Examining and displaying data

Contents

- Advice on entering data
- Describe and summarise data using tables and plots
- Checking the Gaussian (Normality) assumption for continuous data (taking measurements)
- Investigating missing data

Entering data: example of a data file

Patient ID	Date of death	Cause of death	Time until death
10001G	08/06/2006	Lung cancer	1140
10002h	09/09/1999	Lung cancer, metastatic	361
1003E	28-06-03	Heart disease	66 days
20001Y	08/10/2005	Breast Cancer	-999
202U	14/05/04	Stroke	844
20003F	14th July 2008	LUNG CANCER	5.1 years
20004q	13/01/2004	Stroke	326
20005K	14/05/2005	Lung carcinoma	> 6 months
20006H	02-28-2004	Heart attack	12.6 months

What do you think are the problems with this dataset?

Problems with this datafile

- Subject identifier uses a mixture of letters and numbers. It is usually best to use a simple number.
- Date of death in different formats. Be wary of American date formats (month/day/year), bottom row
- Also, for 14-May-04, Excel might not be able to distinguish whether '04' stands for 1904 or 2004, and an error flag is hidden in the cell; a stats package cannot read this date in as a date
- Lung cancer coded in different ways (lower and upper case letters; 'cancer' or 'carcinoma'). A stats package will read all these as different values. Use a numeric coding instead, and enter 1 to 4 in the datafile:
  - 1=lung cancer
  - 2=heart disease
  - 3=stroke
  - 4=breast cancer
- Row 2 has a comma in the cell. This will cause problems if you use a CSV (comma delimited) file. Try to avoid using commas in data values

Problems with this datafile

The 'time until death' column has several problems:

- -999 will be read as a number unless you tell stats package otherwise (use full stop instead)
- Text (eg 'years', '>') is used in the cell: numeric columns **must only contain numbers**, nothing else. A stats package will usually ignore the cell (and you might not know it has done this!)
- Different scales used, i.e. days or years. All the numbers must be on the same scale

- It is much better if you think carefully about your data before you enter it. Data entry and the statistical analysis will be easier
- It can be very time consuming to have to significantly edit a datafile because the stats package cannot cope with it (usually because of the problems listed above)

Patient ID	Date of death	Cause of death	Follow-up Period
10001G	08/06/2006	Lung cancer	1140
10002h	09/09/1999	Lung cancer, metastatic	361
1003E	28-06-03	Heart disease	66 days
20001Y	08/10/2005	Breast Cancer	-999
202U	14/05/04	Stroke	844
20003F	14th July 2008	LUNG CANCER	5.1 years
20004q	13/01/2004	Stroke	326
20005K	14/05/2005	Lung carcinoma	> 6 months
20006H	02-28-2004	Heart attack	12.6 months

Patient Trial ID	Date of rand.	Date of birth	Sex
20006h	10/02/2003	01/10/1937	male

Here, the 2 datasets would not be merged correctly, because the subject identifier is not identical

Entering data

- Data should be entered in columns, i.e. a column for each variable.
- All subjects/objects should be given a unique identifier to be used in all data sheets. You can then merge several data sets together and ensure that observations are matched correctly.
- Most calculations are performed with numerical variables so use numbers where possible.
- Columns that contain a letter or other non-numeric character could be read as a character/string variable. These may be harder to work with unless the spacing, case and spelling are kept consistent
- Dates are needed for time to event outcomes. Statistical packages can only deal with them if they have a consistent format within a column

Data checking

Before you start displaying or analysing your data, basic checks should be performed.

Look for obvious errors.

Do you really have a patient aged 789? This is probably a typo – a patient of 78, 79 or 89. This sort of error should be checked as early as possible, and corrected before it can affect your analysis.

Dates should also be checked, eg:

• Date of diagnosis should not be before date of treatment

•Date of death should not be before date of last follow up visit

Age	Frequency	Percent	Cumulative Percentage
66	25	8.96	8.96
67	29	10.39	19.35
68	35	12.54	31.90
69	28	10.04	41.94
70	30	10.75	52.69
71	21	7.53	60.22
72	20	7.17	67.38
73	19	6.81	74.19
74	28	10.04	84.23
75	14	5.02	89.25
76	10	3.58	92.83
77	7	2.51	95.34
78	5	1.79	97.13
79	3	1.08	98.21
80	2	0.72	98.92
83	1	0.36	99.28
84	1	0.36	99.64
789	1	0.36	100.00
Total	279	100	

