UNIVERSITY OF LIVERPOOL

The Royal Liverpool and
Broadgreen University Hospitals
NHS Trust

NHS

Collecting Data for Research

A Statistical Perspective

Professor Azzam F. G. Taktak

Data Format

- Is the data:
  - Continuous – e.g. age, birthweight, blood pressure, height
  - Categorical – e.g. gender, smoking status, marital status, ethnic origin
  - Ordinal – e.g. TNM stage, GCS, BIS index

Raw Data

Make sure you store raw data as much as possible. For example if you have the following set of measurements:

180, 130, 70

Do not store simply as:

High, Medium, Low

You might forget the thresholds used or decide to use different thresholds at a later stage

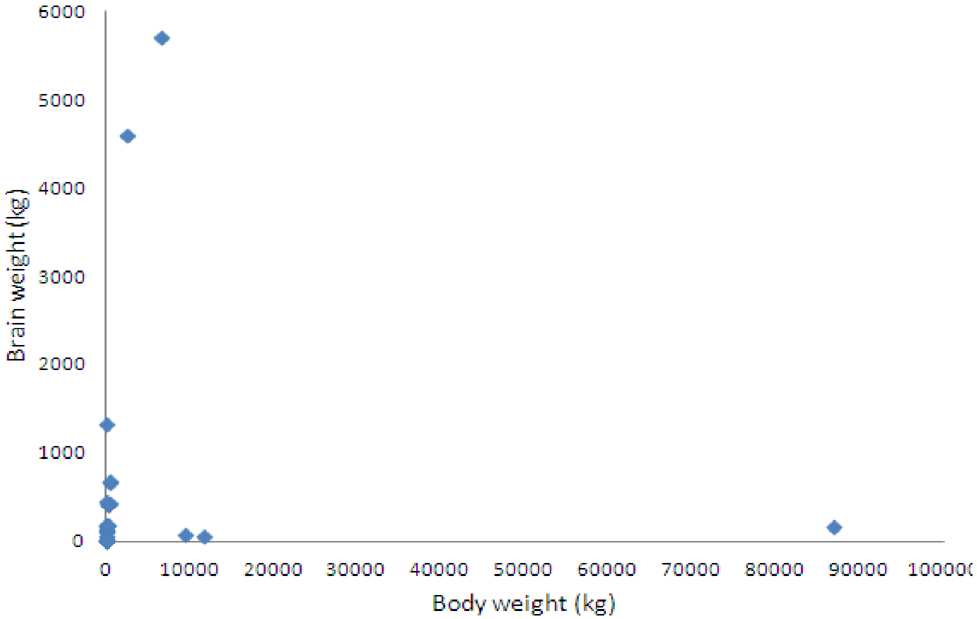
Accuracy

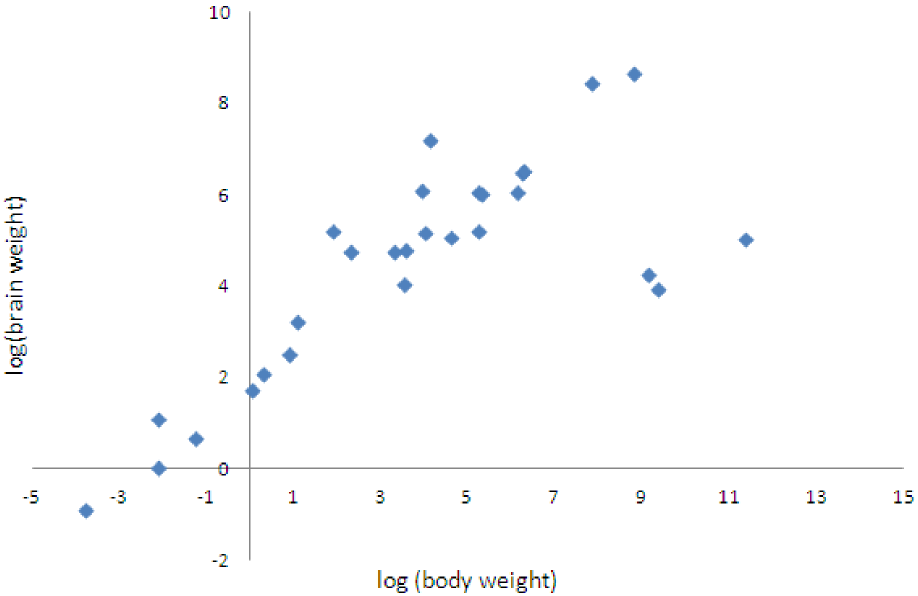
- How many decimal places does your data really have?
  - Average length of gestation = 9.3346785 months!
- Do not round up numbers too early:
  - 18.34807 rounded too early will give 18.35 and rounded second time will give 18.4 instead of 18.3
- Preserve the original number of decimal points:
  - 70.6003 should be rounded to 70.60 rather than 70.6

