Unit for Medical Statistics

GUIDANCE FOR CODING AND PREPARING DATA FOR COMPUTER ENTRY

General tips

- Use an ID number rather than a name as the identifier to maintain confidentiality. The actual names and corresponding numbers should be stored separately and securely

- Even if the study is anonymous, still include an ID for each form for tracking purposes – sometimes data analysis can throw up a query which may be resolved if the specific original form can be checked

- Most statistical programs require data to be analysed as numeric data so categorical data needs to be coded

- Use intuitive codes if possible, eg, use 1/0 for yes/no such that responses given as 'yes' are coded 1 and those given as 'no' are coded 0. This also has the added advantage that the variable's 1/0 values can be simply summed to give the number of positive responses

- Use codes 1,2,3 etc where data fall into more than 2 categories

- If the first category is a 'null category' such as when recording pain as 'no pain, mild pain, moderate pain and severe pain', it may be sensible to use the codes 0,1,2,3

- **It is essential to keep a record of the codes for current and later reference** – for example in an adjacent spreadsheet

Dealing with missing data

- Missing data are sometimes given a special code such as 9 with the appropriate number of 9s such that the code could not be a real response. For example: for a yes/no response then 9 could indicate a missing value; for a height recorded in cm, 999 could be used, as this is not a possible value

- Computer packages may use a dot '.' to denote a missing value

- It may be important to distinguish between data which are simply missing from the original source and data which the data extractor failed to record. This can be achieved using different codes

- Sometimes a response to a question may be 'not applicable' such as when asking the number of cigarettes smoked when the respondent has already answered 'no' to a question about whether they currently smoke. It may be helpful to code such responses differently, for example using 8s rather than 9s.

Example of data with codes

Does the patient currently smoke?

□ Yes	(=1)
□ No	(=0)

This is a single yes/no question

Patient’s legal marital status

□ Married	(=1)
□ Single	(=2)
□ Widowed	(=3)
□ Divorced	(=4)
□ Separated	(=5)

This is a single question with multiple options

Patient’s self-reported pain level

□ None	(=0)
□ Slight pain	(=1)
□ Moderate pain	(=2)
□ Severe pain	(=3)

This is a single question with multiple options

Patient’s current medication for pain

□ TCA	(=0/1)
□ Anti-epileptic	(=0/1)
□ Topical analgesic	(=0/1)
□ Opioid	(=0/1)
□ NSAID	(=0/1)
□ Other	(=0/1)

This is a multiple question for which each option is no or yes since patients may be taking more than one drug or none at all

SETTING UP A SPREADSHEET FOR DATA ENTRY

Key points for most situations:

- Each column is a separate variable
- Each row is a separate subject
- Give variables meaningful names
- For serial measurements give them the same name with a different trailing digit eh height1 height2 etc
- Only use numeric data and code data before data entry. (This is not possible if free form text is needed but note that this will not be analysable).
- Don't mix comments in with the data in the same spreadsheet unless the comments are in a specific and labelled column. It is best to put any comments in a separate spreadsheet so the data can be easily transferred into a statistical program for analysis
- Don't split data sets up into separate spreadsheets if they are from the same study for example for men and women. If you have data for say, men and women, then define a variable that specifies 'sex' and join the data sets together
- Check the data for impossible and/or inconsistent values before analysis

Example of a spreadsheet with different types of data

idnum	sex	gestation	gestdays	bweight	smoking	apgar1
1	1	25+5	180	0.884	0	3
2	1	30+2	212	1.26	0	9
3	2	32+0	224	1.558	1	9
4	2	30+5	215	1.5	0	9
5	1	30+4	214	1.158	0	6

- The first row of the spreadsheet gives the variable names
- Each of the subsequent 5 rows give the data for each of 5 subjects
- Each column represents one variable
- *idnum* is the unique subject identifier
- *sex* denotes the sex of the baby and has 2 possible values, 1 and 2. To interpret the data we need to know the coding, ie, to know that in this case, 1=male, 2=female

- *gestation* is the gestational age of the baby and is recorded in 'weeks + days'. This format is commonly used for descriptive purposes but is not suitable for data analysis.

