Descriptive Measures

Professor Dr Abubakir M. Saleh

Biostatistics NUR 304

Fall semester

4th week

Outline

- **Measure of central tendency**

➢ Mean , Median, Mode

- **Measure of dispersion**

➢ Standard deviation, range, Box plot

Objectives

At the end of this lecture, students should be able to :

- calculate mean , standard deviation, mode, range ,median , and quartiles.
- Draw box plot

Summarizing measures for quantitative variables

- There are two main measures for quantitative variables:

1. Measure of central tendency (a value that "places" in the middle of data)

2. Measure of spread (a range that indicates how widely values are spread above and below the middle)

- The letter ***n*** is used for the number of observations or values of the variable.
  - E.g. measuring the hemoglobin of 40 persons, n=40

- The letter ***x*** is used for the values themselves
  - E.g. x= 13.2

The mean

- The most commonly used measure of the central value of a distribution is the arithmetic mean, or the average.

- It is the sum of the observations divided by the number of observations.

- The formula for calculation of the arithmetic mean is

  Mean= $\frac{\sum x}{n}$ = $\overline{\mathrm{X}}$

- The ∑ symbol indicates that the values of x must be added together.

Mean

- **Example**

  Calculate the arithmetic mean for:

- The weight of 8 newborn infants:

  2.75 2.86 3.37 2.76 2.62 3.49 3.05 3.12 Kgs

$$\overline{\mathrm{X}} = \frac{\sum x}{n}$$

2.75+2.86+3.37+2.76+2.62+3.49+3.05+3.12 = 24.02

24.02÷8= 3.0025 Kgs

Mean= 3.0025 = $\overline{\mathrm{X}}$

The standard deviation

- This is the measure of spread used together with the mean.

- It is based on the deviations of the observations from the mean (the difference between each observation and the mean)

X	$x - \bar{x}$	$x - \bar{x}$	
2.75	2.75-3.0025	-0.2525	
2.86	2.86-3.0025	-0.1425	
3.37	3.37-3.0025	0.3675	
2.76	2.76-3.0025	-0.2425	
2.62	2.62-3.0025	-0.3825	
3.49	3.49-3.0025	0.4875	
3.05	3.05-3.0025	0.0475	
3.12	3.12-3.0025	0.1175	
24.02			

The standard deviation

- This is the measure of spread used together with the mean.

- It is based on the deviations of the observations from the mean (the difference between each observation and the mean)

- These deviations are squared and added.

$\bar{x}$

The standard deviation

- **Example**

  Calculation of standard deviation (SD) of the distribution of body weight of infants

X	$x - \bar{x}$	$x - \bar{x}$	$(x - \bar{x})^2$
2.75	2.75−3.0025	-0.2525	0.0638
2.86	2.86−3.0025	-0.1425	0.0203
3.37	3.37−3.0025	0.3675	0.1351
2.76	2.76−3.0025	-0.2425	0.0588
2.62	2.62−3.0025	-0.3825	0.1463
3.49	3.49−3.0025	0.4875	0.2377
3.05	3.05−3.0025	0.0475	0.0023
3.12	3.12−3.0025	0.1175	0.0138
24.02		0	**0.6781**

The standard deviation

- This is the measure of spread used together with the mean.
- It is based on the deviations of the observations from the mean (the difference between each observation and the mean)
- These deviations are squared and added.
- The result is divided by (n-1). The result of this is called the **variance**.
- The standard deviation is the square root of the variance.

  Standard deviation = $\sqrt{\text{variance}}$

The standard deviation

- Mean $= \overline{x} = \frac{\sum x}{n} = 3.0025$
- Variance $= \frac{\sum (x - \overline{x})^2}{n - 1} = \frac{0.6781}{8-1} = 0.0968$
- Standard deviation $=$ SD $= \sqrt{\text{variance}} = \sqrt{0.0968} = 0.3112$

- The abbreviation SD is often used for the standard deviation.

The standard deviation

- We do not usually calculate SDs by hand. It is described to give a feel of what the SD is.

- Both the SD and the mean can be obtained on calculator.

- SD and mean are also given by many computer programs.

