Describing the data

Dr Moses Ngari

Statistics

- A science pertaining to the collection, analysis, interpretation & presentation of data.

How do we achieve this?

- Describing the data, basic features and obtain simple summaries, freq

Descriptive statistics

- Making inferences about a pop from the data(sample)

Inferential statistics

What is data

- Data → Observations → Variables
- Typically, observations are rows and variables are columns

ID	Age	Sex	Hb	….	….	i^{th} var
001	23	M	7.5			
002	34	F	8.7			
003	29	F	9.7			
…	…	…	…			
N^{th} obs						

Classes of variables

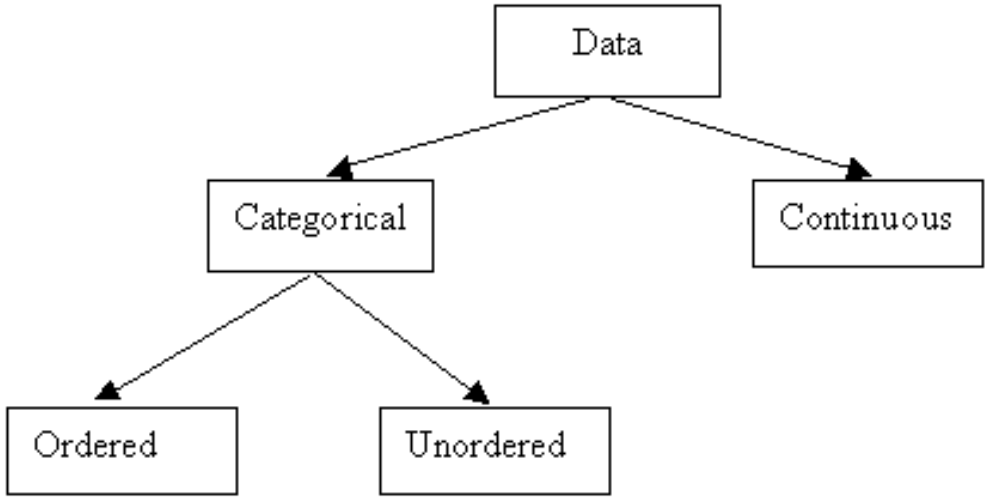
Outcomes & Exposures

- An **Outcome** is the variable of interest
  - E.g. death, illness, recovery of health, positive on laboratory test

- An **Exposure** is a variable that we wish to investigate *in relationship to* an outcome
  - E.g. age, weight, infected with *p. falciparum*

Variable types

- Variables have a type



- The type of a variable determines how we explore them

Categorical variables

- Data that consist of only **small number** of values corresponding to a **specific category value or label**.

Ask yourself whether you can state out loud all the possible values of your variable without taking a breath. If you can, you have a pretty good indication that your data are categorical. In a study of breast feeding in pre-term infants, there are a variety of categorical variables:

- Sex (male/female);
- Working status(employed, not employed);
- Education level (None, primary, secondary, tertiary)
- Marital status (single, married, divorced, widowed)

Continuous variables

- Data that consist of a **large number of values**, with **no particular category label** attached to any particular data value.

Ask yourself if your data can conceptually take on any value inside some interval. If it can, you have a good indication that your data are continuous. In the same study of breast feeding in pre-term infants, there are a variety of continuous variables:

- the infant's **birth weight in grams**;
- **Age in years**;
- Math scores from 0 to 100
- CD4 counts
- Haemoglobin level in g/dl

Describing Data

- Distributions
  - Tables and Graphs
- Statistics
  - Mean, median
  - Range, quartiles, variance & standard deviation

Measures of Data: Completeness and validity

Prior to any kind of statistical analysis, one must

1. Determine completeness of data
2. Identify data elements that are questionable
3. Recode questionable data elements to “missing” or “unknown” values

Data types and measures

- The way that we can summarize data depends on the type of the data
  - e.g. what is the range of a categorical variable?
    - Does order change this?