- *gestdays* is the gestational age in whole days and is suitable for data analysis

- *bweight* is the baby's birthweight in kg

- *smoking* is the smoking status of the mother and is recorded as 0/1 to indicate no/yes

- *apgar1* is the apgar score at 1 minute which can take any integer value between 0 and 10

- **Note that the variable names do not contain any spaces (eg 'painam' or 'pain_am' to represent morning pain score would be allowed but 'pain am' is not). In this way the variable names will be transfered directly into a statistical program**

Summary tips for data entry into spreadsheets

- Liaise with the analyst/statistician beforehand

- Use one row per subject

- Use one column per variable

- Don't leave gaps in the spreadsheet or insert comments amongst data – put any comments at the beginning or at the end

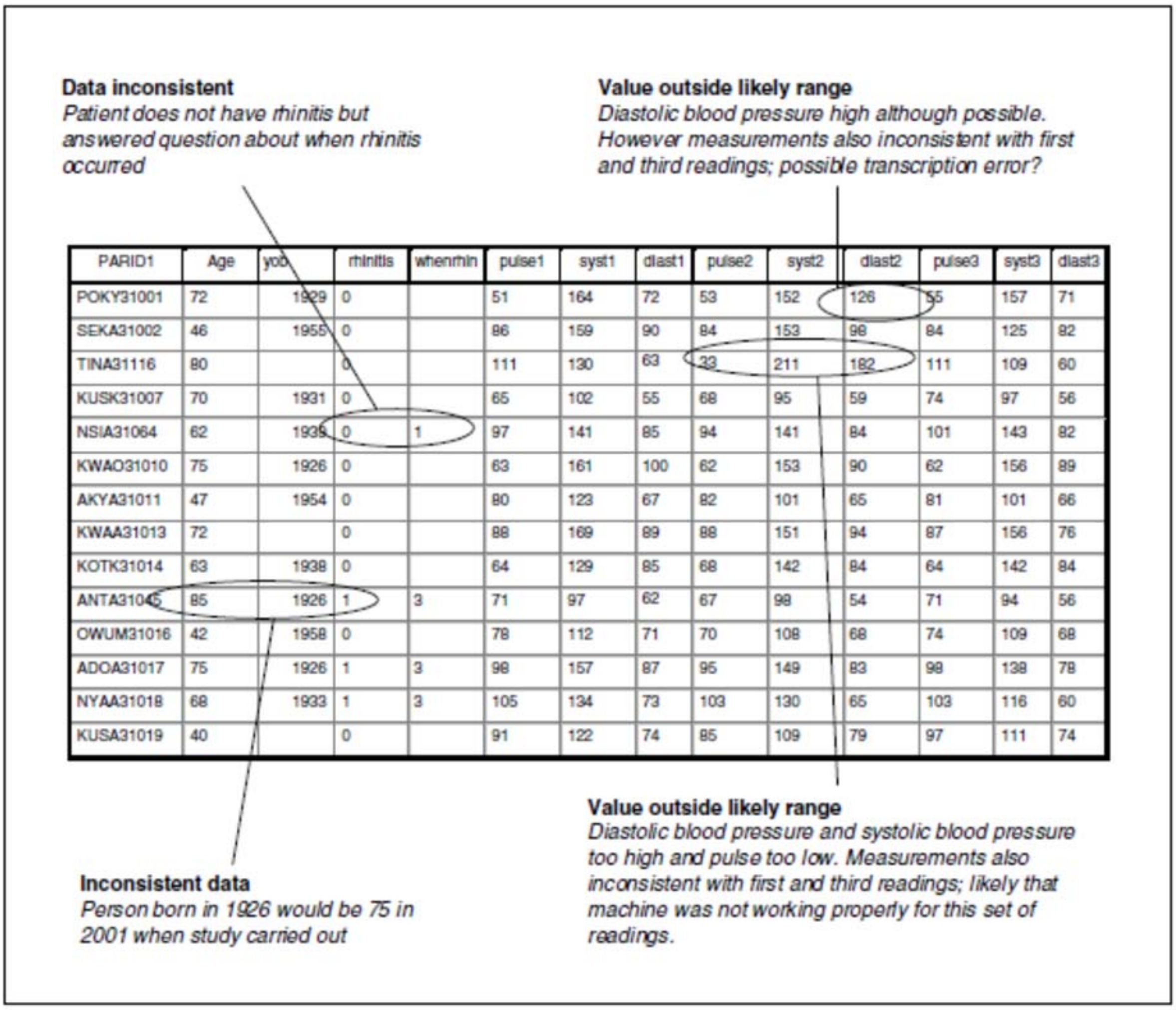
- Wherever possible avoid using non-numerical data in the cells