Types of data file

Data file	Comments
Excel	Columns of data; easy to enter data; you can enter whatever you like in each column (unless you format the column); sometimes not easily read into statistics packages
CSV (comma delimited); Tab delimited	Each variable clearly separated by a comma or space; easily read by statistics packages; you can enter whatever you like in each column; problems when missing data are not identified as such
Database system (eg in SPSS, Access)	Easy to enter data; forces you to a fixed format for each column; if in Access you may need to output to a CSV or Tab delimited file to be read by stats package; if in SPSS, no need to transfer data

Displaying ‘counting people’ outcomes

- This can be summarized as:
  - The number of subjects (absolute frequency)
  - The percentage of subjects (relative frequency)
- And displayed as:
  - A frequency table
  - A bar chart

Presenting the data in tables allows you to easily see how many patients are in each group. The number of patients in each group will be one of the factors that influences the analyses.

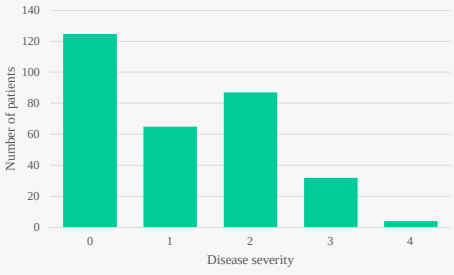
Frequency tables

Disease severity	% (N)
0	39.9 (125)
1	20.8 (65)
2	27.8 (87)
3	10.2 (32)
4	1.3 (4)
Total	**100 (313)**

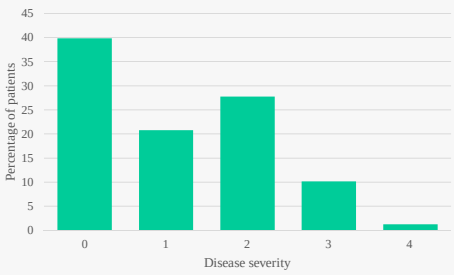
It is almost always better to compare groups using percentages, not absolute numbers (to allow for different groups with different numbers of subjects)

Displaying ‘counting people’ outcomes

Bar chart showing the absolute frequencies (observed counts)

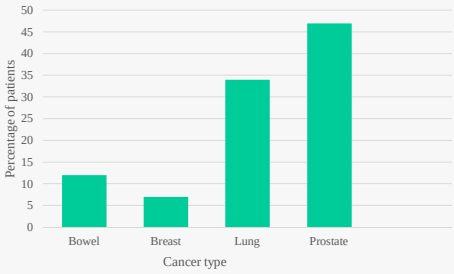


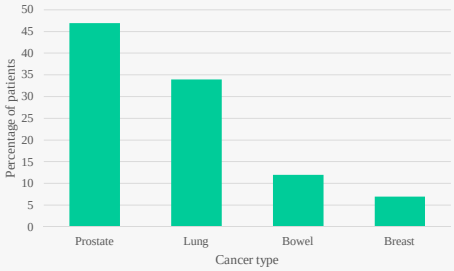
Bar chart showing the relative frequencies (percentages)



Always make sure the percentages in bar charts and tables (for each outcome) sum to 100 (not summing to 100 is a common mistake made, often spotted by an eagle eyed journal reviewer!)

Counting people endpoint with no natural ordering





Cancer	%
Bowel	12
Breast	7
Lung	34
Prostate	47

Which bar chart do you prefer, top (alphabetical order) or bottom (size of frequency)?

Combining cells

Disease severity	% (N)
0	39.9 (125)
1	20.8 (65)
2	27.8 (87)
3	10.2 (32)
4	1.3 (4)
Total	**100 (313)**

If a cell has a small number, then consider combining with an adjacent cell for the stats analysis (eg 3 & 4)

Disease severity	% (N)
0	39.9 (125)
1	20.8 (65)
2	27.8 (87)
3-4	11.5 (36)
Total	**100 (313)**

But don't combine too many cells, as this loses information

Disease severity	% (N)
0	37.5 (125)
1	19.5 (65)
2	26.1 (87)
3-4	10.8 (36)
Unknown	6.0 (20)
Total	**100 (333)**