- However, we can always show a frequency distribution

Measures of data

- Distribution
- Centrality
- Variability

Distribution of data

Continuous variable

Value	Frequency	Percent	Cumulative Percent
20	1	0.03	0.03
21	4	0.13	0.17
22	3	0.10	0.27
23	4	0.13	0.40
24	3	0.10	0.50
25	1	0.03	0.53
35	1	0.03	0.57
36	3	0.10	0.67
37	1	0.03	0.70
38	6	0.20	0.90
39	2	0.07	0.97
40	1	0.03	1.00

Distribution of data

Categorical variable

Value	Frequency	Percent	Cumulative Percent
Male	15	50	50
Female	15	50	100

Statistical measures
Centrality

There are three averages

1. Mean
   - What we all “know” as the average

   i.e. sum of values/number of values
2. Median
   - The value that divides a distribution into two equal parts
3. Mode
   - Shoe sizes

Measures of central tendency

12, 0, 5, 13, 0, 0, 5, 10, 5, 1, 5, 6, 7, 5, 7, 8, 10, 5, 11,14

- arrange in order

0, 0, 0, 1, 5, 5, 5, 5, 5, 5, 6 , 7, 7, 8, 10, 10, 11, 12, 13,14

- Mode is most frequently occuring number
  - mode rarely used
- Median is value that divides data in half

Measures of central tendency

12, 0, 5, 13, 0, 0, 5, 10, 5, 1, 5, 6, 7, 5, 7, 8, 10, 5, 11,14

- Mean
  - sum of all observations divided by number of observations

$$\frac{\sum xi}{n}$$

Library books illustrations

- Students borrow and return books from a university library regularly

14, 13, 12, 11, 17, 20, 14, 16, 12, 12, 11, 9, 18, 21

- One student borrows a book and forgets to return it
- They re-discover the book 1 year later when moving out of the student accommodation and decide to return the book

14, 13, 12, 11, 17, 20, 14, 16, 12, 12, 11, 9, 18, 21, 365

- What do you think will happen to the mean and median of a data set on borrowing periods?

library books illustration

- The mean is the preferred measure of central tendency when describing a data that do not have outliers.
- A major disadvantage is that it is affected by outliers (i.e. single observations which are very extreme compared with most observations and whose inclusion or exclusion changes results noticeably). -In the presence of outliers, the median is the preferred measure of central tendency

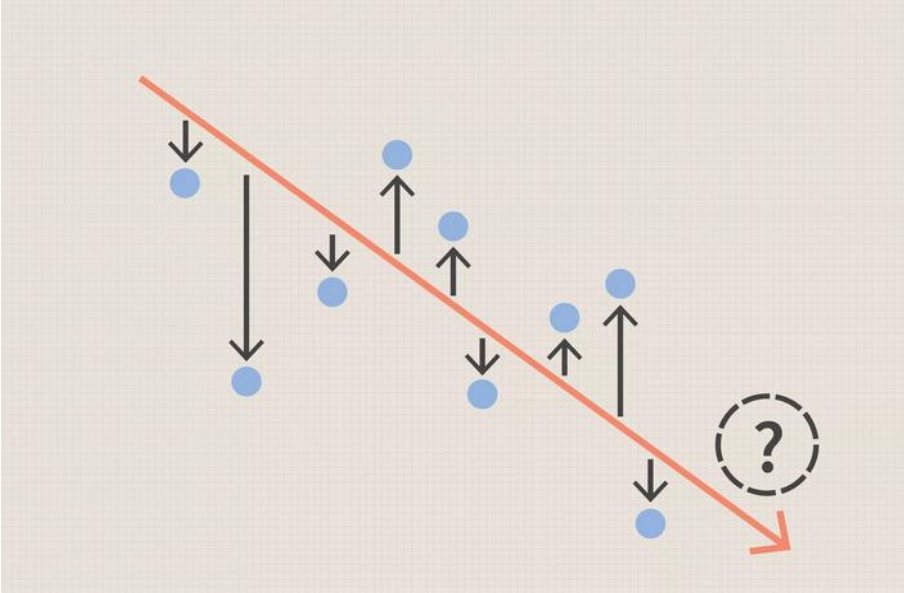
library books illustrations

- Calculation of the median does not involve the use of all available data and is therefore has less power than the mean.