- Use a dot rather than a blank space to indicate any missing data unless there are specific codes for different types of missing data

- Keep a formal record of the coding used for each variable in the study

CHECKING DATA FOR ERRORS

Example



Data inconsistent
Patient does not have rhinitis but answered question about when rhinitis occurred

Value outside likely range
Diastolic blood pressure high although possible. However measurements also inconsistent with first and third readings; possible transcription error?

PARID1	Age	yob	rhinitis	whenrhin	pulse1	syst1	diast1	pulse2	syst2	diast2	pulse3	syst3	diast3
POKY31001	72	1929	0		51	164	72	53	152	126	55	157	71
SEKA31002	46	1955	0		86	159	90	84	153	98	84	125	82
TINA31116	80		0		111	130	63	33	211	182	111	109	60
KUSK31007	70	1931	0		65	102	55	68	95	59	74	97	56
NSIA31064	62	1939	0	1	97	141	85	94	141	84	101	143	82
KWAO31010	75	1926	0		63	161	100	62	153	90	62	156	89
AKYA31011	47	1954	0		80	123	67	82	101	65	81	101	66
KWAA31013	72		0		88	169	89	88	151	94	87	156	76
KOTK31014	63	1938	0		64	129	85	68	142	84	64	142	84
ANTA31045	85	1926	1	3	71	97	62	67	98	54	71	94	56
OWUM31016	42	1958	0		78	112	71	70	108	68	74	109	68
ADOA31017	75	1926	1	3	98	157	87	95	149	83	98	138	78
NYAA31018	68	1933	1	3	105	134	73	103	130	65	103	116	60
KUSA31019	40		0		91	122	74	85	109	79	97	111	74

Value outside likely range
Diastolic blood pressure and systolic blood pressure too high and pulse too low. Measurements also inconsistent with first and third readings; likely that machine was not working properly for this set of readings.

