Applied Statistical Analysis with Missing Data
Sensitivity analysis and reporting

Morten Frydenberg ©
Department of Biostatisics, Aarhus Univ, Denmark

Sensitivity analysis

Guidelines for reporting analysis of data with missings

Adjusting for missing observation by weights

What you have learned

Evaluation

Goodbye

Sensitivity analysis

In general, a sensitivity analysis is an **alternative analysis** of the data, where you try to see how sensitive you results are to minor or large deviations from the assumptions behind the statistical analysis you have made.

There are at least **two types** of sensitivity analysis that you should make, when you have missing data (and use MI).

A: Try different plausible **imputation models**.
If these give very different results, then you probably have not really understood the structure of the data correctly.

B: Try modelling the missing data in a **MNAR way** and se how this influences the complete data analysis.
A **unrealistic** model could be the "worst-case scenario", where all missing data are set to extremes: highest/lowest birth weight, mother non-smoker/ 20 cigs per day.

Sensitivity analysis – type B

A worst-case sensitivity analysis is **seldom interesting** as it is not plausible.

Often authors try to make **"mental"** sensitivity analyses based on what they learned long ago for a **two by two** table:

> *If small children have a higher risk of not being recorded, then we have underestimated the effect of smoking during pregnancy.*

But often the setting is **too complicated** for the logic based on a two by two table to be correct.

With modern computers **you can do much better**.

You can **simulate** the missing data from any distribution you want, e.g. assuming that the missing birth weight are **on average** 150g lower than the observed birth weights.
And analysis this new data set to how things change.

Morten Frydenberg Missing data - lecture 5 3

Sensitivity analysis – type B

- mis. birth weight 150 g **higher/lower**, than observed (given sex, age parity....?)
- mis. birth weight had larger **variance** than observed (given sex, age parity....?)
- missing data is mainly among "unhealthy – strong", where there is **no relationship** between smoking and birth weight
- mis. income is 10% lower/higher, than what is found **based on the imputation model**
- mis health score is 20% higher (you do not turn up if **well**) than we see in the rest of the data.
- mis health score is 20% lower (you do not turn up if **not well**) than we see in the rest of the data.
- mis health scores is 20% higher among women on drugE.

Morten Frydenberg Missing data - lecture 5 4

An old example!

Case-control study of squamous cell cancer of the oral cavity in Denmark.
Bundgaard, Wildt, Frydenberg, Elbrond, & Nielsen; *Cancer Causes and Control* **1995**,6 57-67

161 cancer cases – each matched to **four** controls.
Primary exposures were: number of teeth, alcohol and smoking habits – data collected by a mailed questionnaire.
All cases, but only 400 controls **(85%)** returned the questionnaire.

Complete data analysis showed that cancer was highly associated with alcohol, smoking and having few teeth.

Did we have MCAR? – **no!**
missing was definitely higher among alcoholics.

The complete data analysis was biased!!

An old example!

Case-control study of squamous cell cancer of the oral cavity in Denmark.
Bundgaard, Wildt, Frydenberg, Elbrond, & Nielsen; *Cancer Causes and Control* **1995**,6 57-67

Did we have MAR? – **no!**
We do not believe that participation is independent of alcohol habits given just age and sex,
which was all we knew about the non-responders.

As we expect that the use of alcohol (and smoking) was higher among missing controls compared to those who participated – the study (probably) **overestimated** the association between cancer and alcohol – **but how much**?

We made **two sensitivity studies**
-A: The missing controls were sampled among **cases**
-B: The missing controls were sampled among **controls** but the use of alcohol and smoking were **doubled**.

An old example!

Table 8. Results of two simulation studies evaluating the effect of nonrespondents

		Comp case	A	B
		Actual estimates OR[c]	Simulation (i)[a] 1,000 simulations geometric average OR[c]	Simulation (ii)[b] 1,000 simulations geometric average OR[c]
Current alcohol	1-5 drinks/day	1.1	1.2	0.9
	6+ drinks/day	9.7[d]	4.5[d]	7.2[d]
Current tobacco	1-20 g/day	2.0[d]	1.8[d]	1.2
	21+ g/day	6.3[d]	3.7[d]	4.4[d]
Number of teeth	5-14	1.5	1.4	1.3
	0-4	2.1[d]	1.9[d]	1.6[d]

[a]Pseudo-controls taken as random cases.
[b]Pseudo-controls taken as random control, but with double the tobacco and alcohol consumption.
[c]OR = odds ratio (reference: 0 drinks/day, 0 g tobacco/day, 15-32 teeth).
[d]$P < 0.05$.