Example of last week
40 Hb measurements

7.2
13.7
14.2
13.5
8.3
12.2
13.9
11.3
12.3
11.9

14.6
11.7
12.9
11.7
12.1
12.5
9.4
10.2
14.9
14.3

10.5
10.6
11.5
15.2
11.2
11.4
12.6
11.4
12.7
13.1

13.6
10.9
13.4
12.1
10.2
14.5
8.7
9.5
12.5
13.2

- Mean =12.04

x	**$X-\overline{X}$**	**$X-\overline{X}$**	**$(x-\overline{X})^2$**
7.2	7.2 – 12.04	-4.84	23.4
8.3	8.3 – 12.04	-3.74	14.0
8.7	8.7 – 12.04	-3.34	11.2
9.4	9.4 – 12.04	-2.64	7.0
9.5	9.5 – 12.04	-2.54	6.5
10.2	10.2 – 12.04	-1.84	3.4
10.2	10.2 – 12.04	-1.84	3.4
10.5	10.5 – 12.04	-1.54	2.4
10.6	10.6 – 12.04	-1.44	2.1
10.9	10.9 – 12.04	-1.14	1.3
11.2	11.2 – 12.04	-0.84	0.7
11.3	11.3 – 12.04	-0.74	0.5
11.4	11.4 – 12.04	-0.64	0.4
11.4	11.4 – 12.04	-0.64	0.4
11.5	11.5 – 12.04	-0.54	0.3
11.7	11.7 – 12.04	-0.34	0.1
11.7	11.7 – 12.04	-0.34	0.1
11.9	11.9 – 12.04	-0.14	0.0
12.1	12.1 – 12.04	0.06	0.0
12.1	12.1 – 12.04	0.06	0.0
12.2	12.2 – 12.04	0.16	0.0
12.3	12.3 – 12.04	0.26	0.1
12.5	12.5 – 12.04	0.46	0.2
12.5	12.5 – 12.04	0.46	0.2
12.6	12.6 – 12.04	0.56	0.3
12.7	12.7 – 12.04	0.66	0.4
12.9	12.9 – 12.04	0.86	0.7
13.1	13.1 – 12.04	1.06	1.1
13.2	13.2 – 12.04	1.16	1.3
13.4	13.4 – 12.04	1.36	1.8
13.5	13.5 – 12.04	1.46	2.1
13.6	13.6 – 12.04	1.56	2.4
13.7	13.7 – 12.04	1.66	2.8
13.9	13.9 – 12.04	1.86	3.5
14.2	14.2 – 12.04	2.16	4.7
14.3	14.3 – 12.04	2.26	5.1
14.5	14.5 – 12.04	2.46	6.1
14.6	14.6 – 12.04	2.56	6.6
14.9	14.9 – 12.04	2.86	8.2
15.2	15.2 – 12.04	3.16	10.0
481.6			**134.8**

- Mean = $\frac{\sum x}{n}$ =12.04
- Variance = $\frac{\sum (\mathrm{x} - \overline{\mathrm{x}})^2}{\mathrm{n} - 1}$ = $\frac{134.8}{40\text{-}1}$ = 3.46
- Standard deviation = SD = $\sqrt{\text{variance}}$ = $\sqrt{3.46}$ = 1.86

Mean & SD

- Provided the distribution is roughly symmetrical the mean is a good measure of the central value, and the standard deviation is a measure of spread.

- One standard deviation either side of the mean includes roughly 70% of the distribution and two standard deviations includes roughly 95%.

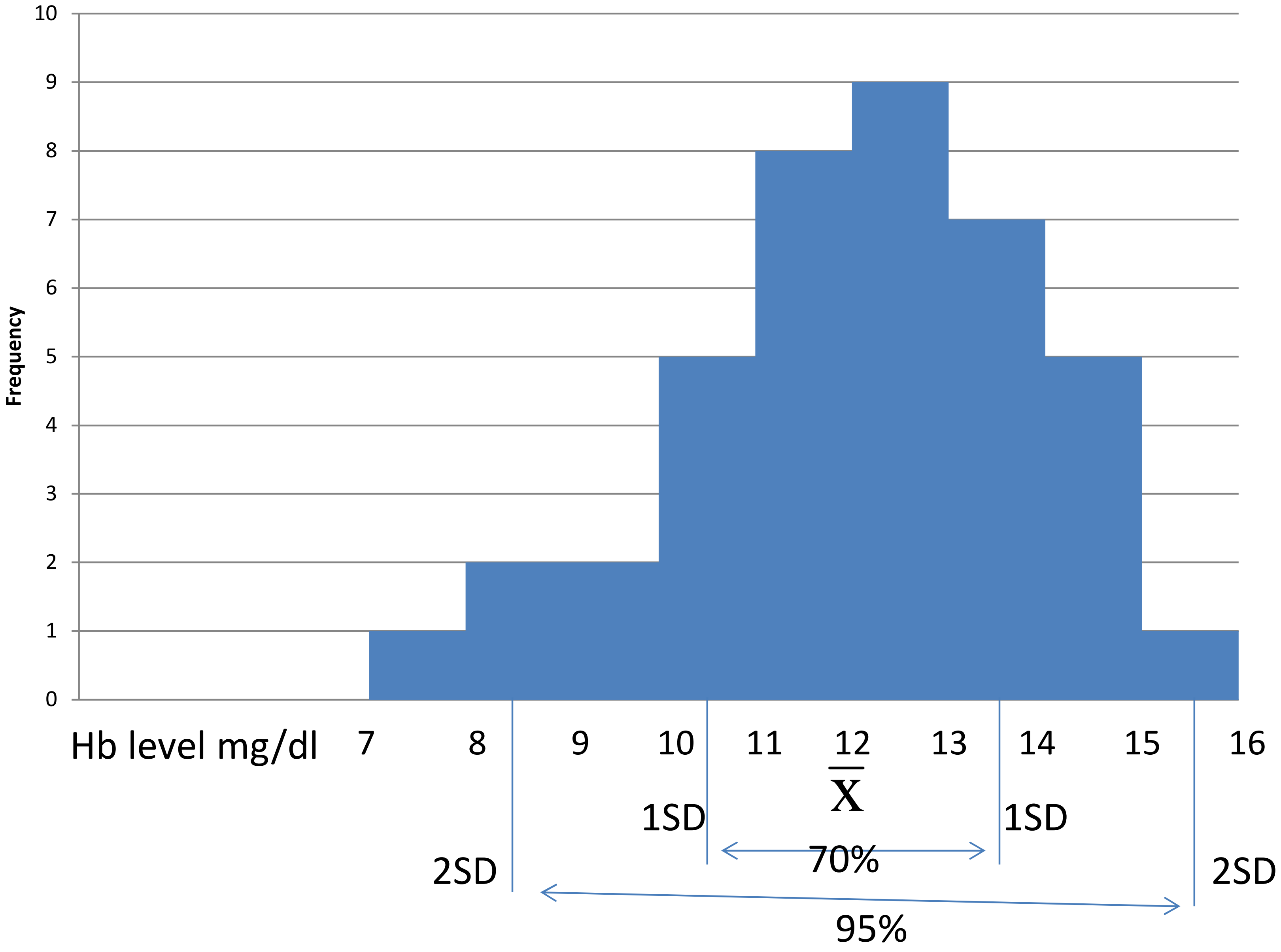
- About two-thirds (70%) of the data values will fall within one standard deviation of the mean value.