Measures of dispersion

- They give us an idea of the variation or spread of values around the central one.
  - Variance and standard deviation
  - Range
  - Interquartile range

Variance



$$\frac{\sum(xi - x)^2}{(n - 1)}$$

Variance

- Can be defined in terms of how close the scores are to the middle of the distribution(mean)
- This variability or variance can be measured in terms of how far observations are from the mean on average i.e. how far, on average, each observation deviates from the mean.
- variance = by dividing the sum of squares of these deviations by (n-1).
- The formula for the variance is:

$$\frac{\sum(xi - x)^2}{(n-1)}$$

Variance

Xi	Xi-x (mean=33)	$(x_i - x)^2$
10	-23	529
20	-13	169
20	-13	169
20	-13	169
30	3	9
30	3	9
40	7	49
50	17	289
50	17	289
60	27	729
Total		2,410

$$\text{Variance} = \frac{\Sigma(x_i - x)^2}{(n-1)}$$

2410/9 = 267.77 = 268

Variance is in square units then this variance is 268 square cm.....kinda weird to talk of square cm

Standard deviation

- Standard Deviation (SD): A measure of the average spread of values about the mean.
- It is usually more convenient to express the variation in terms of the original, unsquared units (e.g. grams), i.e. to take the square root of the variance.
- This is then called the standard deviation (SD).
- A small standard deviation indicates that most values lie very close to the mean

Variance to standard deviation

Xi	Xi-x (mean=33)	$(x_i - x)^2$
10	-23	529
20	-13	169
20	-13	169
20	-13	169
30	3	9
30	3	9
40	7	49
50	17	289
50	17	289
60	27	729
Total		2,410

$$\text{Variance} = \frac{\Sigma(x_i - x)^2}{(n-1)}$$

2410/9 = 267.77=268

Variance is in square units
, then this variance is 268 square cm

Standard deviation is √268 = 16.4 cm

Often data summarised as

Mean ± standard deviation

In this case it's 33 ± 16.4

Measures of dispersion: Range

- The interval between the largest and smallest

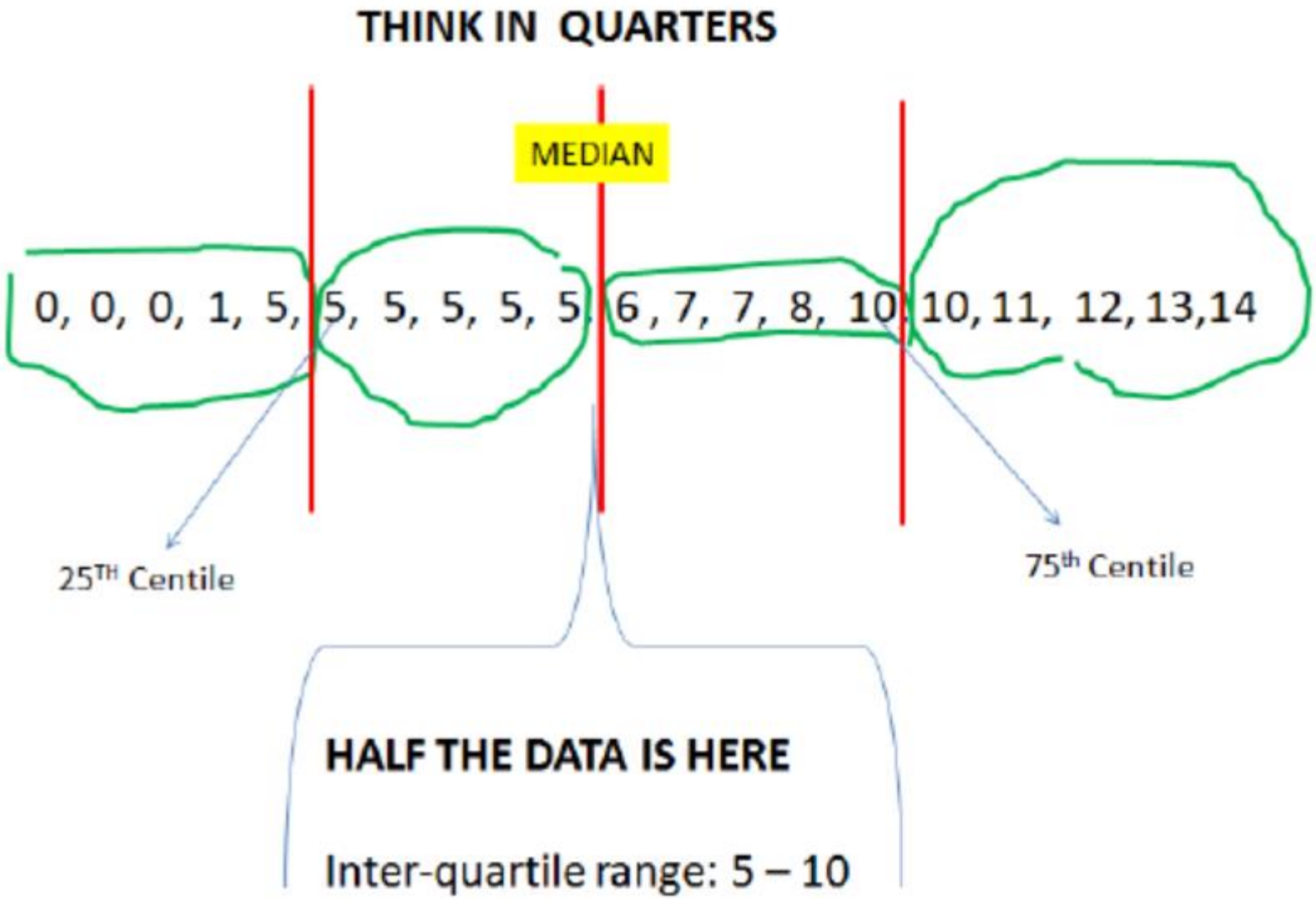
0, 0, 0, 1, 5, 5, 5, 5, 5, 5, 6 , 7, 7, 8, 10, 10, 11, 12, 13,14