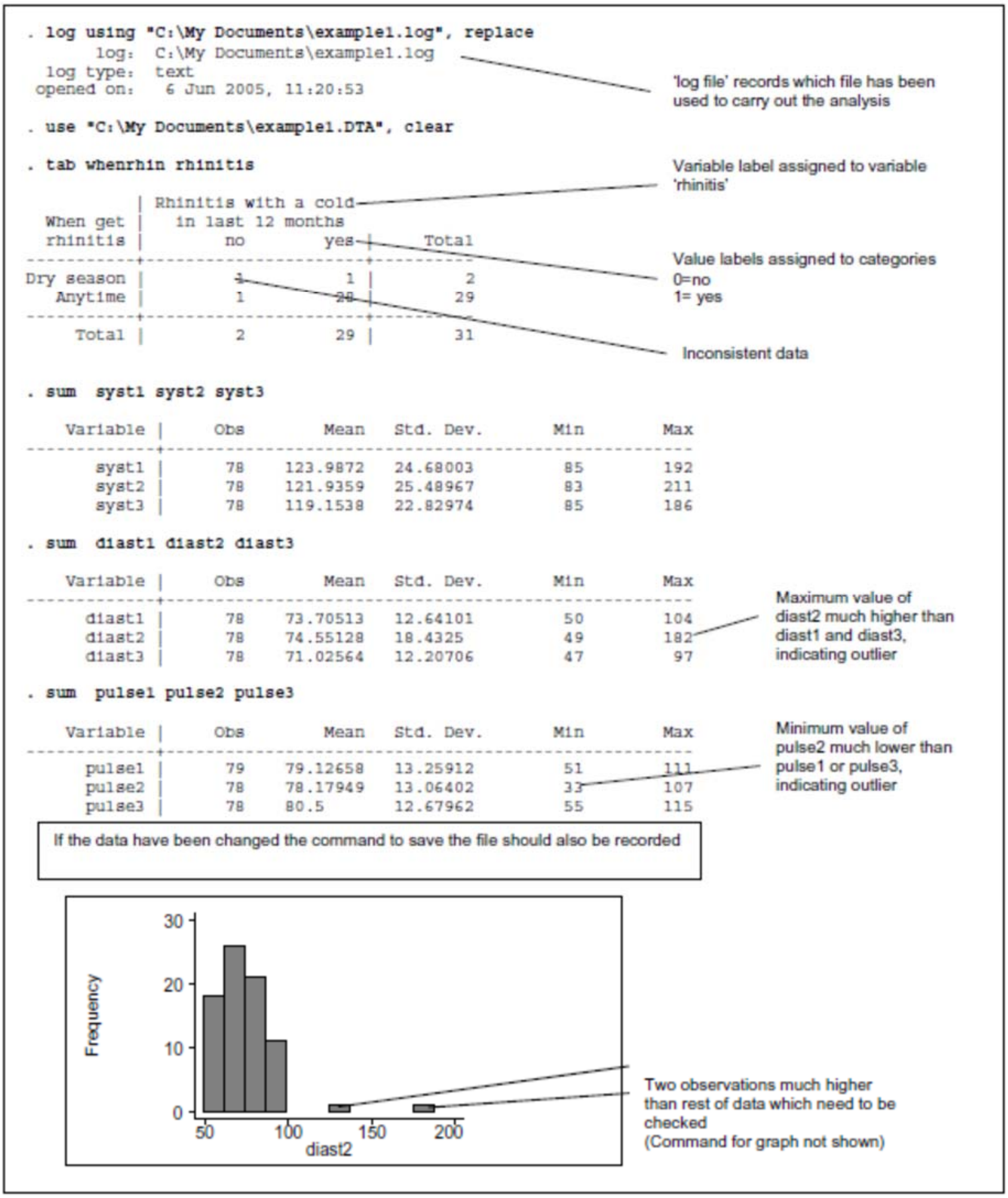
Inconsistent data
Person born in 1926 would be 75 in 2001 when study carried out

Examples of data checking using a statistical program (Stata used here but other programs will do this too)

Key things to look for to check for errors

- Extreme or outlying values
- Inconsistent values in a given subject

Remember to save the data when changes are made but always save each previous version, and keep a record of changes that are made



```
. log using "C:\My Documents\example1.log", replace
      log:  C:\My Documents\example1.log
 log type:  text
opened on:   6 Jun 2005, 11:20:53

. use "C:\My Documents\example1.DTA", clear

. tab whenrhin rhinitis

           | Rhinitis with a cold
  When get |   in last 12 months
  rhinitis |        no        yes |     Total
-----------+----------------------+----------
Dry season |         1          1 |         2
   Anytime |         1         28 |        29
-----------+----------------------+----------
     Total |         2         29 |        31

. sum  syst1 syst2 syst3

    Variable |       Obs        Mean   Std. Dev.       Min        Max
-------------+-------------------------------------------------------
       syst1 |        78    123.9872   24.68003         85        192
       syst2 |        78    121.9359   25.48967         83        211
       syst3 |        78    119.1538   22.82974         85        186

. sum  diast1 diast2 diast3

    Variable |       Obs        Mean   Std. Dev.       Min        Max
-------------+-------------------------------------------------------
      diast1 |        78    73.70513   12.64101         50        104
      diast2 |        78    74.55128    18.4325         49        182
      diast3 |        78    71.02564   12.20706         47         97

. sum  pulse1 pulse2 pulse3

    Variable |       Obs        Mean   Std. Dev.       Min        Max
-------------+-------------------------------------------------------
      pulse1 |        79    79.12658   13.25912         51        111
      pulse2 |        78    78.17949   13.06402         33        107
      pulse3 |        78        80.5   12.67962         55        115
```

'log file' records which file has been used to carry out the analysis

Variable label assigned to variable 'rhinitis'

Value labels assigned to categories
0=no
1= yes

Inconsistent data

Maximum value of diast2 much higher than diast1 and diast3, indicating outlier

Minimum value of pulse2 much lower than pulse1 or pulse3, indicating outlier

If the data have been changed the command to save the file should also be recorded

Two observations much higher than rest of data which need to be checked
(Command for graph not shown)