- About 95% of the data values will fall within two standard deviations of the mean value.

Mean=12.04 **SD=1.86**

1 SD = 12.04+1.86=**13.90**
12.04-1.86=**10.18**

2SD = 12.04+2(1.86)=**15.76**
12.04-2(1.86)=**8.33**



70% of 40=28 observations

10.19 to 13.89

1SD

7.2	12.2
8.3	12.3
8.7	12.5
9.4	12.5
9.5	12.6
10.2	12.7
10.2	12.9
10.5	13.1
10.6	13.2
10.9	13.4
11.2	13.5
11.3	13.6
11.4	13.7
11.4	13.9
11.5	14.2
11.7	14.3
11.7	14.5
11.9	14.6
12.1	14.7
12.1	15.2

95 of 40=38 observations

8.34 to 15.74

2SD

7.2	12.2
8.3	12.3
8.7	12.5
9.4	12.5
9.5	12.6
10.2	12.7
10.2	12.9
10.5	13.1
10.6	13.2
10.9	13.4
11.2	13.5
11.3	13.6
11.4	13.7
11.4	13.9
11.5	14.2
11.7	14.3
11.7	14.5
11.9	14.6
12.1	14.7
12.1	15.2

Non Symmetric Distributions

- The mean and SD may be satisfactory for reasonably symmetric distributions, but are less so for distributions that are clearly not symmetric.

Example

The number of days spent in hospital by 17 persons after an operation, arranged in increasing size, were

3 4 4 6 8 8 8 10 10 12 14 14 17 25 27 37 42

- The distribution is asymmetric because low values are closer together and often repeated, compared with the high values.

- The mean is 14.6 days. This is clearly not in the centre of the distribution (12 of 17).

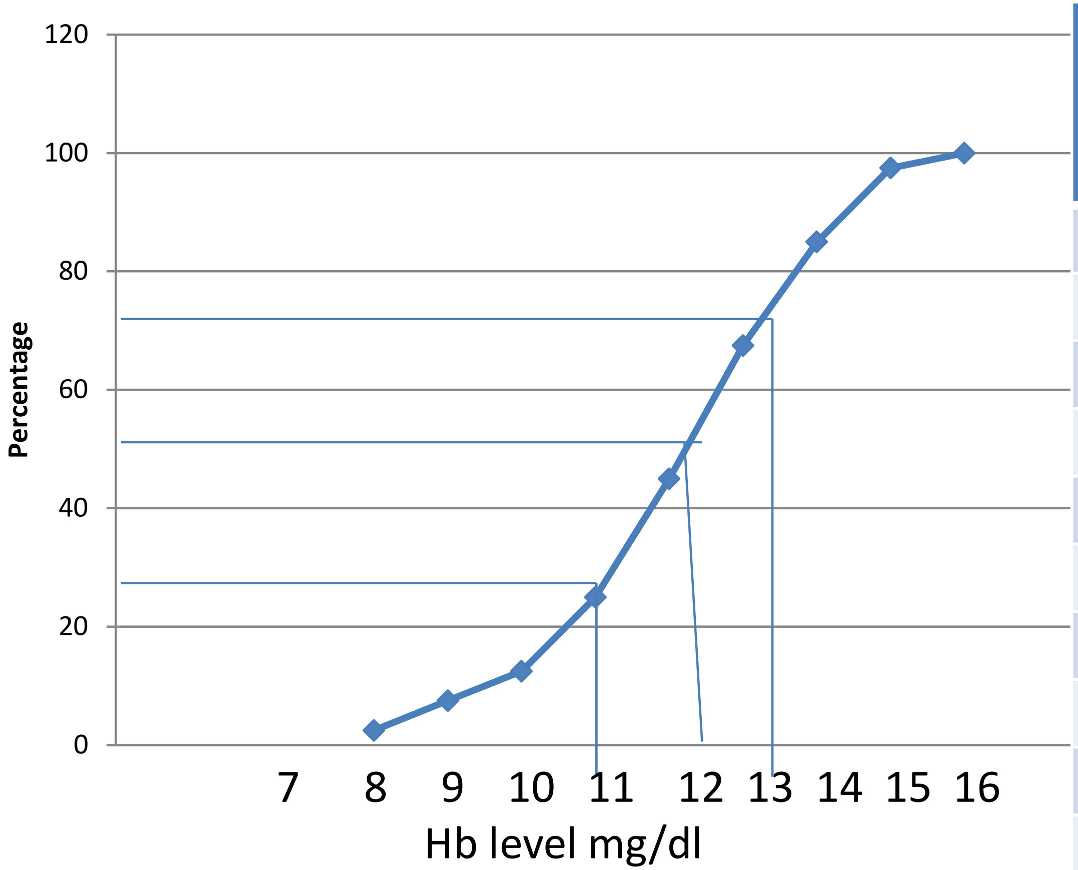
3
4
4
6
8
8
8
10
10
12
14
14
17
25
27
37
42

