Mini Lecture 7:

Checking the data

Outline of Session

- Why check data
- Example
- Different errors and how to check for them
- What to do if you find an error

Outline of Session

- Why check data
- Example
- Different errors and how to check for them
- What to do if you find an error

Why we check data

- Need good quality data in order to obtain good quality evidence
- Errors can easily arise during data collection and entry
- Incorrect or missing data may lead to bias or incorrect results so should be avoided if possible

Outline of Session

- Why check data
- Example
- Different errors and how to check for them
- What to do if you find an error

Example

Q: Do people who start ART with a CD4 count >500 cells/mm^3 experience a higher rate of ART toxicity than people with a CD4 count ≤500 cells/mm^3?

Example

ID	Date_Birth	Sex	Mode_ acquisition	Ethnicity	CD4 count	Toxicity	Date_Death
A1234	26/09/1952	M	MSM	White	209	0	
B1235	01/01/1988	F	Het	White	342	0	
A1236	14/08/1966	M	MSM		578	0	30/07/1910
H1237	04/26/1974	M	IDU		174	0	
X1238	01/11/1996	M	MTCT		621	1	
S1239	15/06/1977	F	Het		3354	0	
B1240	19/10/1981	M	Het	Black African	267	0	
H1241	01/07/1978	F	MSM		351	0	
S1242	22/06,1962	M	MSM		495	1	29/10/2013
H1243	01/01/2001	M	Het	Unknown	357	0	

Outline of Session

- Why check data
- Example
- Different errors and how to check for them
- What to do if you find an error

Types of data checks

1. Formats

Format	Example	
Date	Date of birth, date of death	dd/mm/yyyy *or* mm/dd/yyyy

Types of data checks

1. Formats

Format	Example	
Date	Date of birth, date of death	dd/mm/yyyy *or* mm/dd/yyyy
Character/look up	Mode of acquisition	MSM, Het, MTCT
(Categorical)	Sex	M, F

Types of data checks

1. Formats

Format	Example	
Date	Date of birth, date of death	dd/mm/yyyy *or* mm/dd/yyyy
Character/look up	Mode of acquisition	MSM, Het, MTCT
(Categorical)	Sex	M, F
Number	CD4 count	209, 342, 578, 174 etc.
(Numeric or categorical)	Toxicity	0/1

Types of data checks

1. Formats

Format	Example	
Date	Date of birth, date of death	dd/mm/yyyy *or* mm/dd/yyyy
Character/look up	Mode of acquisition	MSM, Het, MTCT
(Categorical)	Sex	M, F
Number	CD4 count	209, 342, 578, 174 etc.
(Numeric or categorical)	Toxicity	0/1
Text	Name, hospital number, cause of death	A19235 “Mr X died as a result of post-operative haemorrhage”

Types of data checks

1. Formats

ID	Date_Birth	Sex	Mode_ acquisition	Ethnicity	CD4 count	Toxicity	Date_Death
A1234	26/09/1952	M	MSM	White	209	0	
B1235	01/01/1988	F	Het	White	342	0	
A1236	14/08/1966	M	MSM		578	0	30/07/1910
H1237	04/26/1974	M	IDU		174	0	
X1238	01/11/1996	M	MTCT		621	1	
S1239	15/06/1977	F	Het		3354	0	
B1240	19/10/1981	M	Het	Black African	267	0	
H1241	01/07/1978	F	MSM		351	0	
S1242	22/06,1962	M	MSM		495	1	29/10/2013
H1243	01/01/2001	M	Het	Unknown	357	0	

Types of data checks

1. Formats

How you can check for errors:

- Most statistical packages will select a format based on values
- Tabulate values
- Sort values in order

Types of data checks

2. Outlying or impossible values

- An outlying value is a value that lies far outside the normal range for a numeric variable
- Sometimes this value will be impossible (e.g. CD4 = -7)

Types of data checks

2. Outlying or impossible values

- An outlying value is a value that lies far outside the normal range for a numeric variable
- Sometimes this value will be impossible (e.g. CD4 = -7)

- For categorical (look up) variables it is also possible that a value will be entered that does not align to a pre-specified category