- Range is
- based on only two observations and gives no idea of how the observations are arranged between these two.

Measures of dispersion: Inter-Quartile Range

- Shows spread of the middle 50% of the distribution

0, 0, 0, 1, 5, 5, 5, 5, 5, 5, 6 , 7, 7, 8, 10, 10, 11, 12, 13,14


THINK IN QUARTERS
MEDIAN
0, 0, 0, 1, 5, 5, 5, 5, 5, 5, 6, 7, 7, 8, 10, 10, 11, 12, 13,14
25TH Centile
75th Centile
HALF THE DATA IS HERE
Inter-quartile range: 5 – 10

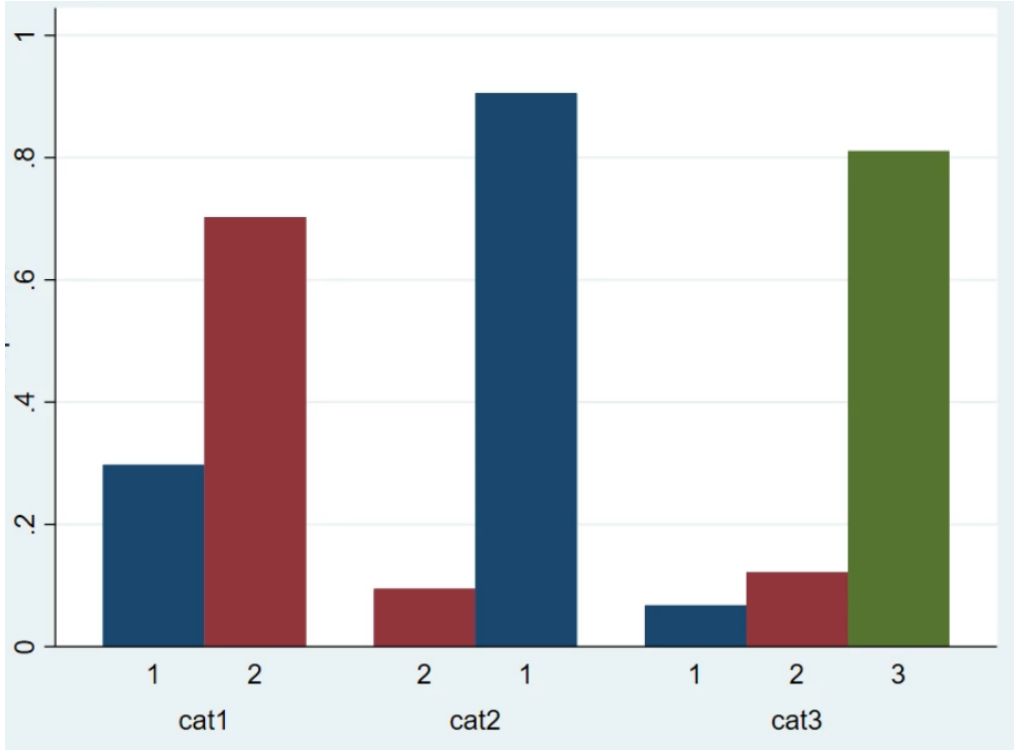
Graphical exploration of data

- Bar graphs (for categorical data).
- Histograms to show the shape of continuous data
- Cumulative distribution frequency
- Box and whisker graphs

- For each graph we can compare our data with what we expect from the population
  - Frequency of values.
  - Distribution or shape.