Follow-up Time

- It is better to record date of treatment and date last seen to calculate follow-up rather than just follow-up time

Data Screening



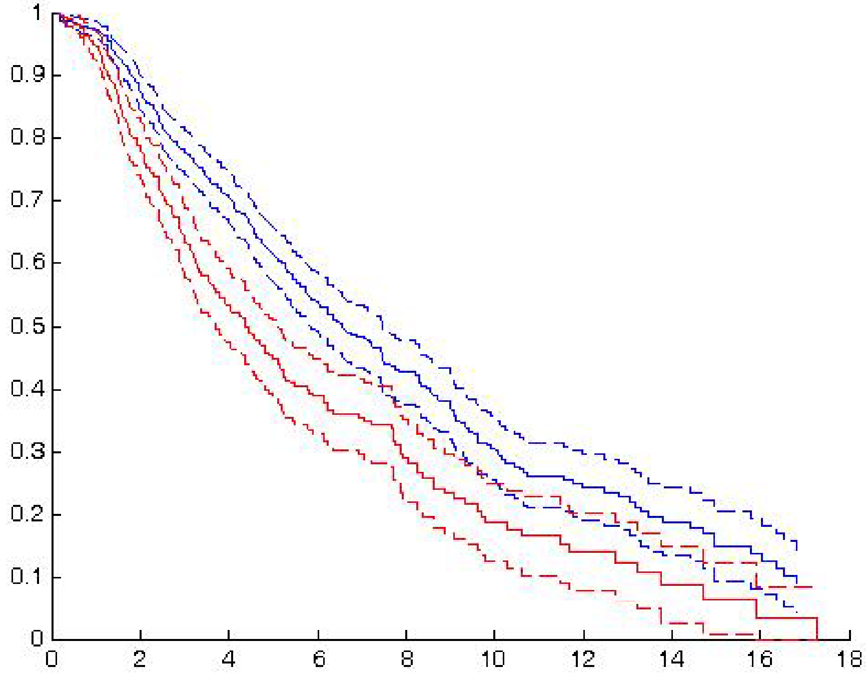


Missing Data

- Do not code missing data as 0 or 999.
  - In a cancer dataset, unknown TNM stage was coded as 99. The median TNM stage in the dataset was calculated as 16

More on Missing Data

- The “missingness” pattern is often informative



Subject ID

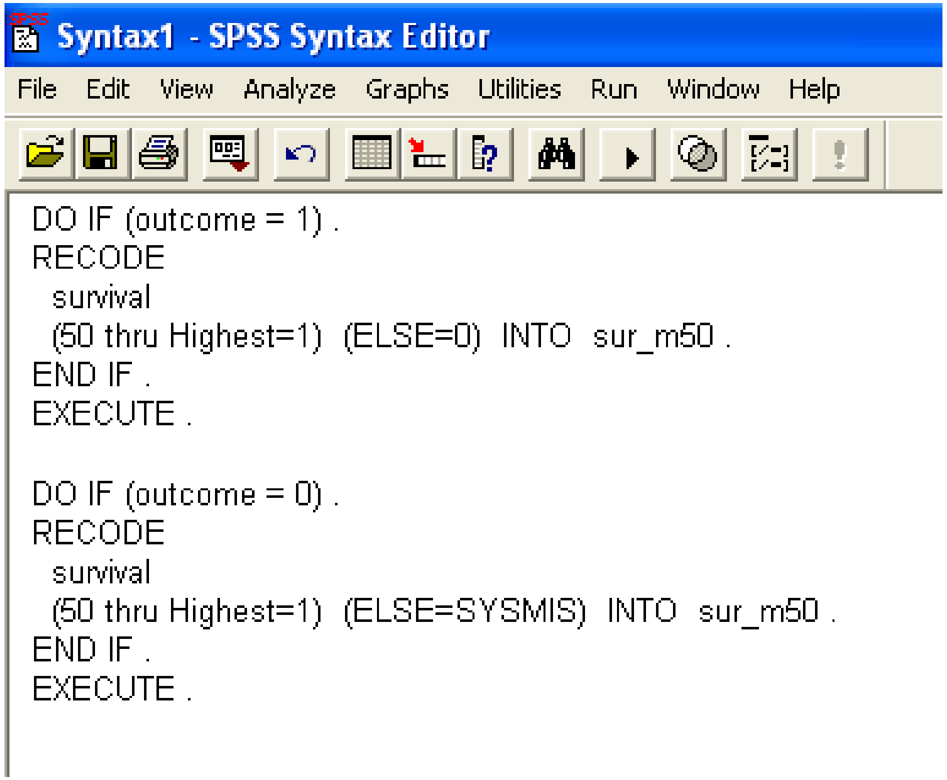
- Legislations often requires that data stored for research purposes are anonymised. However, without subject ID, it is impossible to:
  - query suspicious values
  - ascertain independence of subjects
  - follow subject up
- Pseudo-anonymisation is more appropriate

