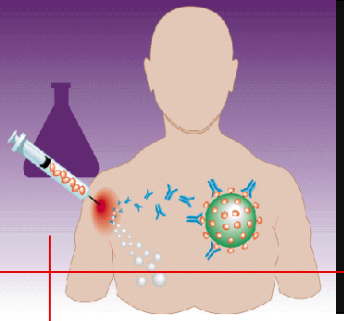
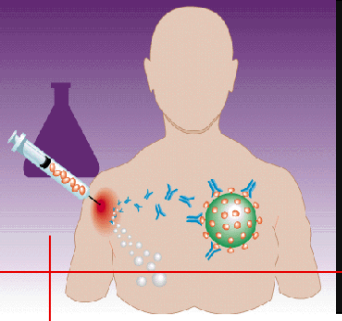
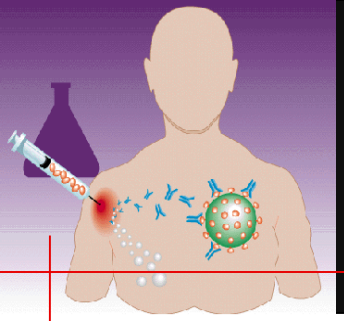


Summarise

- Summarise a large set of data by a few meaningful numbers.
- Single variable analysis
  - For the purpose of describing the data
  - Example; in one year, what kind of cases are treated by the Psychiatric Dept?
  - Tables & diagrams are usually used to describe the data
  - For numerical data, measures of central tendency & spread is usually used

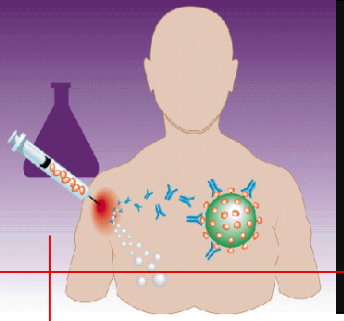


Measurement of Central Tendency & Spread



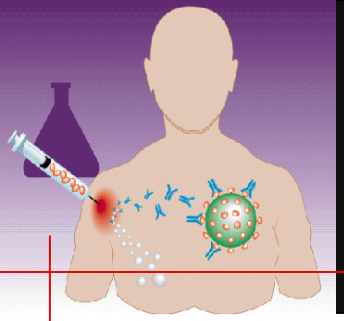
Measures of Central Tendency

- Mean
- Mode
- Median



Measures of Variability

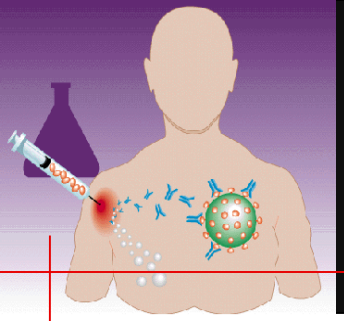
- Standard deviation
- Inter-quartiles
- Skew ness & kurtosis



Mean

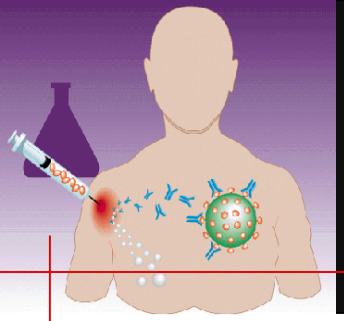
- the average of the data collected
- To calculate the mean, add up the observed values and divide by the number of them.
- A major disadvantage of the mean is that it is sensitive to outlying points

$$\bar{x} = \frac{(\Sigma x)}{n}$$



Mean: Example

- 12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35, 37, 38, 41, 43, 44, 46, 53, 58
- Total of x = 648
- n= 20
- Mean = 648/20 = 32.4