Types of data checks

2. Outlying or impossible values

ID	Date_Birth	Sex	Mode_ acquisition	Ethnicity	CD4 count	Toxicity	Date_Death
A1234	26/09/1952	M	MSM	White	209	0	
B1235	01/01/1988	F	Het	White	342	0	
A1236	14/08/1966	M	MSM		578	0	30/07/1910
H1237	04/26/1974	M	IDU		174	0	
X1238	01/11/1996	M	MTCT		621	1	
S1239	15/06/1977	F	Het		3354	0	
B1240	19/10/1981	M	Het	Black African	267	0	
H1241	01/07/1978	F	MSM		351	0	
S1242	22/06,1962	M	MSM		495	1	29/10/2013
H1243	01/01/2001	M	Het	Unknown	357	0	

Types of data checks

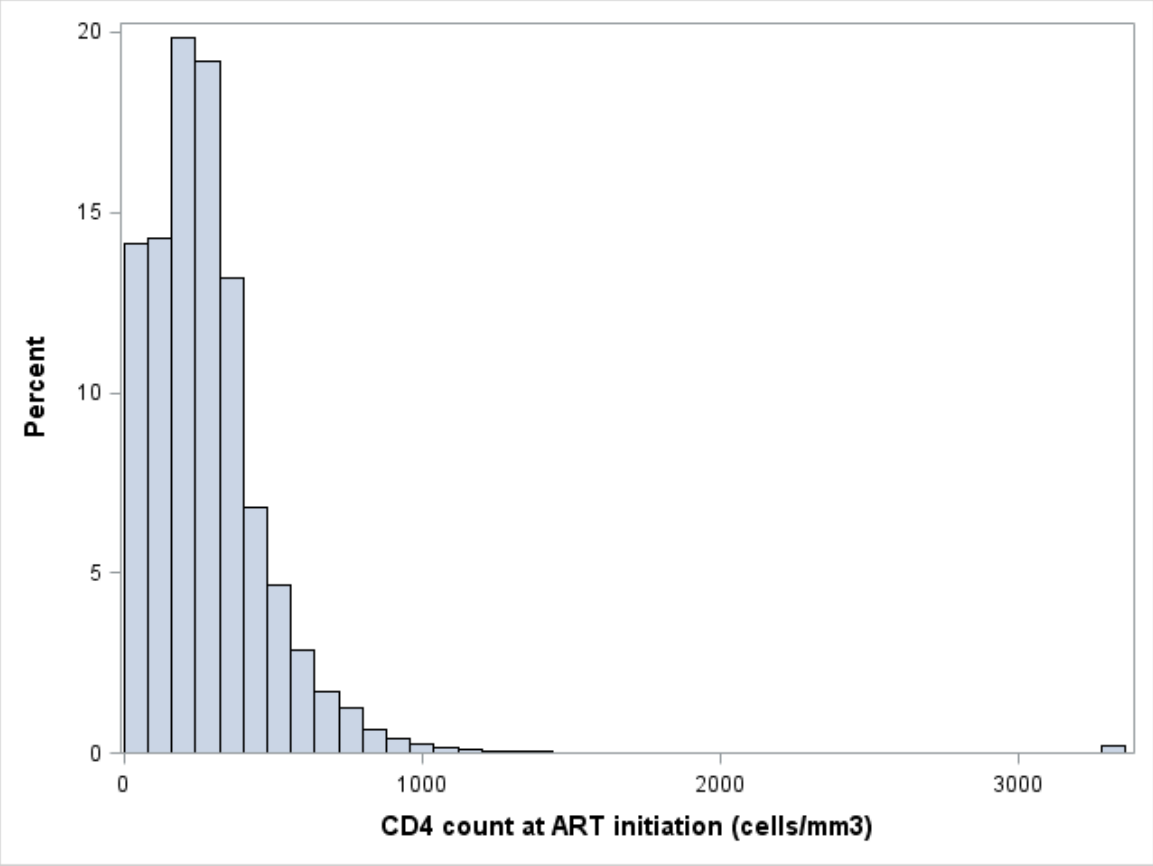
2. Outlying or impossible values

How you can check for errors:

- Plot histogram (numeric)
- Tabulate values (categorical)

Types of data checks

2. Outlying or impossible values



Types of data checks

3. Missing data

- If there are large amounts of missing data or data are missing in some systematic way then this may introduce bias
- Missing data may be entered as “Unknown”

