
STATISTICS
震旦
I ♥
SH

We should now be able to calculate the 3 measures of central location or averages discussed in the previous unit.

These are important but not always sufficient.

As well as an average, it is important to know how the data values are grouped around the average – whether the values are clustered closely or scattered more widely.

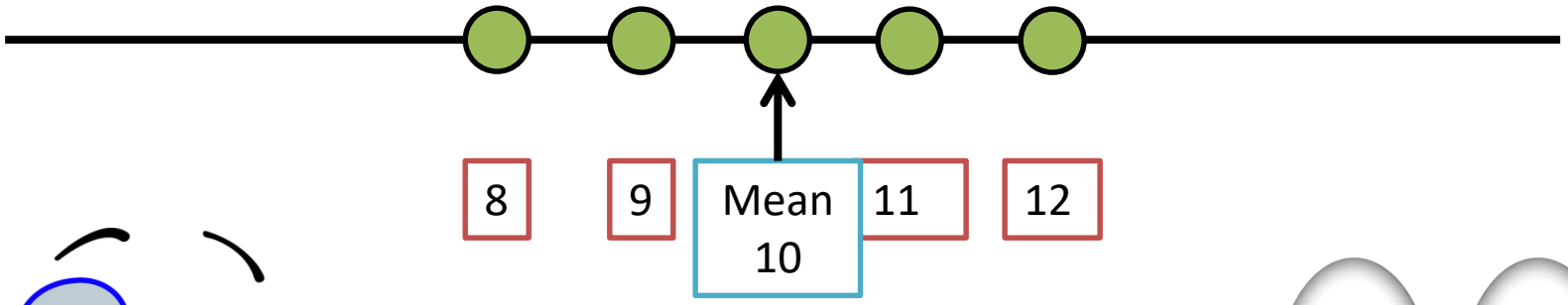
To illustrate this…consider the 2 sets of data:

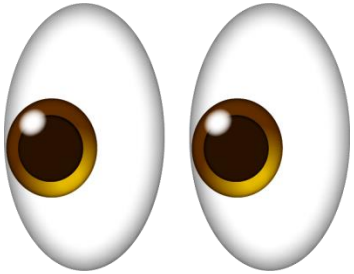
A: 8…9…10…11…12…

B: 4…7…10…13…16…

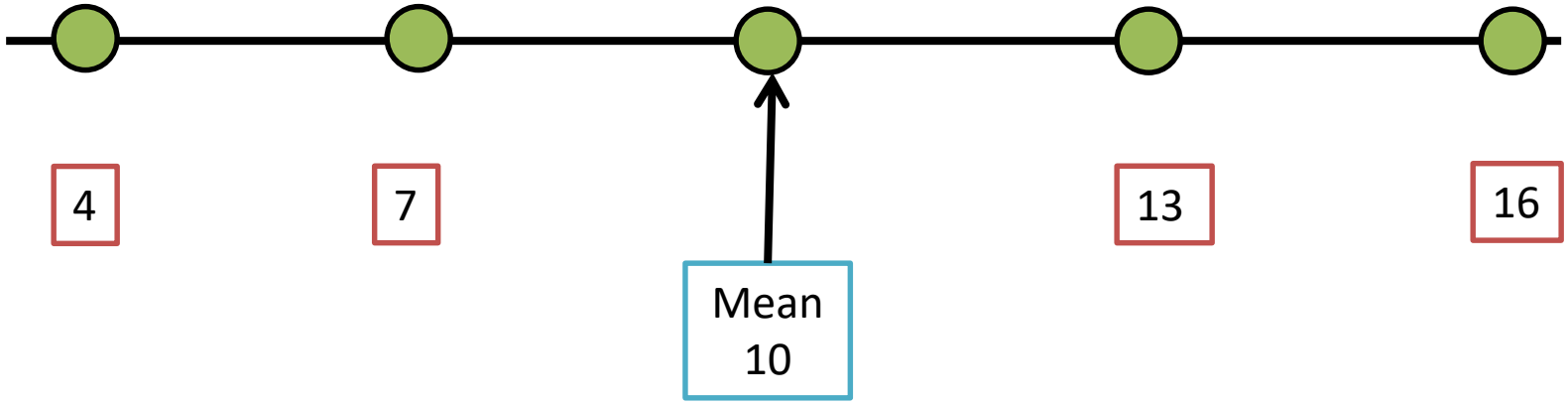
Although the mean of each sample is 10…we can see that the values in B are scattered more widely about the mean than those in A