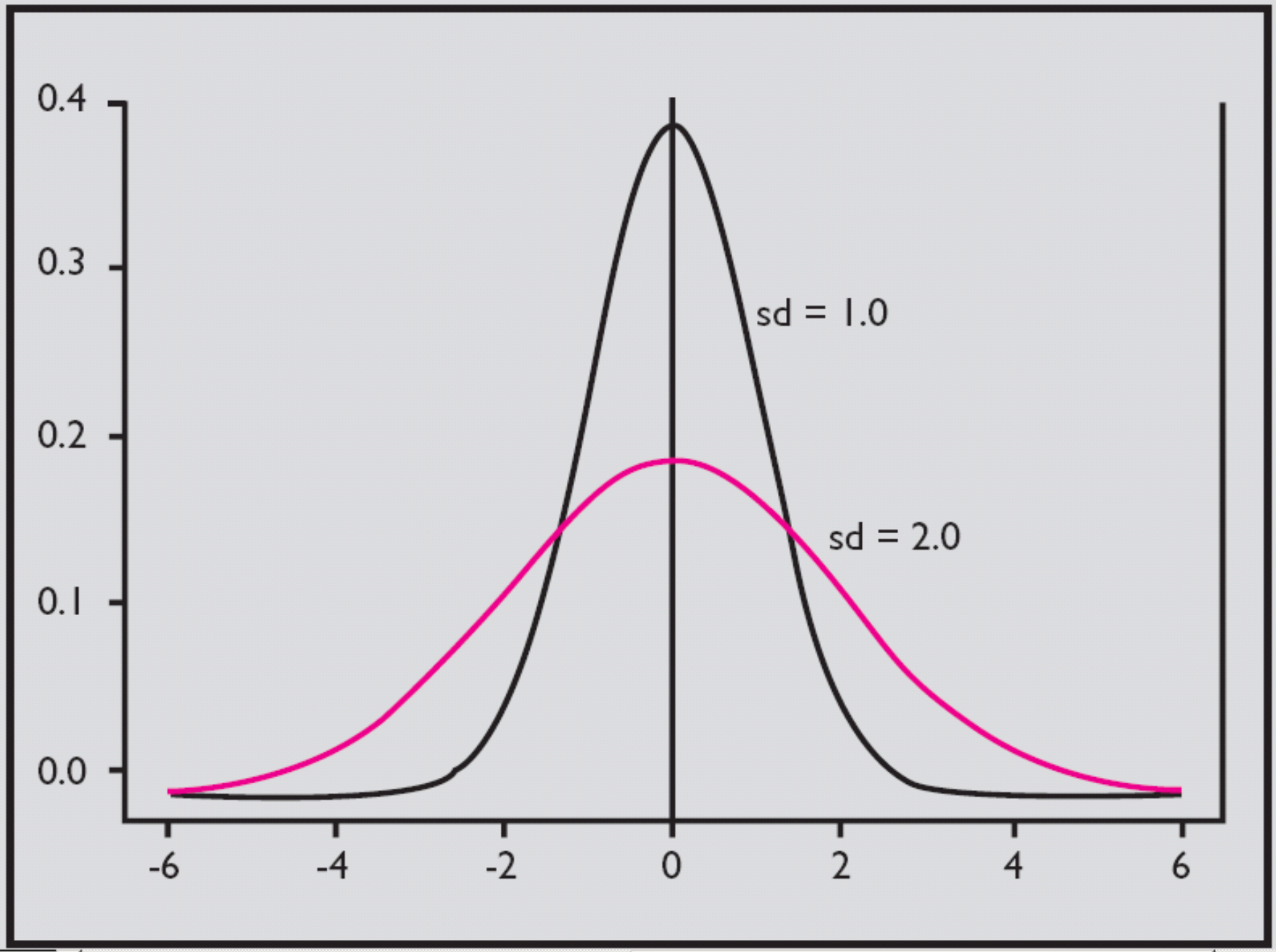


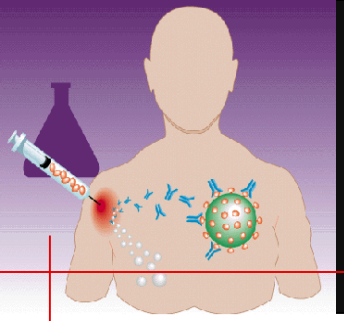
Measures of variation - standard deviation

- tells us how much all the scores in a dataset cluster around the mean. A large S.D. is indicative of a more varied data scores.
- a summary measure of the differences of each observation from the mean.
- If the differences themselves were added up, the positive would exactly balance the negative and so their sum would be zero.
- Consequently the squares of the differences are added.