Transferability of Data

- Store your data in a format that is easily transferable between different software packages, e.g. comma separated text files (.csv)

Session Log

- It is a good idea to keep a log of the session rather than results only.

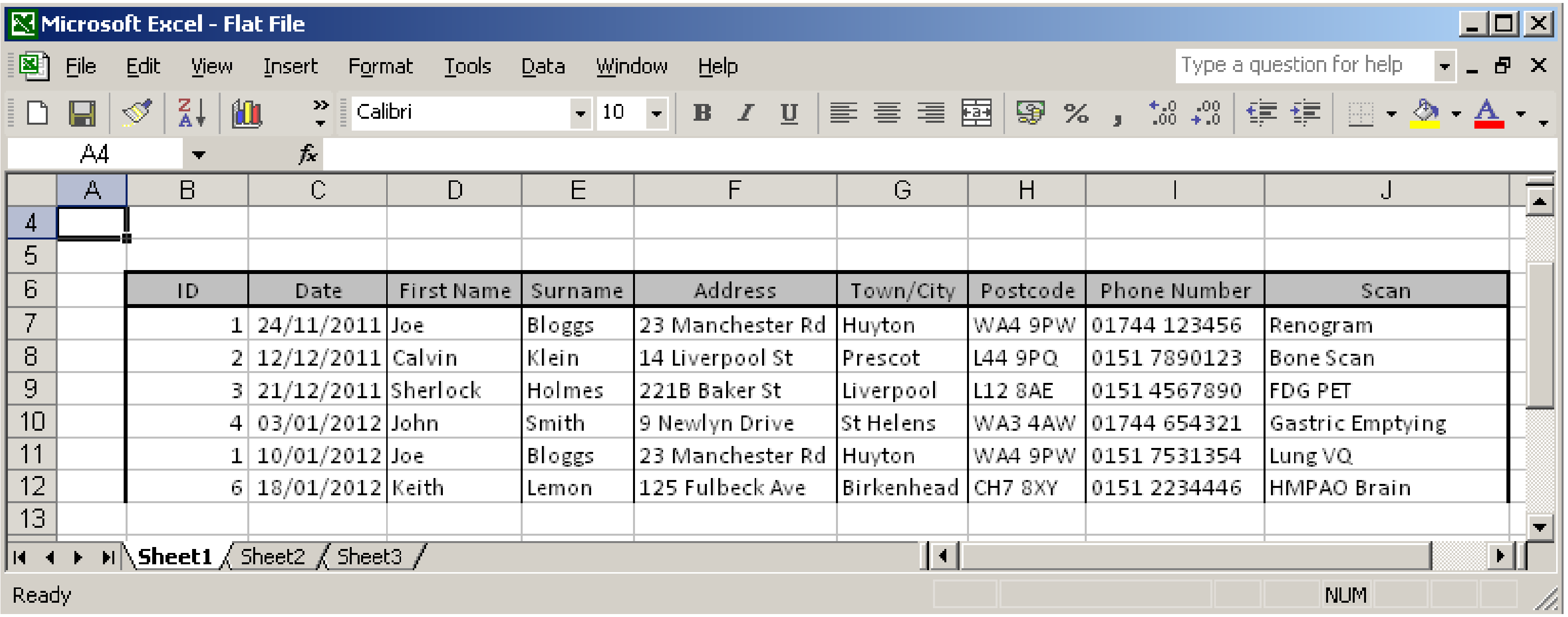


Syntax1 - SPSS Syntax Editor
File Edit View Analyze Graphs Utilities Run Window Help

```
DO IF (outcome = 1) .
RECODE
  survival
  (50 thru Highest=1)  (ELSE=0)  INTO  sur_m50 .
END IF .
EXECUTE .

DO IF (outcome = 0) .
RECODE
  survival
  (50 thru Highest=1)  (ELSE=SYSMIS)  INTO  sur_m50 .
END IF .
EXECUTE .
```

Flat Tables


Microsoft Excel - Flat File
File Edit View Insert Format Tools Data Window Help
Type a question for help
Calibri 10
A4
Sheet1 Sheet2 Sheet3
Ready
NUM

