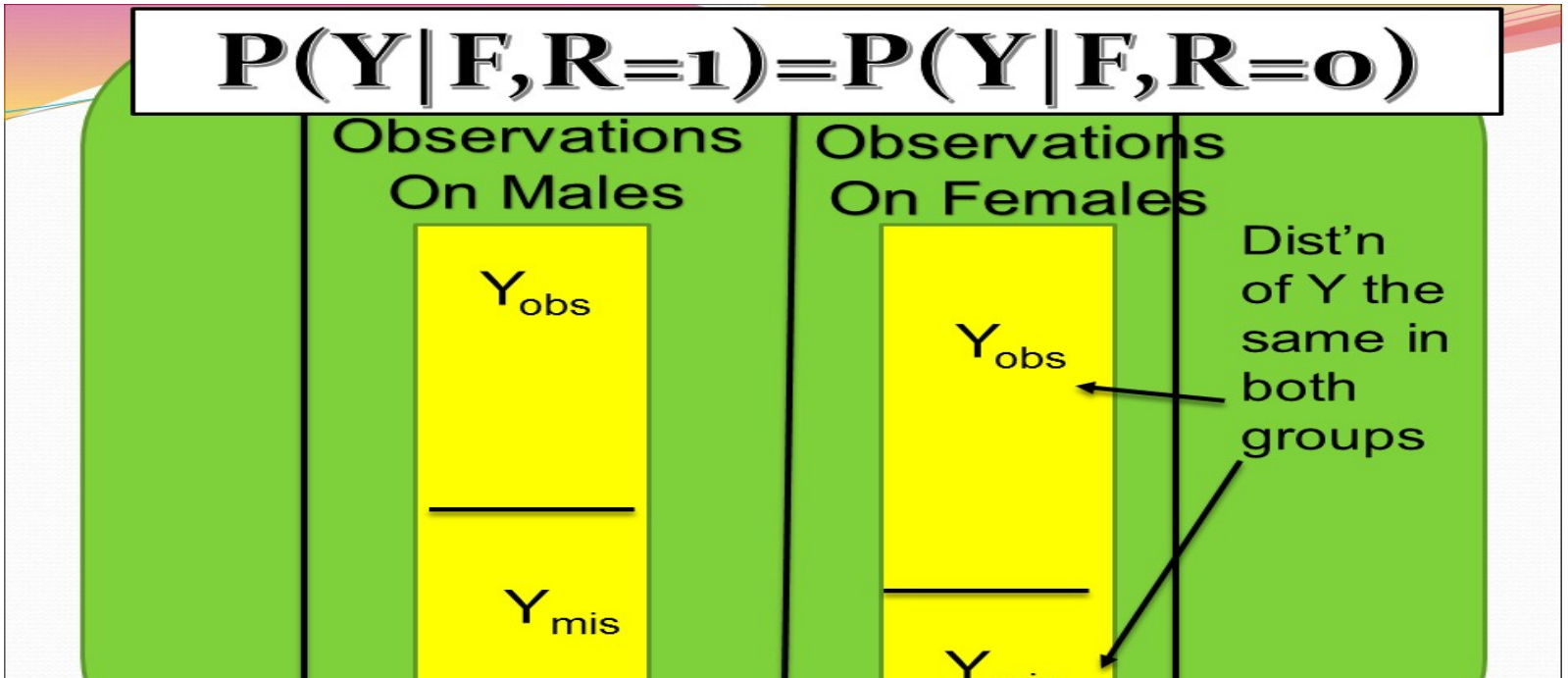
Example: Missing at random (MAR)

In a study on $Y =$ blood pressure, suppose sex of subject is always observed. MAR holds if for females the distribution of Y is the same for $R = 1$ (observed) and $R = 0$ (missing) and similarly for males.



Types of missingness: Missing at random (MAR)

Any systematic difference between missing and observed Y values can be explained by information obtainable from the observed data. For example, missing blood pressure measurements may be lower than observed blood pressures but this difference can be explained by the difference in the distribution of blood pressures between the sexes.



Question. www.socrative.com (calgary)

For what situation is LD unbiased?

A. When estimating a population mean for independent draws from a homogeneous population
B. When estimating a population variance for independent draws from a homogeneous population
C. When observations are missing completely at random.
D. Always, why wouldn't it be?

QUESTION

- How many of you recognize the term MNAR?
- Is it museum of arts in Romania?

Types of missingness: Missing not at random (MNAR)

After the observed data are taken into account, systematic differences between responders and non-responders remain.

MNAR requires a joint model for (Y, R) (more on this later).

Examples: Missing not at random (MNAR)

- The 40% non-responders to the National Household Survey in PEI may differ from responders of similar age/sex with respect to variates of interest.
- In a study of income, some people may be more/less likely to reveal their salaries (Ontario Sunshine List).
- In a study of the effectiveness of a medication to lower blood pressure, people with high blood pressure may be more likely to miss clinic appointments because they have headaches which would result in fewer observations for patients with higher blood pressure.
- In a study of a treatment to reduce depression, some patients may dropout because they believe the treatment is not effective or it is causing side effects.

Question www.socrative.com (calgary)

If the missingness indicator R has a distribution that does not involve the parameter of interest θ, and the distribution of the complete data $f_\theta(y)$ does, then we can ignore the distribution of R in obtaining inferences from incomplete data.

A. TRUE
B. FALSE
C. TRUE AND FALSE
D. (My answer is missing - impute it yourself.)

Biases in Naive Imputation methods for Bivariate Normal Distribution:

"-B"=negative bias, B=pos or negative bias. $MCAR \Rightarrow MAR$

Method	Missing	μ_x	σ_x^2	μ_y	σ_y^2	$\beta_{y\|x}$	$\sigma_{y\|x}^2$	$\beta_{x\|y}$	$\sigma_{x\|y}^2$
LD	Y_{MAR}	0	0	B	-B	0	0	B	-B
LD	X_{MAR}	B	-B	0	0	B	-B	0	0
PD	Y_{MAR}	0	0	B	-B	0	0	B	-B
PD	X_{MAR}	B	-B	0	0	B	-B	0	0
REG	Y_{MAR}	0	0	0	-B	0	0	B	-B
REG	X_{MAR}	0	-B	0	0	B	-B	0	0
St Reg	Y_{MAR}	0	0	0	0	0	0	0	0
St Reg	X_{MAR}	0	0	0	0	0	0	0	0
LOCF	Y_{MCAR}	B	B	B	B	B	B	B	B
Indic	X_{MCAR}	B	B	B	B	B	B	B	B