The median and quartiles of a distribution

- The median is an alternative measure of central value that works better for such non-symmetrical distribution.
- It is the value which halves the distribution, with 50% of the observations below it and 50% above.
- The three values which divide the distribution into quarters are called the quartiles (25%, 50% & 75%).
- The middle quartile (50%) is the median
- The distance between the lower quartile (25%) and the upper quartile (75%) , called the inter-quartile range, is used as a measure of spread.

The median and quartiles

- For a distribution with a large numbers of observations the quartiles are most easily found from the cumulative relative frequency distribution, by reading off the values that correspond to 25%, 50%, and 75%.



Hb measurement class	Frequency	Relative frequency	Cumulative frequency	Relative cumulative frequency
7 to 7.9	1	2.5	1	2.5
8 to 8.9	2	5.0	3	7.5
9 to 9.9	2	5.0	5	12.5
10 to 10.9	5	12.5	10	25.0
11 to 11.9	8	20.0	18	45.0
12 to 12.9	9	22.5	27	67.5
13 to 13.9	7	17.5	34	85.0
14 to 14.9	5	12.5	39	97.5
15 to 15.9	1	2.5.0	40	100
Total	40	100	40	100

The median and quartiles

- For a smaller number of observations the median can be found directly by:
  - Arranging the observations in order from the lowest to the highest value
  - Striking off values at both ends until only one or two remain.

~~7.2~~
8.3
8.7
9.4
9.5
10.2
10.4
10.5
11.5
12.4
~~13.6~~

~~7.2~~
8.3
8.7
9.4
9.5
10.2
10.4
10.5
11.5
12.4
13.6
~~14.2~~

The median and quartiles

- For a smaller number of observations the median can be found directly by:
  - Arranging the observations in order from the lowest to the highest value
  - Striking off values at both ends until only one or two remain.

~~7.2~~
~~8.3~~
8.7
9.4
9.5
10.2
10.4
10.5
11.5
~~12.4~~
~~13.6~~

~~7.2~~
~~8.3~~
8.7
9.4
9.5
10.2
10.4
10.5
11.5
12.4
~~13.6~~
~~14.2~~

The median and quartiles

- For a smaller number of observations the median can be found directly by:
  - Arranging the observations in order from the lowest to the highest value
  - Striking off values at both ends until only one or two remain.
  - If one, this value is the median
  - If two the median is half way between them.

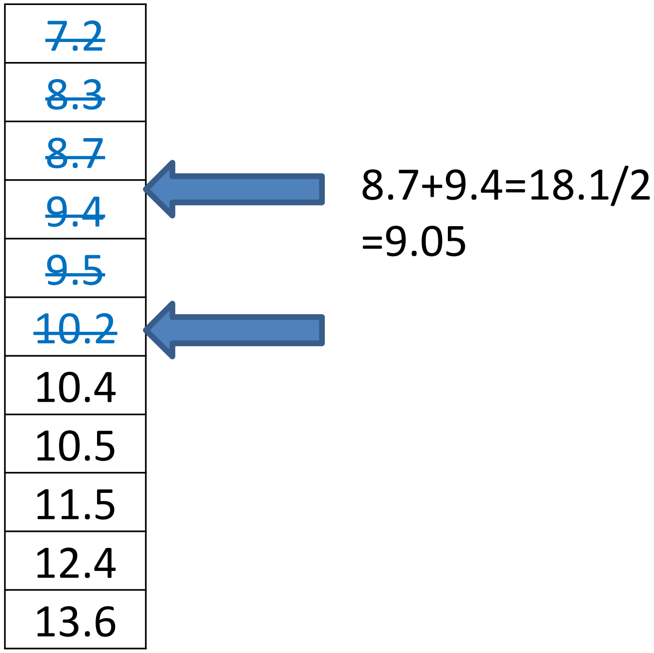
~~7.2~~
~~8.3~~
~~8.7~~
~~9.4~~
~~9.5~~
10.2
~~10.4~~
~~10.5~~
~~11.5~~
~~12.4~~
~~13.6~~

