lec 2:

Biostatistic

Data entry

By the end of this lecture, you should be able to:

- Describe different formats for entering data on to a computer
- Outline the principles of questionnaire design
- Distinguish between single-coded and multi-coded variables
- Describe how to code missing values

When you carry out any study you will almost always need to enter the data into a computer package. Computers are invaluable for improving the accuracy and speed of data collection and analyses , making it easy to check for errors, produces graphical summaries of the data and generate new variables. It is worth spending some time planning data entry – this may save considerable effort at later stages.

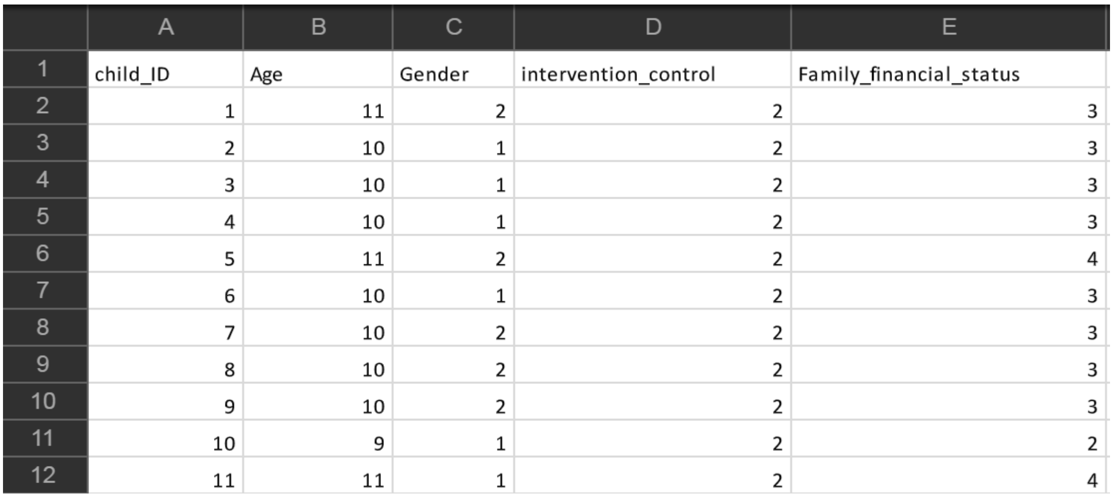
Planning data entry

When collecting data in a study you will often need to use a form or questionnaire for recording the data. If these forms are designed carefully, they can reduce the amount of work that has to be done when entering the data. Generally, these forms/questionnaires include a series of boxes in which the data are recorded – it is usual to have a separate box for each possible digit of the response.

Data entry

Sometimes, data is collected on paper forms, and we need to do data entry into a computer file in preparation for the data analysis.

The goal of any data entry process is to have data arranged in a spreadsheet, like this one:



	A	B	C	D	E
1	child_ID	Age	Gender	intervention_control	Family_financial_status
2	1	11	2	2	3
3	2	10	1	2	3
4	3	10	1	2	3
5	4	10	1	2	3
6	5	11	2	2	4
7	6	10	1	2	3
8	7	10	2	2	3
9	8	10	2	2	3
10	9	10	2	2	3
11	10	9	1	2	2
12	11	11	1	2	4

A well-arranged datasheet should satisfy the following characteristics:
1- Each column represents one variable.

If one variable is measured twice (as before and after an experiment), then it should be recorded in two columns.
If a variable consists of 2 elements (as blood pressure consisting of systolic and diastolic blood pressure), then each element should be recorded in a single column.
2- The unit of measurement is unified in each column.

Height is measured either in meter or in cm, can't be in meter for some patients, and in cm for others.
3- Each row represents a case

The case is the unit of which we collect data, as a patient, a rat, a village, a hospital, etc., depending on each study.
4- Each cell contains only one data point.

It can't include both systolic and diastolic blood pressure, or gestational age in weeks and days.
5- Nominal and ordinal data are coded using numeric codes.

