Applied Statistical Analysis with Missing Data Theory?

Morten Frydenberg © Section of Biostatistics, Aarhus Univ, Denmark

The principles of statistical inference
- The likelihood method
- Estimates, standard errors, confidence intervals and tests
- Bias, coverage probabilities and efficiency

Why are data missing?

Inference ignoring the missing data problem

Different types of missingness

How to attack the missing problem

The Multiple Imputation procedure – an outline

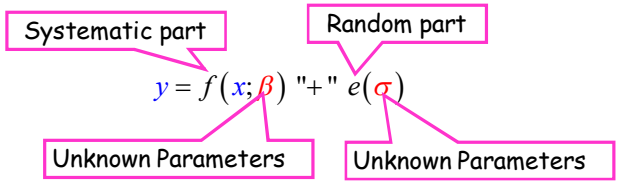
1

The principles of statistical inference

In very general terms the purpose of statistical inference is to estimate the relationship between a response/outcome Y and a set of explanatory variables X.

In order to do that we specify **a statistical model** of the relationship between Y and X.

The model will typically contain a **systematic** and a **random** part with corresponding unknown constants – **parameters**.



2

The principles of statistical inference

The analysis/**substantive** model: $y = f(x;\beta)$ "+" $e(\sigma)$

We will denote the **combined data** by $Z = (Y, X)$.
and combined set of parameters by $\theta = (\beta, \sigma)$.

That is, Z will contain both outcome and explanatory variables, and θ will contain parameters concerning the systematic and the random part of the model.

We will use the data Z **to make inference** concerning θ.

I.e. find
- estimates, standard errors, confidence intervals

and calculate
- test-statistics and p-values

for relevant hypotheses.

3

The maximum likelihood method

Most statistical inference is based on **the likelihood function**:

$$L(\theta \mid z) = p_{Z;\theta}(z;\theta)$$

Based on the likelihood function one can calculate:

- The **M**aximum **L**ikelihood **E**stimate (MLE) of θ
- Approximate **S**tandard **E**rror of the MLE
- Approximate **C**onfidence **I**ntervals (based on MLE and SE)
- Approximate (Wald) **tests** of hypotheses concerning θ

But the method requires that "you" can calculate the probability of the data (**those that are on your hard disc**!) given the parameters.

4

The principles of statistical inference
Bias, coverage probability, and efficiency

Let us denote the **true**, but unknown, parameter value by θ_T

And let $\hat{\theta}_n$ denote an estimator (it could be the MLE) based on n **independent** observations.

The estimate is said to be
unbiased if: Expected value of $\hat{\theta}_n = \theta_T$
asymptotically unbiased if: $\hat{\theta}_n \to \theta_T$ as $n \to \infty$

That is, the estimate gets ever closer to the true value as we get more observations.

Very few estimates are unbiased (notable exception: normal multiple regression), but MLE's are in general asymptotically unbiased.

5

The principles of statistical inference
Bias, coverage probability, and efficiency

The **coverage probability** of a Confidence Interval $(L_n; U_n)$ is $\Pr(L_n < \theta_T < U_n)$

i.e. the probability that the true parameter is contained in the interval.

Ideally a 95% CI will have a coverage probability = 95%.

In practice very few methods will give exactly the stated coverage probability (normal multiple regression), and we will have to hope for **asymptotically correct** coverage probabilities (large data set – correct coverage).

Confidence intervals based on the MLE and the ML SE have in general asymptotically correct coverage probabilities.

6

The principles of statistical inference
Bias, coverage probability, and efficiency

The **width** of a Confidence Interval is $U_n - L_n$

When comparing two (asymptotically) **unbiased** methods of estimation, the method with the **smallest average width** of the confidence interval is said to be the **most efficient**.

In general we prefer :

- estimates that are (asympt.) **unbiased**
- CIs that have (asympt.) **correct coverage probabilities**
- methods that are (asympt.) the **most efficient**

Note:
High **efficiency** corresponds to higher statistical power.
High **efficiency** corresponds to small SE.

7

What do we observe/record when data are missing?

We do not fully observe Z, but rather:

Z_{Obs}

id	y	x_1	x_2	x_3
1	o	o	o	o
2	o	.	o	o
3	.	o	o	o
4	.	.	o	o
5	o	o	o	o
6	o	.	.	o

o observed

$R_{ecorded}$

id	y	x_1	x_2	x_3
1	1	1	1	1
2	1	0	1	1
3	0	1	1	1
4	0	0	1	1
5	1	1	1	1
6	1	0	0	1

8

Missing data

Avoid missing data!!!

If not, then collect as much information on the reason why the observation became missing:

- did the patient refuse to participate?
- is the patient dead?
  Is this missing data?
- did the patient not turn up?
- was the "measurement" never made?
- was the result not registered?

→ **Why???**

- was the value below the detection limit?
  Is this missing data?

9

Missing data – Solutions??

Complete case analysis - version one:
Ignore the problem and only analyse **patients** with information on all relevant variables!!!

Pros: Always possible – transparent model

Cons: Model often wrong
Estimate likely biased
Analysis likely inefficient

Complete case analysis - version two:
Ignore the problem and only use **variables** that are available for all patients!!!

Pros: Always possible - transparent model

Cons: Wrong/irrelevant model
Biased estimate!!

10

Different types of missingness

Three types of problems:

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 based solely on information in the observed data.

If the reason/mechanism behind the missingness is not known, then it is impossible to distinguish between situation 2 and 3.

11

Missing at random???

What Is Meant by "Missing at Random"?

Shaun Seaman, John Galati, Dan Jackson and John Carlin

Statistical Science
2013, Vol. 28, No. 2, 257–268
DOI: 10.1214/13-STS415

Abstract. The concept of missing at random is central in the literature on statistical analysis with missing data. In general, inference using incomplete data should be based not only on observed data values but should also take account of the pattern of missing values. However, it is often said that if data are missing at random, valid inference using likelihood approaches (including Bayesian) can be obtained ignoring the missingness mechanism. Unfortunately, the term "missing at random" has been used inconsistently and not always clearly; there has

12

How to attack the problem with missing data

First try to determine whether you are in situation 1, 2 or 3, by going through the possible missing data mechanisms.

If you are in situation 1 **(MCAR)** then a complete case analysis will be valid, although not the most efficient.

If you are in situation 3 **(MNAR)** then you **have to make additional** assumptions concerning the missing data in order to analyse the observed data. Specific modelling will be required and you will typically **not be able use standard programs**.

If you are in situation 2 **(MAR)** then you **might solve** the problem by **imputation** methods.

In all cases you should supplement your analysis with different types of **sensitivity analyses**.

13

The analyses of data with missing values by imputation

The analysis using imputation have 3 separate components:

1. **The full data model – the analytic/substantive model**
   Specification of how to analyse the data, if it was without missing values.
2. **Imputing the missing values**
   Generate K complete data sets by generating K values of the missing data.
3. **The estimation**
   Find K estimates of θ - one for each of the K 'complete' datasets. The final, overall estimate of θ is found as the average of the K estimates. Calculate a suitable standard error (Rubins formula).

14

Missing At Random and MI

A Multiple Imputation solution requires, that you have a consistent and 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.

15

Will Multiple Imputation work?

The method works if:

- It gives **asymptotic unbaised estimates**
- It gives **Confidence Intervals** with asymptotically correct **confidence intervals**.

It is worthwhile if it is more efficient.

Does it work????

Morning exercise!

16