~~7.2~~
~~8.3~~
~~8.7~~
~~9.4~~
~~9.5~~
10.2
10.4
~~10.5~~
~~11.5~~
~~12.4~~
~~13.6~~
~~14.2~~

10.3

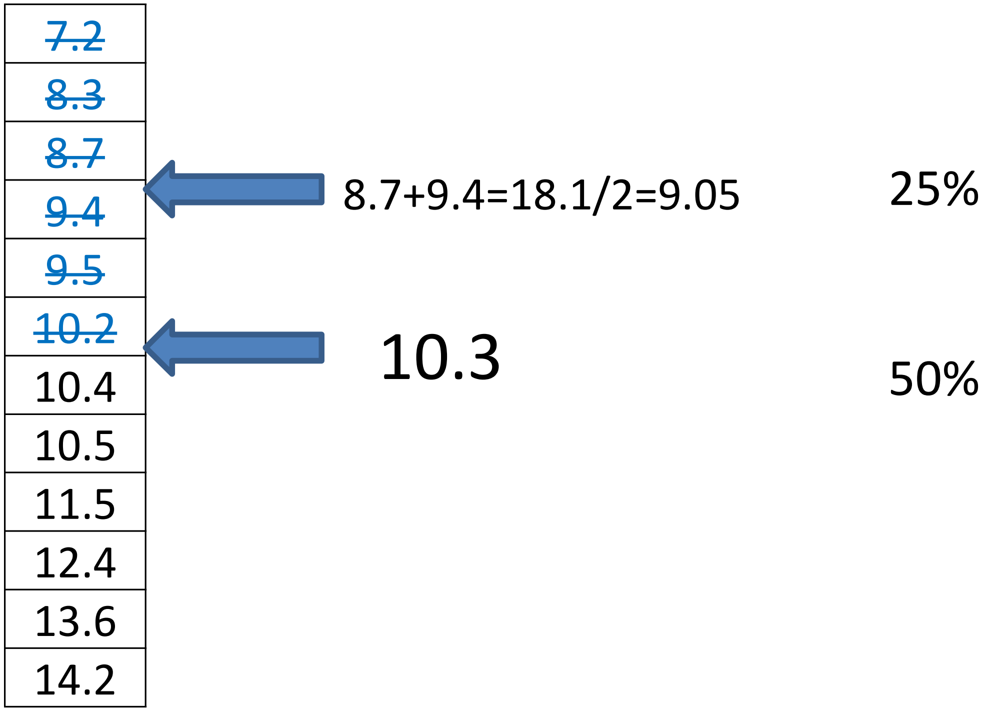
Median and quartiles

- The median is then used to divide the data into two halves
- The medians of each of the halves found in the same way - these are the upper and lower quartiles.
- If the median is the single central value, include it in each half.