Often better to keep a separate row when outcome is unknown

Displaying ‘taking measurements on people’ outcomes

- Histograms
- Scatter plots
- Box plots

Suppose we only have 1 group of people/things, and 1 measurement

Example: Weights (kg) of 100 teenagers

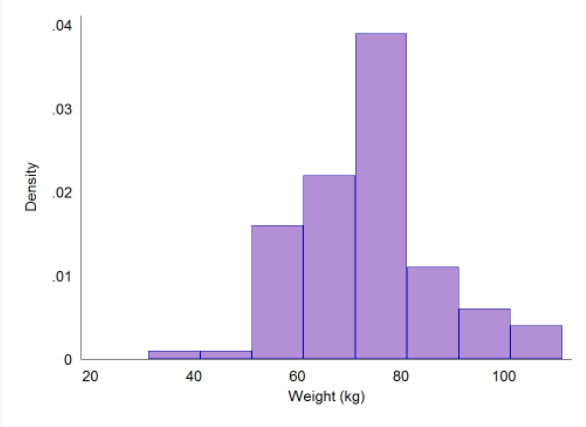
73.4, 66.8, 55.1, 51.6, 95.6, 98.6, 66.8, 43.7, 67.8, 62.3,
79.7, 58.5, 56.3, 62.2, 56.4, 77.6, 84.8, 62.8, 69.9, 58.6,
62.1, 56.9, 69.4, 87.6, 76.9, 71.3, 74.3, 72.2, 53.4, 71.8,
55.8, 55.8, 59.8, 57.9, 103.2, 101.2, 65.6, 71.8, 72.2, 52.2,
99.0, 80.9, 55.8, 54.2, 78.5, 70.1, 80.8, 61.3, 77.9, 58.6,
75.9, 67.5, 69.1, 71.4, 63.5, 93.3, 83.5, 88.3, 96.0, 71.5,
31.2, 91.1, 79.9, 63.3, 76.9, 76.1, 81.8, 74.1, 75.9, 80.6,
87.9, 74.8, 81.1, 70.4, 72.4, 74.6, 105.7, 108.1, 92.5, 62.3,
80.3, 75.6, 81.6, 78.8, 68.5, 83.9, 80.3, 73.2, 72.8, 80.4,
74.2, 73.4, 74.0, 70.7, 62.9, 85.9, 69.3, 81.5, 80.7, 72.2

The easiest way to display this, is to turn this continuous measurement into a categorical one

But choose sensible/obvious categories

Histograms

Weight (kg)	Number
30 ≤ w < 40	1
40 ≤ w < 50	1
50 ≤ w < 60	16
60 ≤ w < 70	19
70 ≤ w < 80	34
80 ≤ w < 90	18
90 ≤ w < 100	7
100 ≤ w < 110	4
Total	**100**



A histogram looks like a bar chart, except there are no gaps between the bars, because it represents a continuous variable (rather than a discrete variable, as with categorical data). Each block is called a 'bin'.

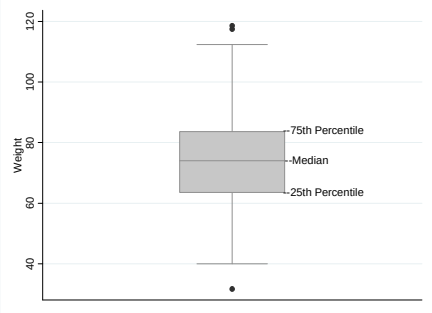
- Although a nice way to see the data, categorising the data can lose information (eg variability) when you do the statistical analyses
- Try the analyses both ways (using the continuous and categorical measure) to see if conclusions are similar
- Sometimes you can get different results (in which case it is often best to use the continuous measurement)

Continuous measurements

- If we want to compare a measurement between ≥2 groups of people/things, we can use:
  - Box plots (shows a summary of the data)
  - Scatterplots (shows all the observed data values)

Boxplots

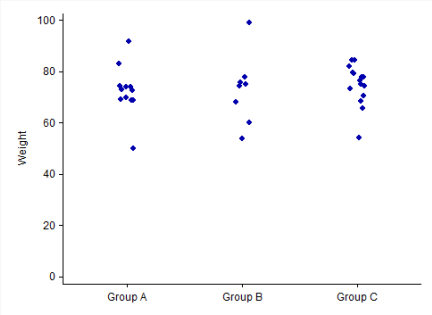
- Shows the median, 25th and 75th percentiles as well as the highest and lowest values.
- Outliers are any point which lies more than 1.5 x interquartile range from the 25th or 75th percentile, and are shown as separate data points.