We use numbers as codes for each category instead of writing the name of the category. For example, we may use 1 as code for males and 2 as code for females. Always keep a codebook for your coded variables where you can find the codes and corresponding values.
Coding of categorical data:
It is better to use numeric codes when entering categorical data, easier, less prone to typing mistakes, and more suitable for the statistical software packages.
It is better to use reasonable codes for each variable as in the following examples:
Severity of disease:

- Mild → 1
- Moderate → 2
- Severe → 3

Severity of Pain:

- No pain → 0
- Mild pain → 1
- Moderate pain → 2
- Severe pain → 3

If binary (Yes/No)

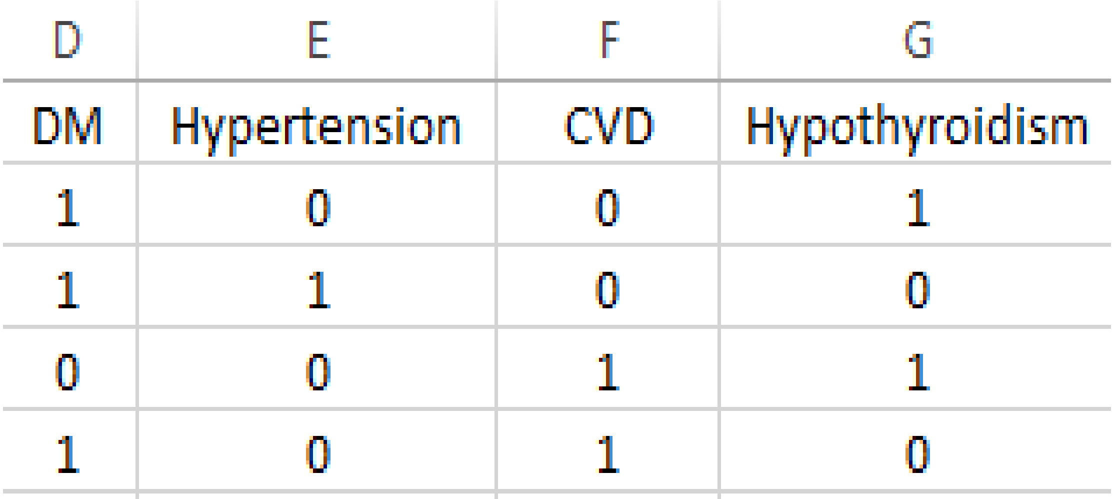
- Yes → 1
- No → 0

- If multiple answers are allowed for one question, use a column for each choice and code it as 1/0 representing Yes/No.

In the data collection form, asking about chronic conditions may be in this way:
Do you have any of the following chronic diseases?

- DM
- Hypertension
- CVD
- Hypothyroidism

But in the data entry, it should be like this:



D	E	F	G
DM	Hypertension	CVD	Hypothyroidism
1	0	0	1
1	1	0	0
0	0	1	1
1	0	1	0

If there is a variable with open answers or a large number of possible answers, we have to evaluate those answers and categorize them into a limited number of categories, so that we can include them in the statistical analysis.

Numerical data

Numerical data should be entered with the same precision as they are measured, and the unit of measurement should be consistent for all observations on a variable. For example, weight should be recorded in kilograms or in pounds, but not both interchangeably

Multiple forms per patient
Sometimes, information is collected on the same patient on more than one occasion. It is important that there is some unique identifier (e.g. a serial number) relating to the individual that will enable

Problems with dates and times

Dates and times should be entered in a consistent manner, e.g. either as day/month/year or month/day/year, but not interchangeably. It is important to find out what format the statistical package can read.

Coding of missing data

It is better to use codes for missing data instead of leaving the cells empty so that we are sure that it is a missing value and not a data entry mistake.

Use impossible values (as codes) that can't be correct for this variable.
For example:

Refused to answer and Not applicable are
not considered the same as missing (we give them other codes such as 998, 997).