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

$$\text{SD} = \sqrt{\frac{\Sigma(x - \bar{x})^2}{n - 1}}$$

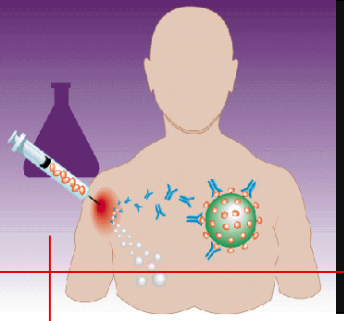

0.4
0.3
0.2
0.1
0.0
sd = 1.0
sd = 2.0
-6
-4
-2
0
2
4
6



sd: Example

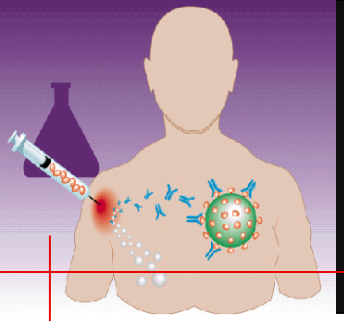
- 12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35, 37, 38, 41, 43, 44, 46, 53, 58
- Mean = 32.4; n = 20
- Total of $(x\text{-mean})^2$ = 3050.8
- Variance = 3050.8/19 = 160.5684
- sd = $160.5684^{0.5}$=12.67

x	(x-mean)^2	x	(x-mean)^2
12	416.16	32	0.16
13	376.36	35	6.76
17	237.16	37	21.16
21	129.96	38	31.36
24	70.56	41	73.96
24	70.56	43	112.36
26	40.96	44	134.56
27	29.16	46	184.96
27	29.16	53	424.36
30	5.76	58	655.36
TOTAL	1405.8	TOTAL	1645



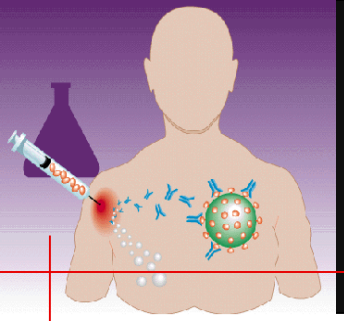
Median

- the ranked value that lies in the middle of the data
- the point which has the property that half the data are greater than it, and half the data are less than it.
- if n is even, average the n/2th largest and the n/2 + 1th largest observations
- "robust" to outliers



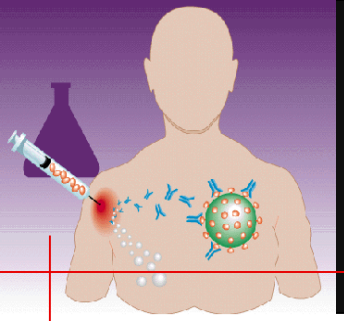
Median:

- 12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35, 37, 38, 41, 43, 44, 46, 53, 58
- (20+1)/2 = 10^{th} which is 30, 11^{th} is 32
- Therefore median is (30 + 32)/2 = 31



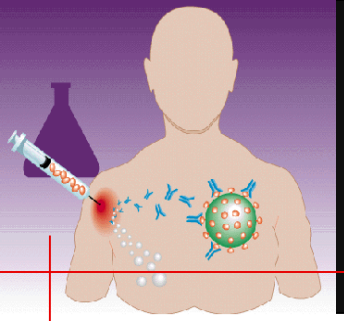
Measures of variation - quartiles

- The range is very susceptible to what are known as *outliers*
- A more robust approach is to divide the distribution of the data into four, and find the points below which are 25%, 50% and 75% of the distribution. These are known as *quartiles,* and the median is the second quartile.



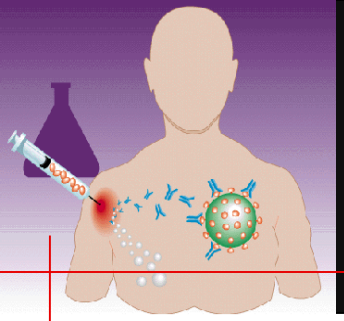
Quartiles

- 12, 13, 17, 21, 24,
  24, 26, 27, 27, 30,
  32, 35, 37, 38, 41,
  43, 44, 46, 53, 58
- 25th percentile 24; (24+24)/2
- 50th percentile 31; (30+32)/2 ; = median
- 75th percentile 42; (41+43)/2
- median 31(24;42)



Mode

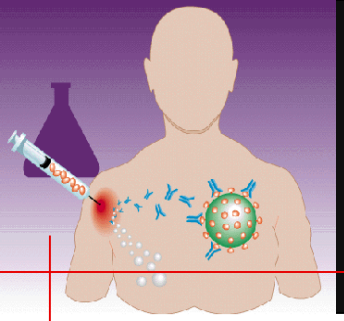
- The most frequent occurring number. E.g. 3, 13, 13, 20, 22, 25: mode = 13.
- It is usually more informative to quote the mode accompanied by the percentage of times it happened; e.g., the mode is 13 with 33% of the occurrences.