~~7.2~~
~~8.3~~
~~8.7~~
~~9.4~~
~~9.5~~
~~10.2~~
10.4
10.5
11.5
12.4
13.6

8.7+9.4=18.1/2
=9.05



~~7.2~~
~~8.3~~
~~8.7~~
~~9.4~~
~~9.5~~
~~10.2~~
10.4
10.5
11.5
12.4
13.6
14.2

8.7+9.4=18.1/2=9.05

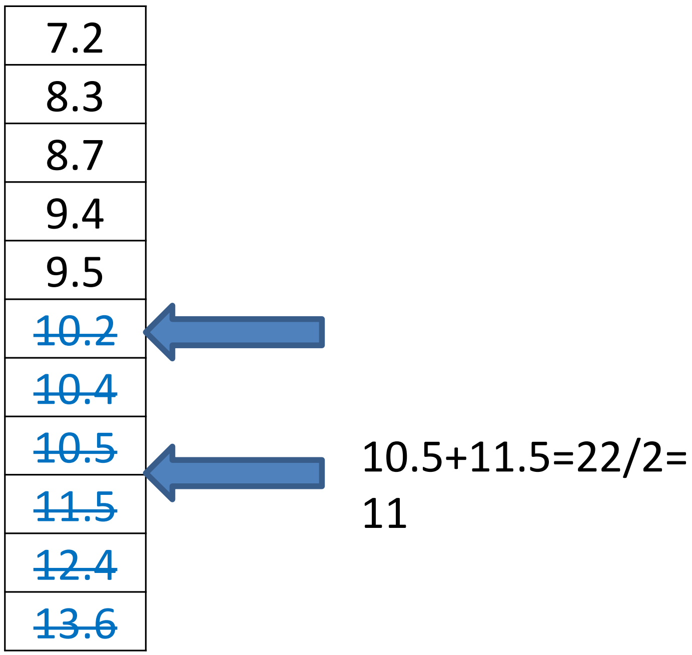
25%

10.3

50%

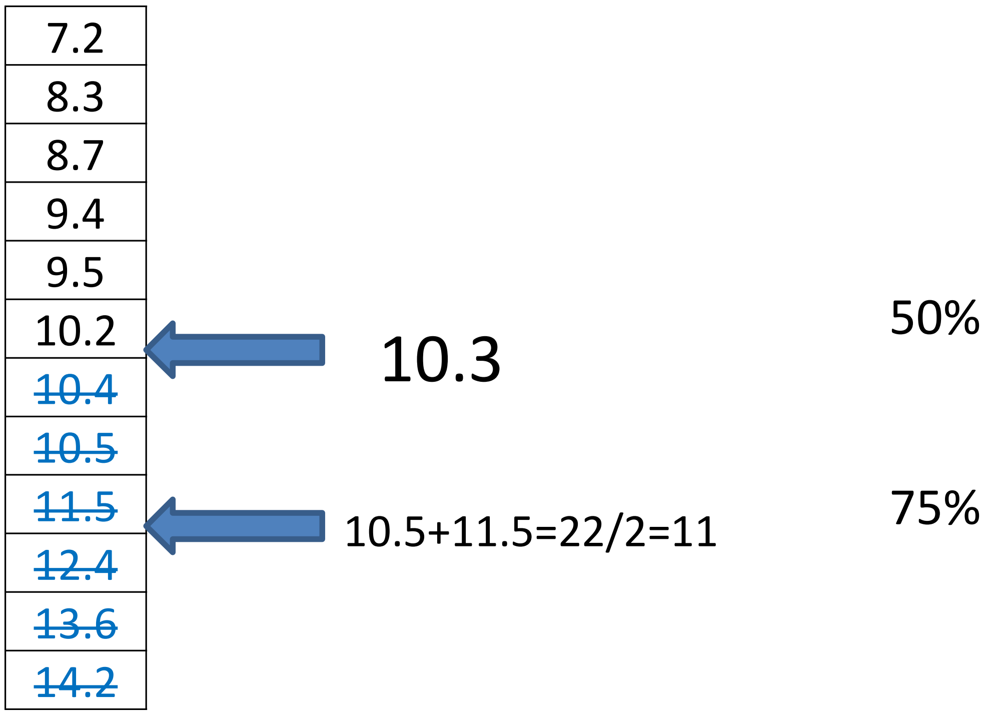
Median and quartiles

- The median is then used to divide the data into two halves
- The medians of each of the halves found in the same way - these are the upper and lower quartiles.
- If the median is the single central value, include it in each half.