ID	Date	First Name	Surname	Address	Town/City	Postcode	Phone Number	Scan
1	24/11/2011	Joe	Bloggs	23 Manchester Rd	Huyton	WA4 9PW	01744 123456	Renogram
2	12/12/2011	Calvin	Klein	14 Liverpool St	Prescot	L44 9PQ	0151 7890123	Bone Scan
3	21/12/2011	Sherlock	Holmes	221B Baker St	Liverpool	L12 8AE	0151 4567890	FDG PET
4	03/01/2012	John	Smith	9 Newlyn Drive	St Helens	WA3 4AW	01744 654321	Gastric Emptying
1	10/01/2012	Joe	Bloggs	23 Manchester Rd	Huyton	WA4 9PW	0151 7531354	Lung VQ
6	18/01/2012	Keith	Lemon	125 Fulbeck Ave	Birkenhead	CH7 8XY	0151 2234446	HMPAO Brain

As more records are added there is **increased chance of duplication.**

A patient has no phone number, lots of repeated empty fields, **inefficient use of memory.**

A field will have to be selected to prescribe the ID number, **potential for non unique records**.

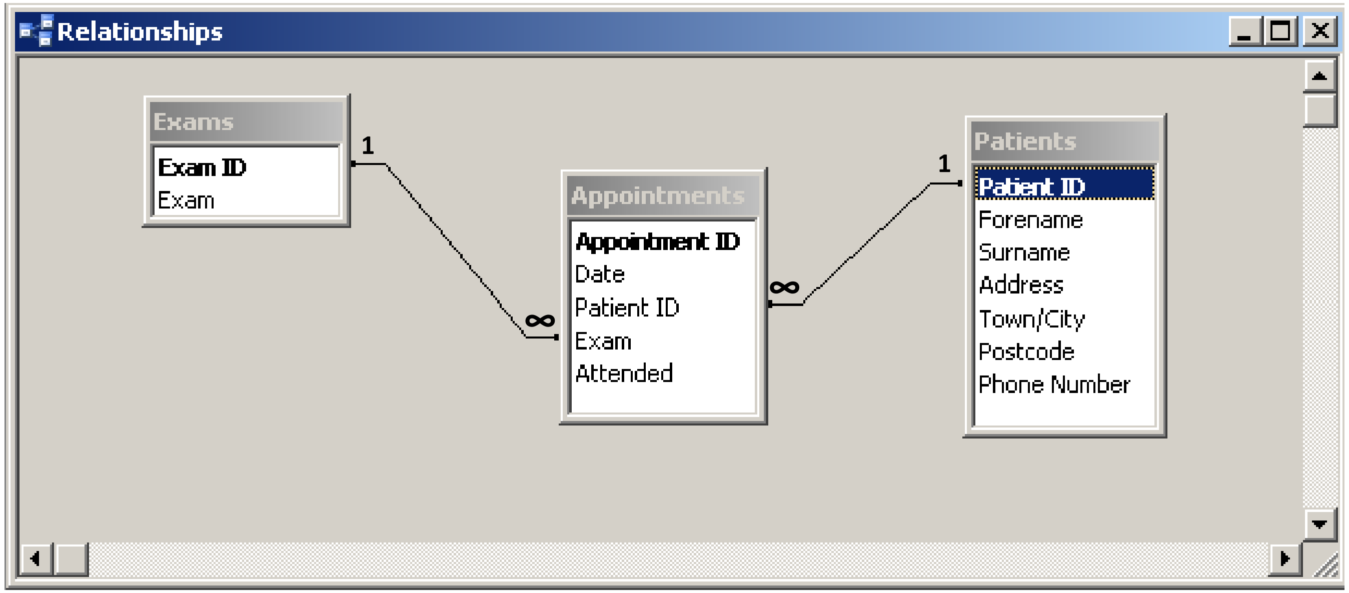
It can handle queries looking at a single fields only, **poor at complex queries.**

Multiple records for each patient making it **harder to update.**

All the information for each patient is in the one table, **poor at limiting access.**

Relational Databases

Hold their data in a number of tables instead of one. Records within the tables are linked (related) to records in other tables by way of common fields.



Relational Databases

- Data is only stored once
  - Avoiding data duplication
  - Bypassing the need for multiple changes modifying/deleting
  - Providing more efficient storage as blank/unnecessary fields are not repeated
- Capable of complex queries
  - Set complex criteria based on multiple fields to select/insert/update/delete/create/drop/calculate table records
- Increased Security
  - Easy to limit which staff see which parts of the data