Sample A



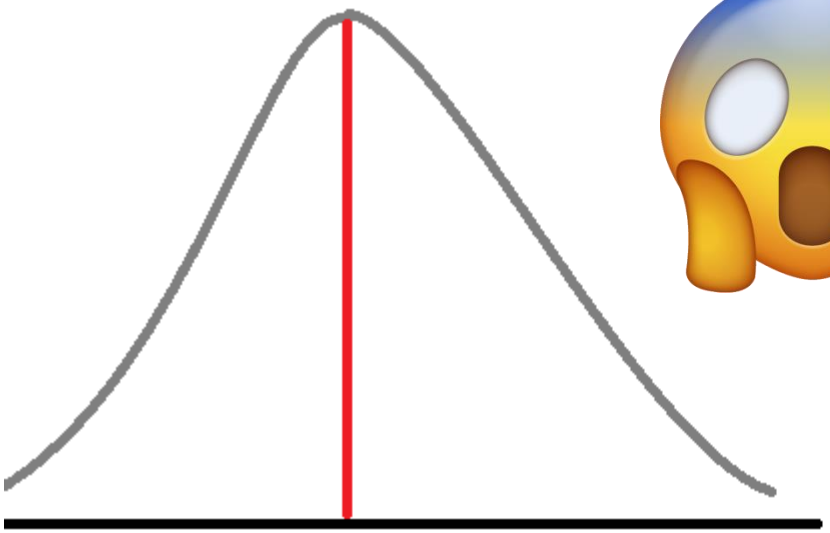


Sample B

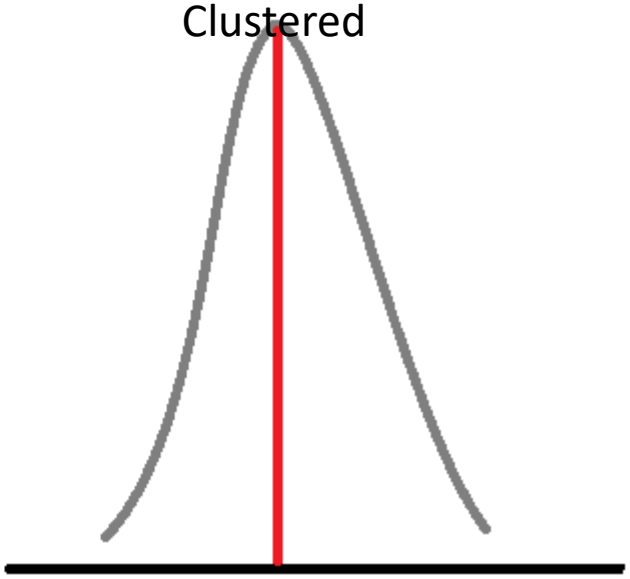


As a further illustration…see the graphs below. The first figure shows that the data is more widely spread about their central value where as the second figure shows the data values tend to cluster together very close to their central value.

Sample B
Scattered



Sample A
Clustered



Both of these diagrams demonstrate that a second numerical measure may be needed in order to describe the variability of the individual values.

Several different measures are used to indicate the extent of the spread, dispersion or variability of a set of data, and in this unit we will be considering 2 types of measures of variability:

1. **Distance measures**
2. **Measures of average deviation**


DISTANCE MEASURES

The two indicators in this section are the **RANGE**
(Highest value – Lowest value)

and the

INTER QUARTILE RANGE (75% - 25%).

The range:

The simplest distance measure of dispersion is the range.

The range is the difference between the largest and the smaller values in a set of data.

$$\boldsymbol{Range = Highest\ value\ - Lowest\ value}$$

$$\boldsymbol{R = x_H - x_L}$$

The range is simple to calculate and interpret. However it is based solely of the 2 extreme values in a set of data and ignores all the other values. It can therefore be greatly distorted by very high or very low values.

The temperature in degrees Celsius on 10 consecutive days in Brisbane are:

27…26…28…25…29…29…30…31…28…27…

Find the range.