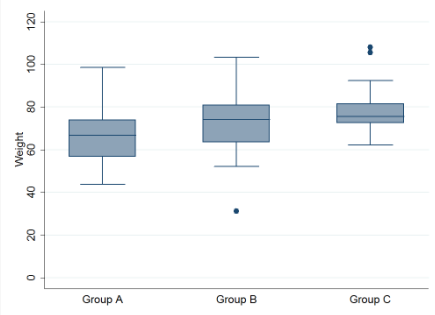


Displaying continuous data: box and scatter plots

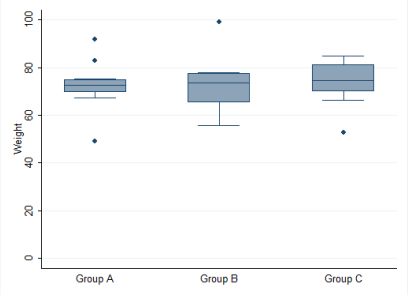
Group	N	Mean	Std. Deviation
A	32	67.4	12.7
B	37	73.8	15.1
C	31	78.4	10.1
Total	100	73.2	13.6

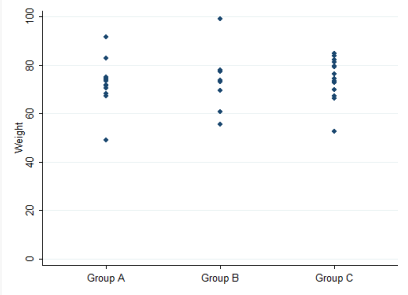
A scatter plot and a boxplot of the weight (in kg) of 54 patients, divided into three groups.





Displaying continuous data: box and scatter plots





Boxplots can be difficult to interpret, or misleading, when the group sizes are small. Here the groups contain 12, 8, and 15 people, and so it would be better to display all the data (right).

Checking for Gaussian (Normal) distribution

•Many statistical tests rely on the assumption that continuous data (i.e. taking measurements on people) is Normally distributed.

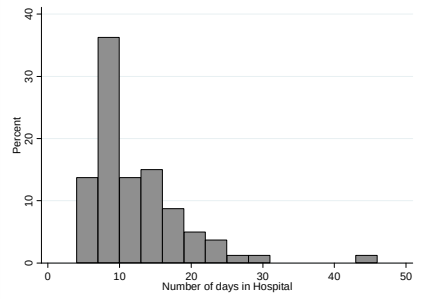
•A histogram can help, but you need many subjects in order to reliably assess whether the shape is symmetric or not. Histograms do not work well with small samples (eg n<50), and may be sensitive to the width of the bins.

• It's best to use a Normal probability plot (sometimes called a Q-Q plot); easy to interpret, even with a small number of subjects

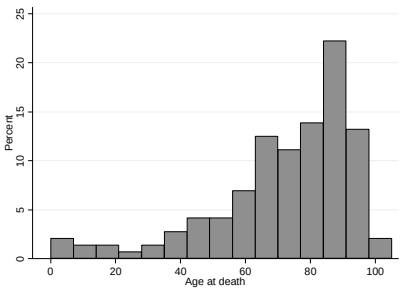
Histograms

Histograms can be used to get a quick idea of the shape of the distribution.

- Variables which are Normally distributed should have a "bell-shaped" curve.
- Variables which are skewed will have a non-symmetric shape, with fewer observations in one tail than the other.



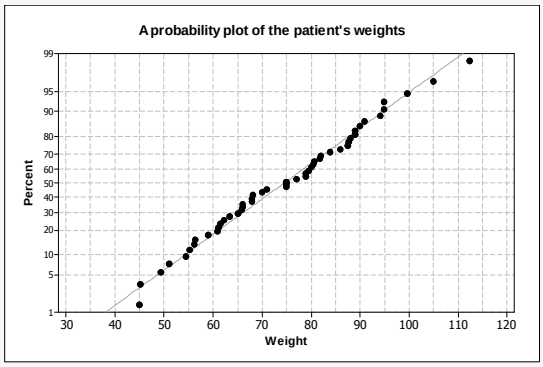
Positively skewed data: A histogram of the number of days spent in hospital



Negatively skewed data: A histogram of the age at death (from all causes).

Normal probability plots (for ‘taking measurements’ data)

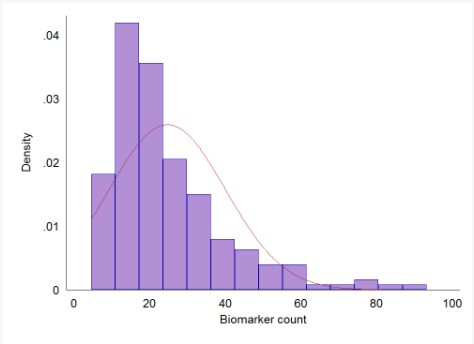
- A normal probability plot is a graphical technique used to decide whether the a set of continuous data is approximately normally distributed.

