Lecture 1 - issues raised by missing data, a systematic approach, and missingness mechanisms

Multiple imputation for missing data

Jonathan Bartlett (thestatsgeek.com)

Oslo, November 2019

Missing data - what's the big deal and a systematic approach

Missingness mechanisms

Course aims

- Understand the effects of missing data on statistical analyses
- Learn about the assumptions under which simple methods for handling missing data are valid
- Learn about principled statistical methods for handling missing data, specifically multiple imputation

Missing data - what's the big deal and a systematic approach

Why is this necessary?

- Missing data commonly arise in empirical research.
- They cause a loss of information, and arguably more importantly, may introduce bias into inferences.
- They are often inadequately handled in both observational and experimental studies.
- For example, (Karahalios et al. 2012) reviewed the reporting and handling of missing data in longitudinal measurements in cohort studies.
- They found that reporting of missing data was inconsistent and inappropriate statistical methods continue to be used (in this field at least).
- Scientific journals and bodies increasingly recognise the importance of careful handling of missing data.

Missing data in trials - the problem and its prevention

- A US National Research Council (NRC) report was recently published on the prevention and treatment of missing data in trials (Council 2010; Little et al. 2012).
- They noted that missing data have seriously compromised inferences from clinical trials in the past.
- They concluded that the assumption that analysis methods can compensate for missing data are not justified.
- The panel therefore recommended strategies for minimizing missing data in trials.

Missing data in trials - six recommended principles (steps)

Based on (Little et al. 2012)

1. Find out if values are missing are relevant for the intended analysis.
2. Formulate a well defined causal primary measure of treatment effect.
3. Document and investigate the reasons for missing data.
4. Decide on a primary set of assumptions about missing data.
5. Perform an analysis using a statistical method which is valid under the assumption chosen in 4.
6. Perform a sensitivity analysis to explore robustness to plausible deviations from the assumption in 4.

A principled approach

- We will attempt to follow such an approach.
- Thinking more generally, outside of clinical trials, step 2. consists of specifying our substantive model or quantitiy of interest.

Example

- e.g. consider the following break down of smoking status (for males in THIN from (Marston et al. 2010).
- Our objective is to estimate the marginal distribution of smoking status in the population.

Smoking status	n (% of sample)	(% of those observed)
Non	82,479 (36)	(48)
Ex	30,294 (13)	(18)
Current	57,599 (25)	(34)
Missing	56,661 (25)	n/a

- Are the %s in the last column unbiased estimates?

Missingness mechanisms

Rubin's classification

- Our first step is to think about the mechanism causing a variable (e.g. smoking status) to be missing.
- Rubin developed a classification for missing data 'mechanisms' (Rubin 1976).
- We introduce the three types in a very simple setting.
- We assume we have one fully observed variable Y_1 (age), and one partially observed variable Y_2 (blood pressure (BP)).
- We will let R indicate whether Y_2 is observed ($R = 1$) or is missing ($R = 0$).
- Note Y_2 is not necessarily the 'outcome' in our final analysis.

Missing completely at random

- The missing values in BP (Y_2) are said to be missing completely at random (MCAR) if missingness is independent of BP (Y_2) and age (Y_1).
- i.e. those subjects with missing BP do not differ systematically (in terms of BP or age) to those with BP observed.
- In terms of the missingness indicator R, MCAR means

$$P(R = 1|Y_1, Y_2) = P(R = 1)$$

Example - blood pressure (simulated data)

To illustrate, we consider some simulated data on age (categorised) and systolic blood pressure.

```r
summary(bpObs)
```

```
##          ageCat          bp                      rCat
##  30-50 years:100   Min.   : 60.1   BP missing : 74
##  50-70 years:100   1st Qu.:112.3   BP observed:126
##                    Median :127.0
##                    Mean   :129.1
##                    3rd Qu.:149.1
##                    Max.   :189.5
##                    NA's   :74
```

Checking MCAR

- With the observed data, we could investigate whether age Y_1 is associated with missingness of blood presure (R).
- If it is, we can conclude the data are not MCAR.
- If it is not, the data are consistent with MCAR, although it is still possible that it is MNAR.
- It is possible (though arguably unlikely in this case) that BP is associated with missingness in BP, even if age is not.