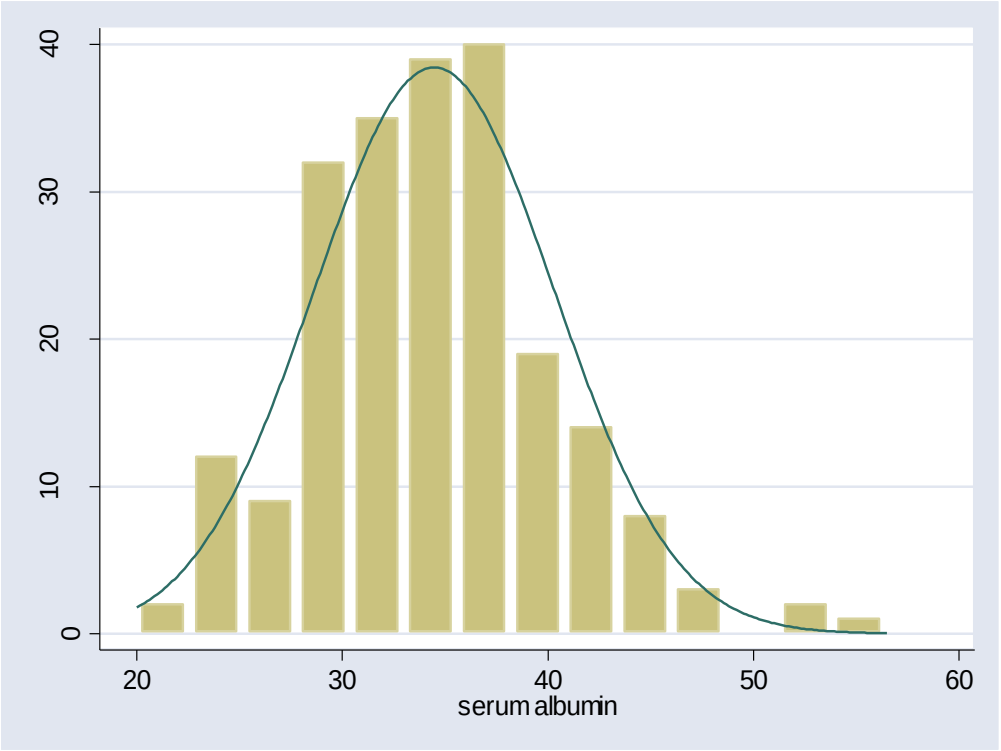
Bar charts for categorical data

- Bar charts show the number or frequency in each category.



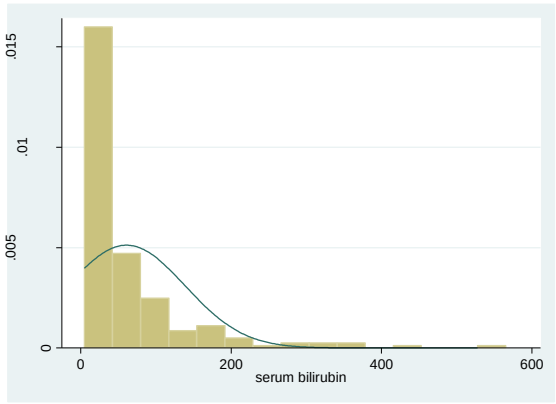
Histograms of continuous data

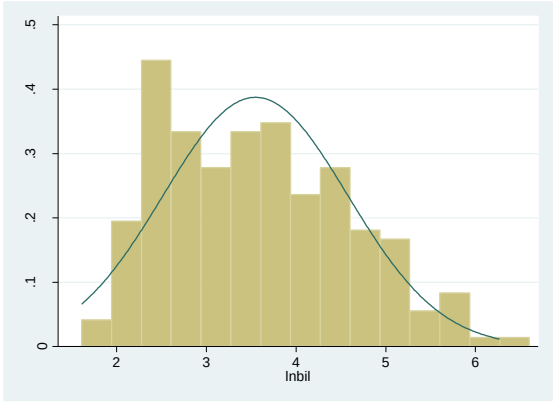
- For continuous variables, a histogram is the best way to explore the assumption of normality.
- Can show number of frequency (the shape is the same)