Re-arrange the temperatures is ascending order:

25…26…27…27…28…28…29…29..30…31…

$$\boldsymbol{R = x_H - x_L}$$
$$\boldsymbol{= 31 - 25}$$
$$\boldsymbol{6^oC}$$

Interquartile range:

The Interquartile range is a more meaningful measure of range as it ignores extreme values by finding the interval that contains the middle 50% of the data values.

The interquartile range is defined using quartiles by the difference.

$$\boldsymbol{IQR = upper\ quartile\ - lower\ quartile}$$

$$\boldsymbol{IQR = Q_3 - Q_1}$$

The highest and lowest 25% of the data have been discarded

The semi-interquartile range

This semi interquartile range is half of the interquartile range and measure the average distance from the center of the set of data values.

$$\frac{1}{2}(Q_3 - Q_1)$$

1. Find the range of:

a) 12…15…10…19…16…

b) 2…5…1…1…4…8…8…7…7…6…9…4…

c) 89…78…61…90…34…70…99…38…

d) 2.02…1.07…1.08…1.05…1.01…1.11…1.08…

2. Find the range and the IQR of:

a) 8…9…7…10…5…6…4…

b) 10…10…3…6…9…9…5…4…3…6…5…14…

3. The weekly wages in dollars paid to employees at a certain company are:

400	425	390	450	370	415	380	415	420	455

Find the range of wages


MEASURES OF DEVIATION

One measure of dispersion is called the variance as it can be used to measure the variability of the data values about the mean.

Find the mean and variance of the values:

2…3…5…6…8…

$$Sx = 2 + 3 + 5 + 6 + 8$$
$$= 24$$

$$Sx^2 = 2^2 + 3^2 + 5^2 + 6^2 + 8^2$$
$$= 4 + 9 + 25 + 36 + 64$$
$$= 138$$

The mean of the sample of n = 5 values is 4.8

The variance of the sample is…

$$S^2 = \frac{1}{n-1}\left[Sx^2 - \frac{(Sx)^2}{n}\right]$$

$$= \frac{1}{5-1}\left[138 - \frac{24^2}{5}\right]$$

$$= \frac{1}{4} \times 22.8$$

$$= 5.7$$

Sample standard deviation

The variance is closely related to another measure variability…the STANDARD DEVIATION.

Just as for the mean…we often want to measure variability using the same units as the original data (not the unit squared)

Such a measure can be obtained by taking the positive square root of the variance.

The square root of the variance is called the standard deviation….denoted by SD.

Since the variance is measured in square units…the SD is measured in the same units as the original data.

The SD is the most important and most frequently used measure of variability.

Looking at our previous example…the variance is 5.7.

The standard deviation is therefore the square root of this…

$$\boldsymbol{SD = \sqrt{5.7}}$$
$$\boldsymbol{= 2.387\ldots}$$

The population variance is denoted by σ^2

$$\boldsymbol{\sigma^2 = \frac{S(X-\mu)^2}{N}}$$

Where N = population size

The population SD denoted by σ is the square root of the population variance.

1. Use the 2 difference formulae for s^2 to calculate the variance of the sample 2…3…4…5…6…

2. Calculate the variance and SD of:

a) 2…5…7…12…9…5…3…4…1…11…8…

b) 12…24…18…29…51…10…32…

3. Calculate the variance of the population 5…8…10…12…15…

4. The travel claim in dollars made on a certain day by 6 employees is:

137…152…105…145…135…129…

Calculate the mean and SD

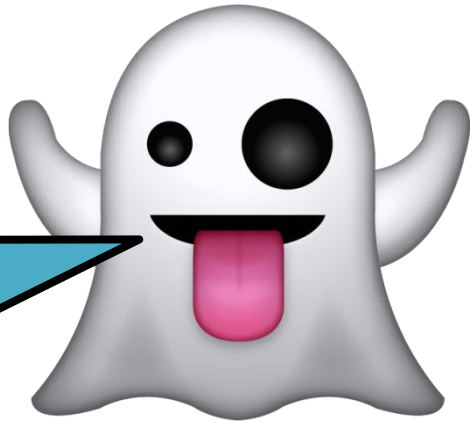
Some applications of variance and SD

We can use our knowledge about dispersion to control the future variability of data values.

For example…the SD deviation can be used in quality control to monitor a particular manufacturing process by measuring the error involved in the output being produced. If there is too much variation, we can identify the causes of the variation and adjust the process to produce goods of higher quality.