7.2
8.3
8.7
9.4
9.5
~~10.2~~
~~10.4~~
~~10.5~~
~~11.5~~
~~12.4~~
~~13.6~~

10.5+11.5=22/2= 11



7.2
8.3
8.7
9.4
9.5
10.2
~~10.4~~
~~10.5~~
~~11.5~~
~~12.4~~
~~13.6~~
~~14.2~~

10.3

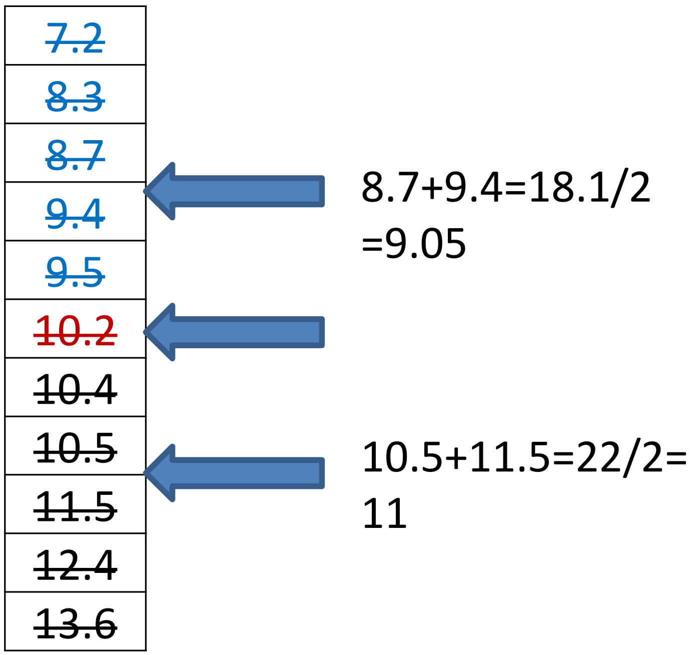
10.5+11.5=22/2=11

50%

75%

Median and quartiles

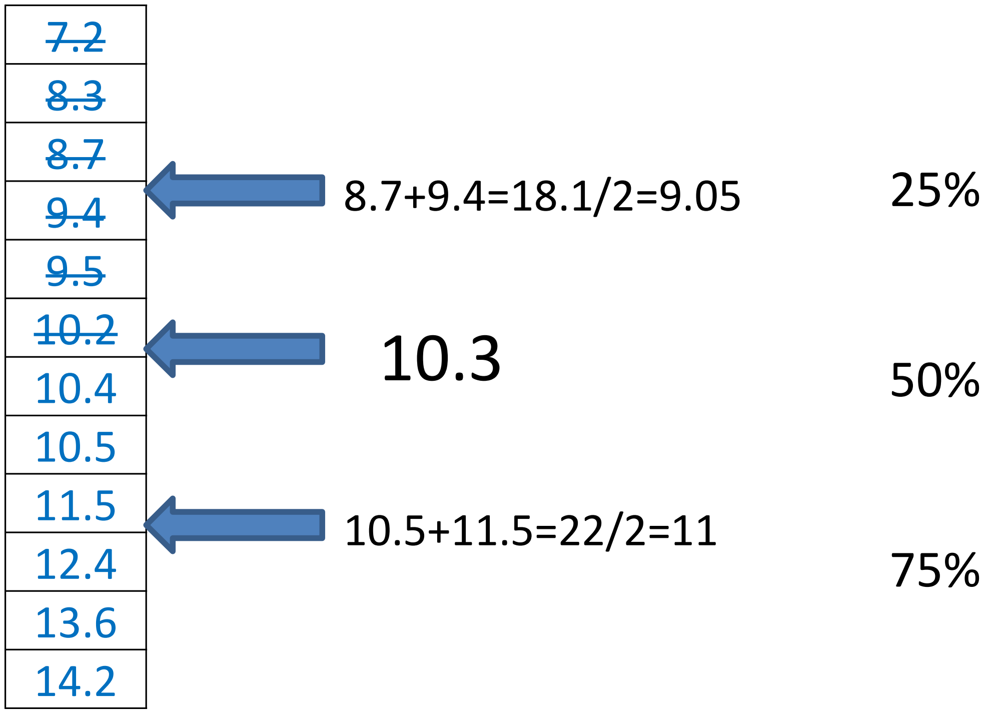
- The median is then used to divide the data into two halves
- The medians of each of the halves found in the same way - these are the upper and lower quartiles.
- If the median is the single central value, include it in each half.



~~7.2~~
~~8.3~~
~~8.7~~
~~9.4~~
~~9.5~~
~~10.2~~
~~10.4~~
~~10.5~~
~~11.5~~
~~12.4~~
~~13.6~~

8.7+9.4=18.1/2 =9.05

10.5+11.5=22/2= 11



~~7.2~~
~~8.3~~
~~8.7~~
~~9.4~~
~~9.5~~
~~10.2~~
10.4
10.5
11.5
12.4
13.6
14.2

8.7+9.4=18.1/2=9.05

25%

10.3

50%

10.5+11.5=22/2=11

75%

Mode

- It is the value that is observed most frequently
- It is commonly used for large number of observations
- With small number of observations, there may be no mode.
- If there are 2 modes, it is called bimodal.
- Mode=8

3
4
4
6
8
8
8
10
10
12
14
14
17
25
27
37
42

Range

- Range is the difference between the lowest observed value and the highest.

  Minimum and maximum

- The sample may have a large range even when the majority of observations are fairly close.

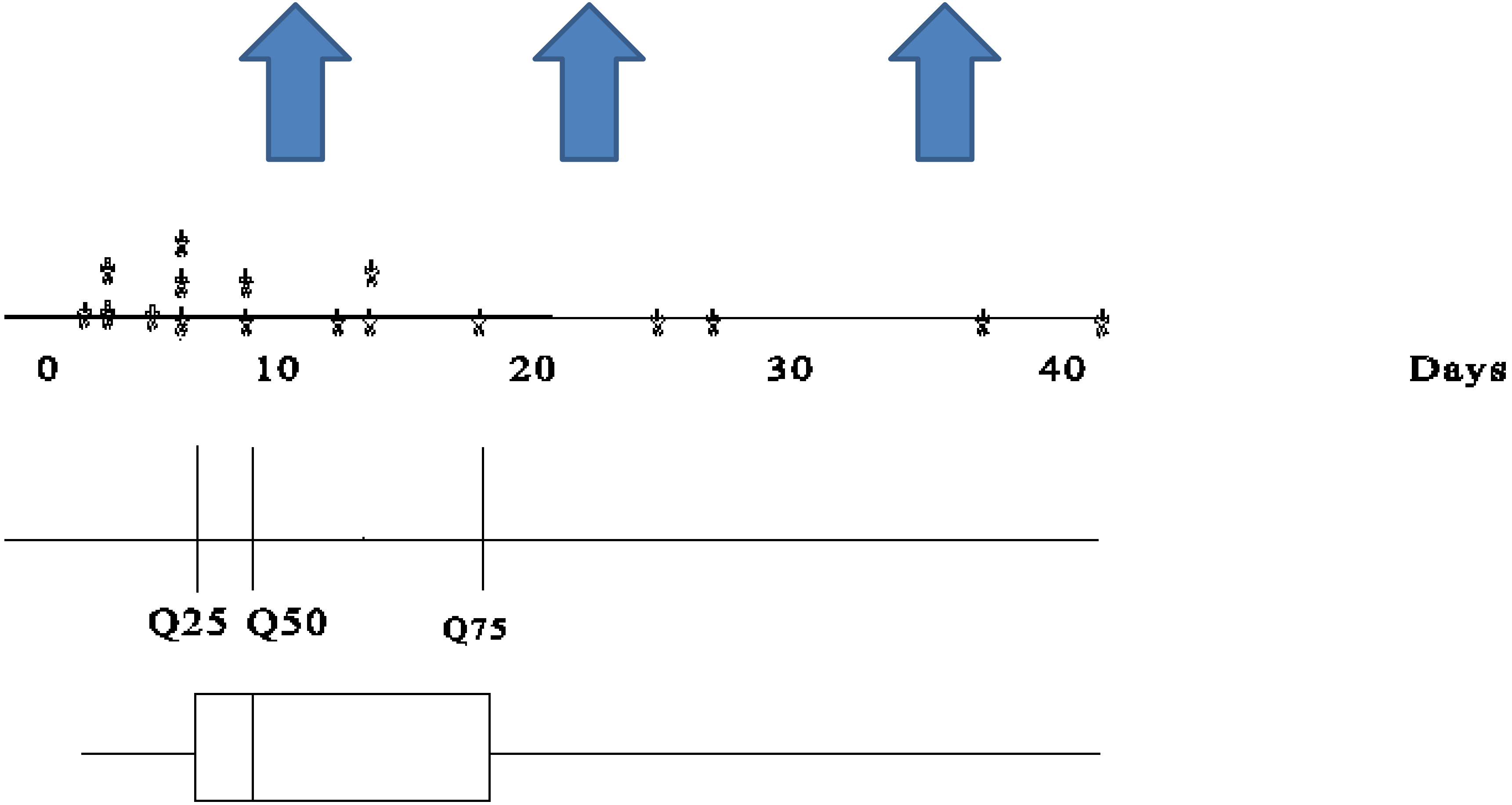
- Range (3 to 42)