Checking MCAR

To examine whether BP is plausibly MCAR, we compare the proportion of missingness between the two age categories:

	BP missing	BP observed	Sum
30-50 years	53	47	100
50-70 years	21	79	100
Sum	74	126	200

Testing MCAR

We can formally test MCAR, e.g. with a chi-squared test:

```
chisq.test(table(bpObs$ageCat, is.na(bpObs$bp)))
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table(bpObs$ageCat, is.na(bpObs$bp))
## X-squared = 20.613, df = 1, p-value = 5.62e-06
```

Here we have strong evidence to reject MCAR.

Missing at random

- BP (Y_2) is missing at random (MAR) given age (Y_1) if missingness is independent of BP (Y_2) given age (Y_1).
- This means that amongst subjects of the same age, missingness in BP is independent of BP.
- In terms of the missingness indicator R, MAR means

$$P(R = 1 | Y_1, Y_2) = P(R = 1 | Y_1)$$

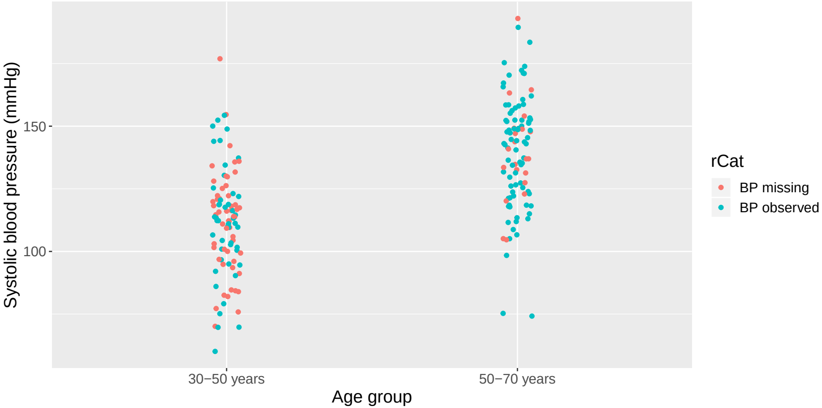
- The name is unfortunate. MAR does **not** mean data are missing completely randomly!

Checking MAR

- We cannot check whether MAR holds based on the observed data.
- To do this we would need to check whether, within categories of age, those with missing BP had higher/lower BP than those with it observed.

Blood pressure MAR given age

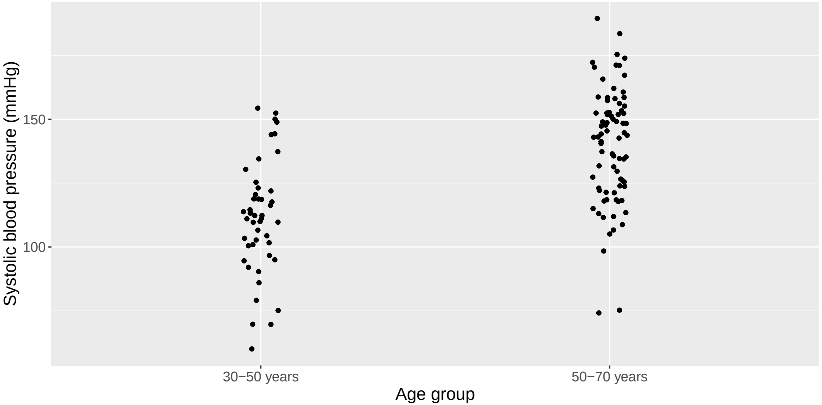
Using the full/complete data:



From this MAR appears plausible - within age categories, the distributions of observed and missing BP look similar.