Another important application of variability occurs in finance where the variance is used to measure risk (as well as sales…profits…return on investments etc.)

When faced with alternative investments, the underlying premise is that larger variances correspond to higher risks.



Suppose the annual rates of return over the past 5 years for two funds are:

Fund A: **6.4…-1.2…7.1…-3.2…11.9**

Fund B: **8.1…-2.5…4.9…10.2…-0.7**

Which fund has the higher level of risk?

Solution

Use a calculator in statistical mode to find:

Fund A	Fund B
$n_A = 5$	$n_B = 5$
$SX_A = 21$	$SX_B = 20$
$SX_A^2 = 244.66$	$SX_B^2 = 200.4$
$Xbar_A = 4.2\%$	$Xbar_B = 4\%$
$S_A^2 = 39.115(\%^2)$	$S_A^2 = 30.1(\%^2)$

On the basis of these results, we can claim that fund A has a higher level of risk than fund B. In addition, observe that fund A has a higher average rate of return

PRACTICAL INTERPRETATION OF SD

The empirical rule specifies the approximate percentage of data values that lie within 1…2 or 3 standard deviations of the mean.

If a bell shaped distribution has mean μ and variance σ^2…the empirical rules states that:

1. Approximately 68% of the data values are within 1SD of the mean.

$[\mu- \sigma, \mu+\sigma]$

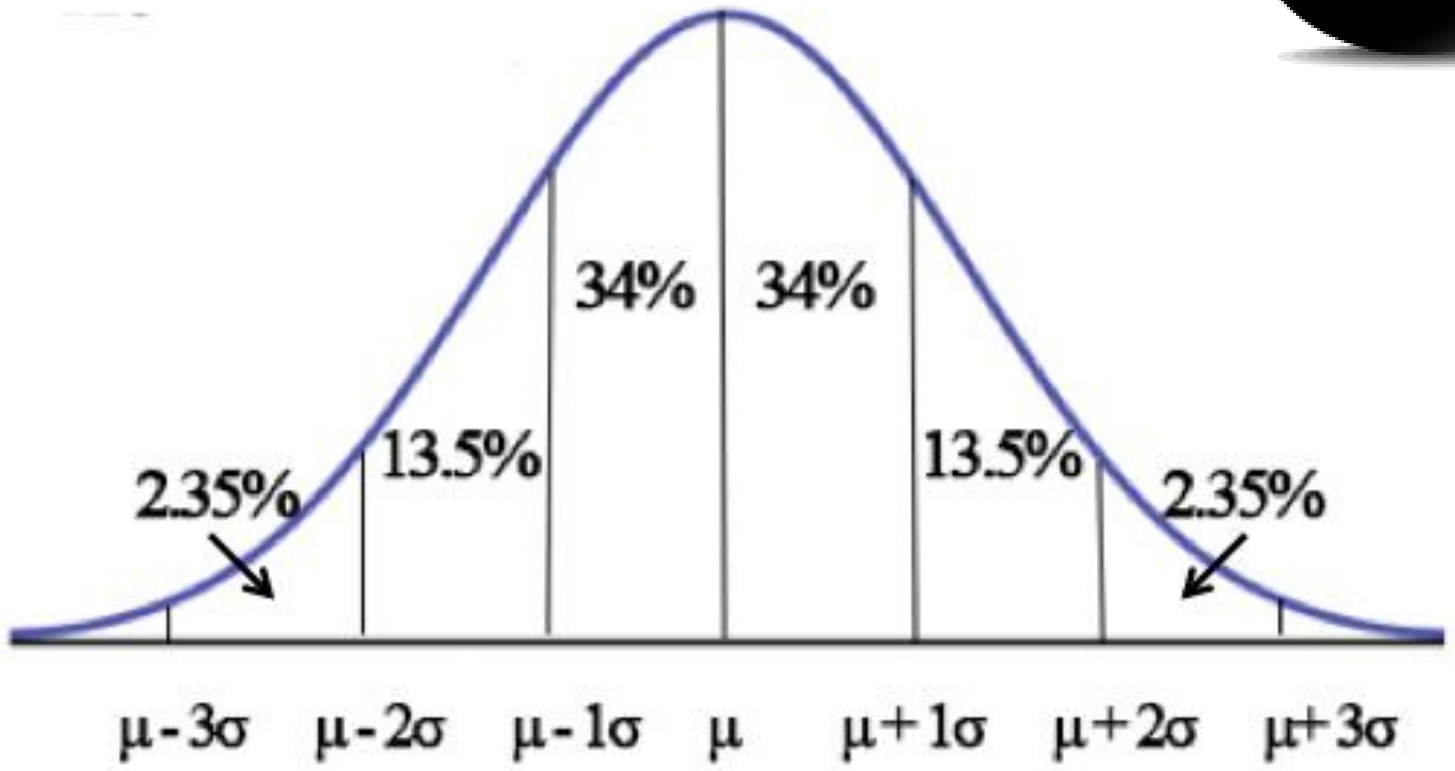
2. Approximately 95% of the data values are within 2 SD's of the mean and thus lie in the interval…