Mode: Example

- 12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35, 37, 38, 41, 43, 44, 46, 53, 58

- Modes are 24 (10%) & 27 (10%)



Mean or Median?

- Which measure of central tendency should we use?
- if the distribution is normal, the mean±sd will be the measure to be presented, otherwise the median±IQR should be more appropriate. (i.e. 31(24;42))
- IQR = 75^{th} centile – 25^{th} centile
  (i.e. 42 – 24 = 18)

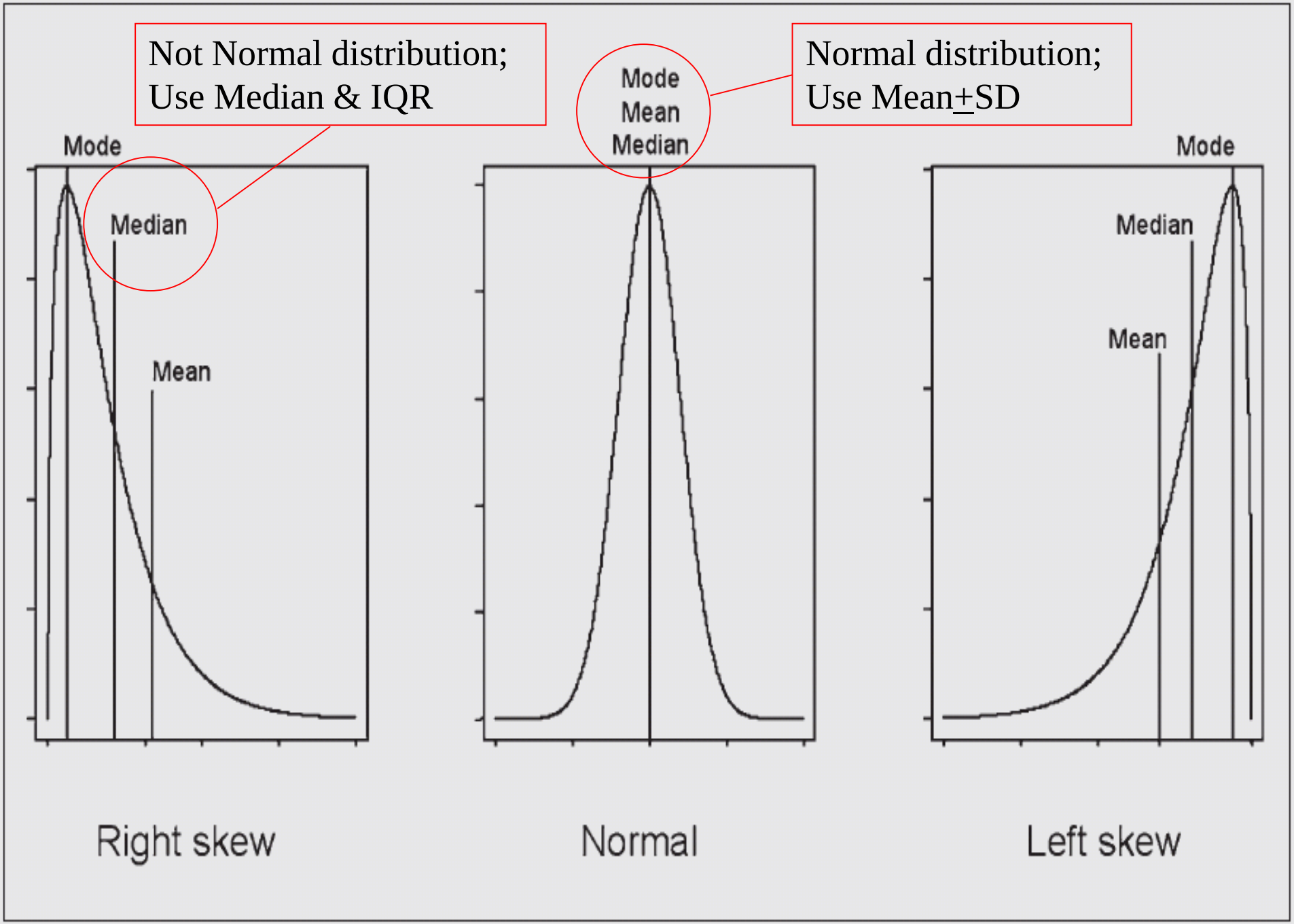


Fig. 2 Distributions of Quantitative Data.