Blood pressure MAR given age

But in reality all we get to see is:



Analysis assuming MAR

- If we are willing to assume data are MAR, we can construct unbiased estimates using a variety of statistical methods.
- e.g. estimate overall mean BP by a weighted average of observed BP means, weighting according to overall proportions of age categories:

$$\frac{100 \times 111.4 + 100 \times 139.6}{200} = 125.5$$

- Note this is not the same as crude average observed BP.

```r
mean(bpObs$bp, na.rm=TRUE)
```

```
## [1] 129.1174
```

A different representation of MAR

- We have defined MCAR and MAR in terms of how $P(R=1|Y_2, Y_1)$ depends on age (Y_1) and BP (Y_2).
- From the plot, we see that MAR can also be viewed in terms of the conditional distribution of BP (Y_2) given age (Y_1).
- MAR implies that

$$f(Y_2|Y_1, R=0) = f(Y_2|Y_1, R=1) = f(Y_2|Y_1)$$

- That is, the distribution of BP (Y_2), given age (Y_1), is the same whether or not BP (Y_2) is observed.
- This key consequence of MAR is directly exploited by **multiple imputation**.

Missing not at random

- If data are neither MCAR nor MAR, they are missing not at random (MNAR).
- This means the chance of seeing Y_2 depends on Y_2, even after conditioning on Y_1.
- Equivalently, $f(Y_2|Y_1, R=0) \neq f(Y_2|Y_1, R=1)$.
- MNAR is much more difficult to handle. Essentially the data cannot tell us how the missing values differ to the observed values (given Y_1).
- We are thus led to conducting sensitivity analyses.

An MNAR analysis of mean blood pressure

- Suppose that, within age categories, the missing BPs are 10mmHg higher than the observed BPs.
- Given this assumption, we can estimate mean BP by assuming the mean of the missing BPs are 10mmHg higher than predicted by MAR:

$$\frac{47 \times 111.4 + 53 \times 121.4 + 79 \times 139.6 + 21 \times 149.6}{200} = 129.2$$

- Note that we must specify how we think the missing BPs differ to the observed values, based on our contextual knowledge.
- The data **cannot** tell us how large this difference is!

Summary

- Missing data introduce ambiguity into the analysis, beyond the familiar sampling imprecision.
- Extra assumptions about the missingness mechanism are needed to ensure valid estimates and inferences.
- These assumptions can rarely be verified from the data at hand.
- It is sensible to consider carefully possible missingness mechanisms, and formulate appropriate analyses.
- Because we cannot be sure about the type of missingness mechanism at work, sensitivity analyses are important.

Summary continued

- Missingness mechanisms fall into three broad classes: MCAR, MAR and MNAR.
- Under MCAR, we obtain valid estimates and inferences by analysing the subset of subjects with no missing values.
- Under MAR, we must allow for variables (somehow) which predict missingness.
- MAR analyses can be done in a number of ways.
- Multiple imputation is one such approach, which we will explore in this course.

References I

Council, National Research. 2010. *The Prevention and Treatment of Missing Data in Clinical Trials*. Washington, DC: National Academies Press.

Karahalios, Amalia, Laura Baglietto, John B Carlin, Dallas R English, and Julie A Simpson. 2012. "A Review of the Reporting and Handling of Missing Data in Cohort Studies with Repeated Assessment of Exposure Measures." *BMC Medical Research Methodology* 12 (1). BioMed Central: 96.

Little, Roderick J., Ralph D'Agostino, Michael L. Cohen, Kay Dickersin, Scott S. Emerson, John T. Farrar, Constantine Frangakis, et al. 2012. "The Prevention and Treatment of Missing Data in Clinical Trials." *New England Journal of Medicine* 367 (14): 1355–60. https://doi.org/10.1056/NEJMsr1203730.

References II

Marston, L., J. R. Carpenter, K. R. Walters, R. W. Morris, I. Nazareth, and I. Petersen. 2010. "Issues in Multiple Imputation of Missing Data for Large General Practice Clinical Databases." *Pharmacoepidemiology and Drug Safety* 19: 618–26.

Rubin, D B. 1976. "Inference and missing data." *Biometrika* 63: 581–92.