Types of data checks

3. Missing data

ID	Date_Birth	Sex	Mode_ acquisition	Ethnicity	CD4 count	Toxicity	Date_Death
A1234	26/09/1952	M	MSM	White	209	0	
B1235	01/01/1988	F	Het	White	342	0	
A1236	14/08/1966	M	MSM		578	0	30/07/1910
H1237	04/26/1974	M	IDU		174	0	
X1238	01/11/1996	M	MTCT		621	1	
S1239	15/06/1977	F	Het		3354	0	
B1240	19/10/1981	M	Het	Black African	267	0	
H1241	01/07/1978	F	MSM		351	0	
S1242	22/06,1962	M	MSM		495	1	29/10/2013
H1243	01/01/2001	M	Het	Unknown	357	0	

Types of data checks

3. Missing data

How you can check for errors:

- Summary statistics (provided in statistical packages e.g. SAS, Stata, SPSS)

Types of data checks

4. Ordering of dates

- Certain events must follow a specific order

E.g. Drug stop dates ≥ Drug start dates

Death > Birth

Types of data checks

4. Ordering of dates

ID	Date_Birth	Sex	Mode_ acquisition	Ethnicity	CD4 count	Toxicity	Date_Death
A1234	26/09/1952	M	MSM	White	209	0	
B1235	01/01/1988	F	Het	White	342	0	
A1236	14/08/1966	M	MSM		578	0	30/07/1910
H1237	04/26/1974	M	IDU		174	0	
X1238	01/11/1996	M	MTCT		621	1	
S1239	15/06/1977	F	Het		3354	0	
B1240	19/10/1981	M	Het	Black African	267	0	
H1241	01/07/1978	F	MSM		351	0	
S1242	22/06,1962	M	MSM		495	1	29/10/2013
H1243	01/01/2001	M	Het	Unknown	357	0	

Types of data checks

4. Ordering of dates

How you can check for errors:

- Calculate time between two dates and look for times <0
- Use a logic function in excel

IF(logical test, value if true, value if false)

IF (Date_Death<Date_Birth, 1, 0)

Types of data checks

5. Incompatible values

- Certain combinations of values for two related variables may prove impossible in practice

E.g.

Ever smoked? [Never, Past, Current]

Number cigarettes smoked daily? [0,1-20,20-40,>40]

Types of data checks

5. Incompatible values

- Certain combinations of values for two related variables may prove impossible in practice

E.g.

Ever smoked? [Never, Past, Current]

Number cigarettes smoked daily? [0,1-20,20-40,>40]

Types of data checks

5. Incompatible values

ID	Date_Birth	Sex	Mode_ acquisition	Ethnicity	CD4 count	Toxicity	Date_Death
A1234	26/09/1952	M	MSM	White	209	0	
B1235	01/01/1988	F	Het	White	342	0	
A1236	14/08/1966	M	MSM		578	0	30/07/1910
H1237	04/26/1974	M	IDU		174	0	
X1238	01/11/1996	M	MTCT		621	1	
S1239	15/06/1977	F	Het		3354	0	
B1240	19/10/1981	M	Het	Black African	267	0	
H1241	01/07/1978	F	MSM		351	0	
S1242	22/06,1962	M	MSM		495	1	29/10/2013
H1243	01/01/2001	M	Het	Unknown	357	0	

Types of data checks

5. Incompatible values

How you can check for errors:

- Cross-tabulate variables

Types of data checks

5. Incompatible values

How you can check for errors:

- Cross-tabulate variables

	MSM	Het	IDU	MTCT	Total
M	3	2	1	1	7
F	1	2	0	0	3
Total	4	4	1	1	10

Types of data checks

5. Incompatible values

How you can check for errors:

- Cross-tabulate variables

	MSM	Het	IDU	MTCT	Total
M	3	2	1	1	7
F	1	2	0	0	3
Total	4	4	1	1	10

Types of data checks

6. Checking assumptions of statistical tests

- Normal distribution (t-test, ANOVA)
- Expected counts (chi-square test)
- Number of observations or events

What to do if you find an error

- For some errors it is obvious what the correct information should be (e.g. dates with incorrect formats)
- If possible you should go back to the data source and try and find out the correct information
- If you cannot find out the correct information, the error should be set to missing

Summary of data checks

1. Formats
2. Outlying values
3. Missing data
4. Ordering of dates
5. Incompatible values
6. Checking statistical assumptions