3
4
4
6
8
8
8
10
10
12
14
14
17
25
27
37
42

Box plot

- The following 5 numbers give a general purpose summary of a distribution for both non-symmetric and symmetric distributions:
- The smallest value
- The lower quartile, Q25
- The median (Q50)
- The upper quartile, Q75
- The largest value
- These numbers are often shown in a figure called a **box plot**.

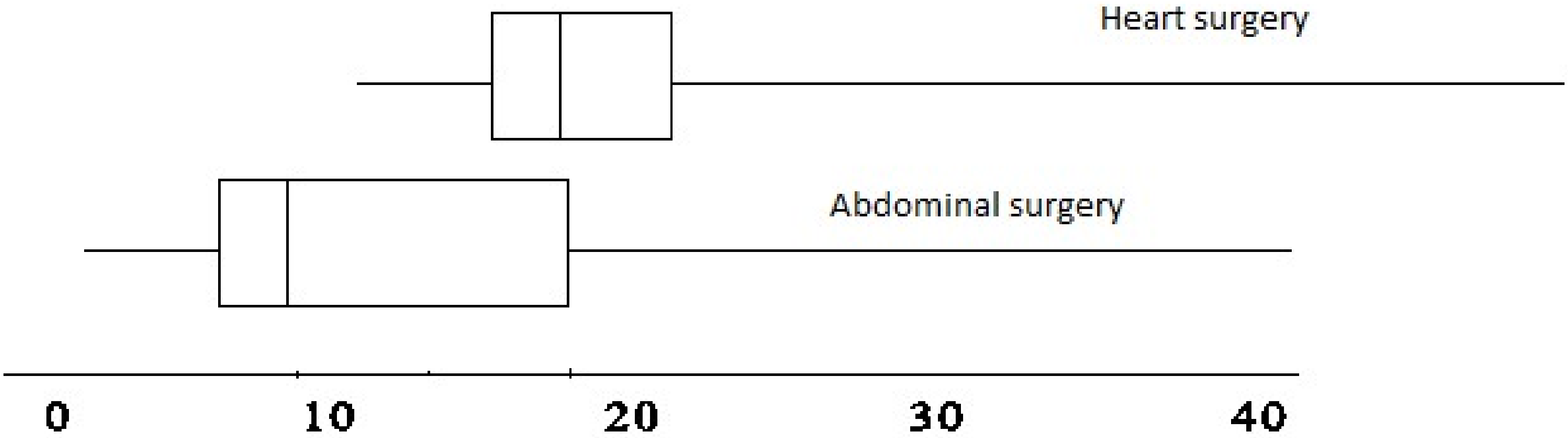
3 4 4 6 8 8 8 10 10 12 14 14 17 25 27 37 42



- The box includes 50% of the distribution (from 25% to 75%)
- The line within the box represents the median
- The horizontal lines join the smallest and largest values to the box.

Box Plot

- Box plots are useful for presenting several distributions in one figure and enable them to be compared easily.



Centiles

- The quartiles are the values which correspond to the cumulative percentages 25, 50 and 75, but there is no need to stick to these percentages.

- Sometimes we need to report the values corresponding to the percentages, e.g. 5% 10%, 25%.

- These are known as the 5th, 10th, 25th percentiles of the distribution.

General rules

- Always report the **number of observations** on which the summary is based, e.g. n=40

- If the central value of a quantitative distribution is measured using the **median**, give the lower and upper **quartiles** as well.

- If the central value of a quantitative distribution is measured using the **mean** give the **standard deviation** as well.

References

- Essential Medical Statistics, by Betty Kirkwood & Jonathan Sterne
(Published by Blackwell)

Statistics Without Tears, a Primer for Non-mathematicians, by Derek Rowntree
(Published by Penguin)

Questions?