Histograms of continuous data

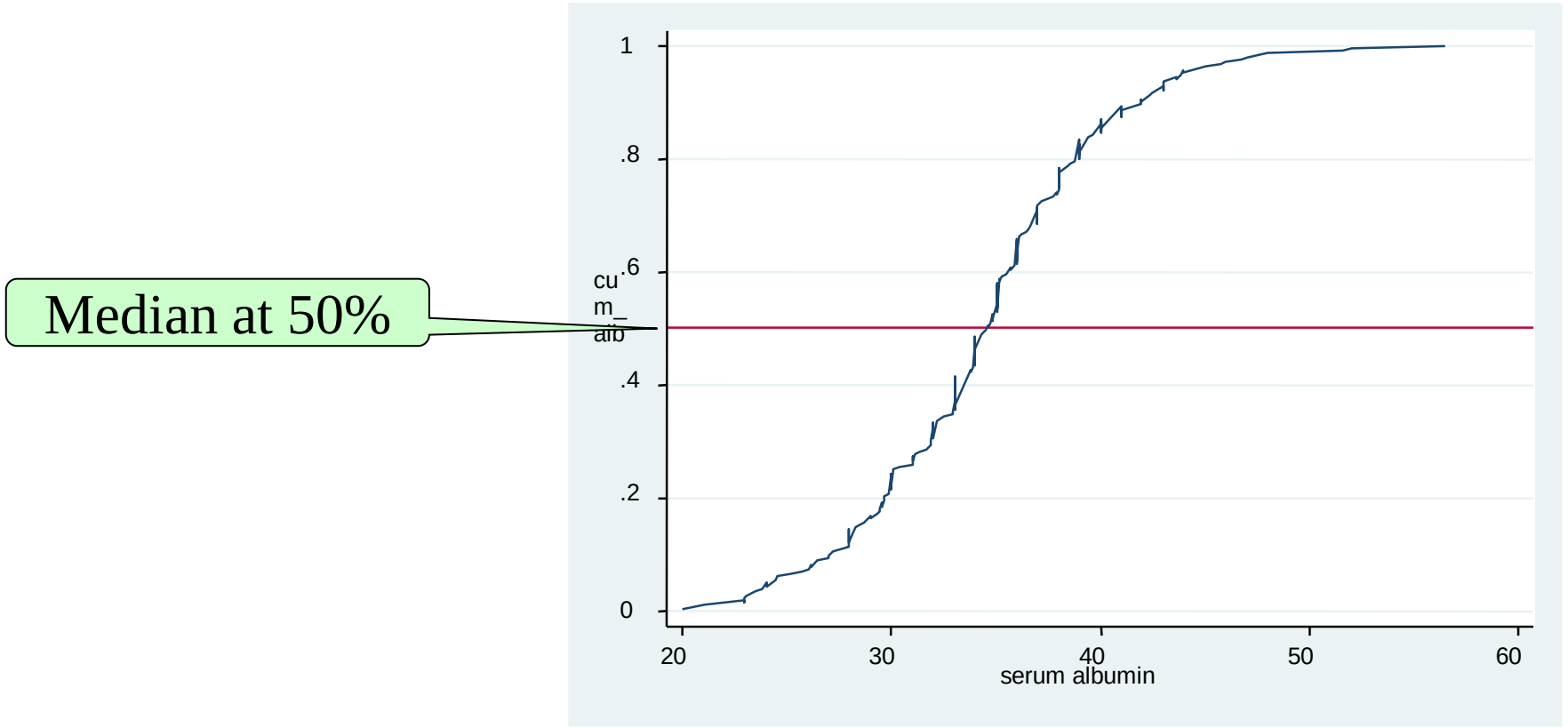
- For counts, cytokine responses and other biological data, a log transformation may be needed





Cumulative distribution functions

- A cumulative frequency graph can be shown. Need to generate a new variable with the cumulative distribution function (CDF)



Box and whisker graphs

- A box and whisker graph shows the median, interquartile range and extreme values.
- It is suitable for skewed data, where it may not be correct to show mean and standard deviations.