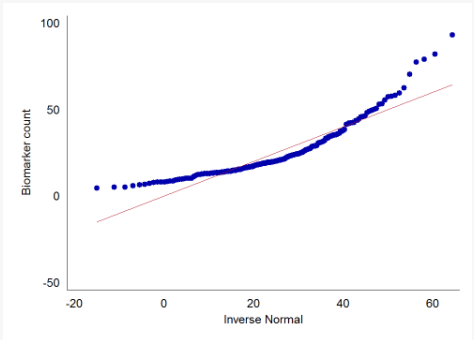


•The points should lie <u>roughly</u> on a straight line.

•Your software package will plot this for you!

What do we do if the data is not Normally distributed?

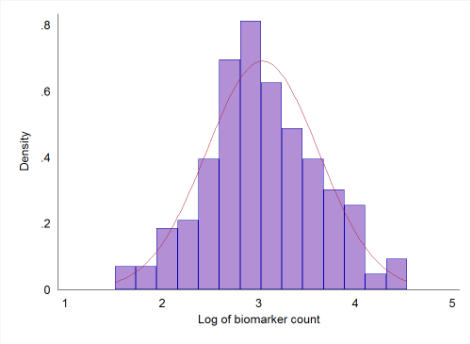


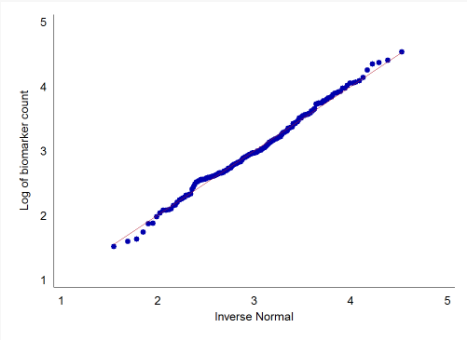


Log transformation:

(other transformations are taking the square root, or reciprocal)

Remember: any conclusions you make from this analysis are on the log of the variable of interest. You'll need to transform back to the original scale before you present your results.





Missing data

- People often use numbers like –9 or -99 (or 09/09/1999 for dates) to indicate missing values, or leave the spreadsheet cell blank
- But you must remember to tell the statistics package that these are missing codes, otherwise it will read them as normal numbers or dates
- When entering data into a spreadsheet or database, it is always best to put a **full stop** to indicate all missing data (all stats packages automatically treat this as missing; no problems with reading data)
- Missing data is almost always a problem when analysing data and can lead to misleading conclusions
- Always try to **avoid** missing data rather than do complicated statistical analysis to handle it. It makes interpretation a bit (lot) simpler

Missing Data

- Why is it missing?
- Design of study was not good enough: follow-up time too long
- Poor design of data collection forms (CRFs)
- Too many time points
- Clinical reason (eg patients are too ill to attend a clinic assessment)

- Is there a specific time point when its always missing?
- Can you go back to records etc and try to get the missing data (always a good option, if possible)?

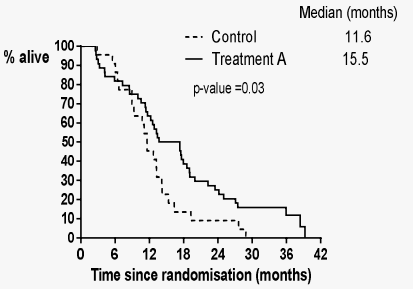
Missing Data

- Missing Completely at Random (MCAR); the missingness does not depend on any factor (therefore OK to just ignore)
- Missing at Random (MAR); the missing data for a factor depends on another factor, but not itself. Eg. in a study of alcohol consumption (factor of interest is alcohol), males are more likely than females to not report their habits. Sometimes possible to ignore these data (as with MCAR)
- Missing Not at Random (MNAR); the missing data for a factor is related to the factor. Eg. in a study of alcohol consumption (factor of interest is alcohol), heavy drinkers are more likely to not report their habits. Such missing data cannot be ignored (there are statistical methods, some complex, that can deal with it)

The importance of checking data, especially in small studies

Randomised trial of 24 patients in Control group, and 44 in Treatment A

Initial analysis



After data errors (wrong survival times) for ~5 patients were corrected