Result of the sensitivity studies:
The association with alcohol and smoking were reduced, but still high.

Guidelines for reporting any analysis potentially affected by missing data.

From: *Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls.*
Sterne et al. BMJ 2009;338 :b2393 doi: 10.1136/bmj.b2393

Report the number of missing values for each variable of interest, or the number of cases with complete data for each important component of the analysis.

Give reasons for missing values if possible, and indicate how many individuals were excluded because of missing data when reporting the flow of participants through the study.

If possible, describe reasons for missing data in terms of other variables (rather than just reporting a universal reason such as treatment failure).

Clarify whether there are important differences between individuals with complete and incomplete data — for example, by providing a table comparing the distributions of key exposure and outcome variables in these different groups.

Describe the type of analysis used to account for missing data (eg., multiple imputation), and the assumptions that were made (eg., missing at random).

If you use imputation.
Provide details of the imputation modelling:

- Report details of the software used and of key settings for the imputation modelling.
- Report the number of imputed datasets that were created (at least 20 may be preferable to reduce sampling variability from the imputation process).
- What variables were included in the imputation procedure? How were non-normally distributed and binary/categorical variables dealt with?
- If statistical interactions were included in the final analyses, were they also included in imputation models?

If a large fraction of the data is imputed, compare observed and imputed values.

Where possible, provide results from analyses restricted to complete cases, for comparison with results based on multiple imputation. If there are important differences between the results, suggest explanations, bearing in mind that analyses of complete cases may suffer more chance variation, and that under the missing at random assumption multiple imputation should correct biases that may arise in complete cases analyses.

Discuss whether the variables included in the imputation model make the missing at random assumption plausible.

It is also desirable to investigate the robustness of key inferences to possible departures from the missing at random assumption, by assuming a range of missing not at random mechanisms in sensitivity analyses.

Adjusting for missing by using sampling weights

An extreme type of missing data is, when you use **stratified sampling**.
E.g. you sample 100 random persons within each five year age and sex strata in the population.

Such a sample is clearly not **MCAR**.

But the data is **MAR**, as you know the age and sex of each of the persons you have not observed.

We could analyse the data with MI using models within each age and sex strata.

Another, valid and simple way is to analyse the data by **weighting** each observed data by **w=(strata size)/100**.

I.e. give a high weight to observation from large strata and a low weight for observations from small strata.

Adjusting for missing by using sampling weights

Using weights **w=(strata size)/100**.

This correspond to

w=1/"chance of being sampled"

or in terms of observed/missing

w=1/"chance of being observed"

The method of using weights can be used in general.

If we have a data set with **only missing outcome** Y and MAR, then we can

1. Model the "chance of being observed" by a logistic regression analysis using the relevant covariates.
2. Use 1/(estimated probability) as weight in the analysis of the observed/complete data set.

Different types of missingness

1 The Easy **MCAR:**
Complete case analysis version one - will give an unbiased estimate of θ.

2 The Tough **MAR:**
It is possible (in theory) to get an unbiased estimate of θ by analysing the observed data correctly.

3 The Unsolvable **MNAR:**
It is impossible to get an unbiased of θ only based on the observed data.

If the reason/mechanism behind the missingness is not known, then it is impossible distinguee situation 2 and 3 (and even 1).

Missing At Random

The chance of observing the unobserved
is independent of the unobserved
given what you have observed

If the data is MAR, then you **might be able** to obtain unbiased estimates with correct standard errors, i.e. valid confidence intervals and p-values by Multiple Imputation.

In MI you (Stata):

- generate (impute) several version of the complete data set.
- analyse each of these by the **complete data model**
- find the average estimate together with its SE and Df by Rubins formula.

Missing At Random and MI

A Multiple Imputation solution requires, that you have an (approximate) valid model for the **unobserved** data given what you have observed.

This requires insight into the missing data mechanism.

As you never know this mechanism:
There is no guaranty the MI method will solve you problem!

But ignoring the problem will not make it disappear.

There are other methods that might solve the missing data problem – but they are in **general more complicated**.

Evaluation

Form
- Location
- 3 consecutive days?

Contents
- Topics you observed that you would like to miss
- Topics you missed

Weights /time spend
- The level should it be different
- Time spend on lectures/exercises
- Time spend on Topics