$[\mu- 2\sigma, \mu+2\sigma]$

3. Approximately 99.7% of the data values are within 3SD's of the mean, and thus lie in the interval…

$[\mu- 3\sigma, \mu+3\sigma]$

THIS IS VERY COOL…!!!



The test scores of 1000 students in a math test are normally distributed with a mean of 60 and a standard deviation of 10. Determine the number of students who scored between:

a) 50 and 70
b) 40 and 80
c) 30 and 90

a) We are given $\mu = 60$ and $\sigma = 10$

Since $\mu - \sigma = 60 - 10 = 50$

And $\mu + \sigma = 60 + 10 = 70$

Approximately 68% or 0.68 x 1000 = 680 students will score between 50 and 70

b) $\mu - 2\sigma = 60 - 20 = 40$

$\mu + 2\sigma = 60 + 20 = 80$

Approximately 95% or 0.95 x 1000 = 950 of the students will score between 40 and 80.

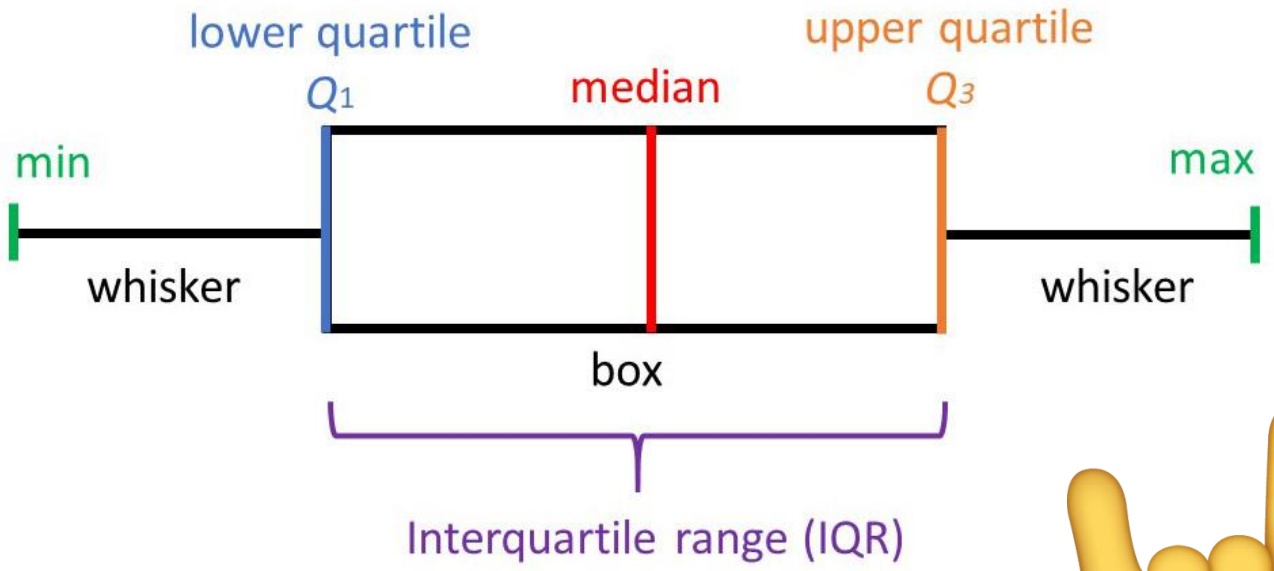
c) $\mu - 3\sigma = 60 - 30 = 30$

$\mu + 3\sigma = 60 + 30 = 90$

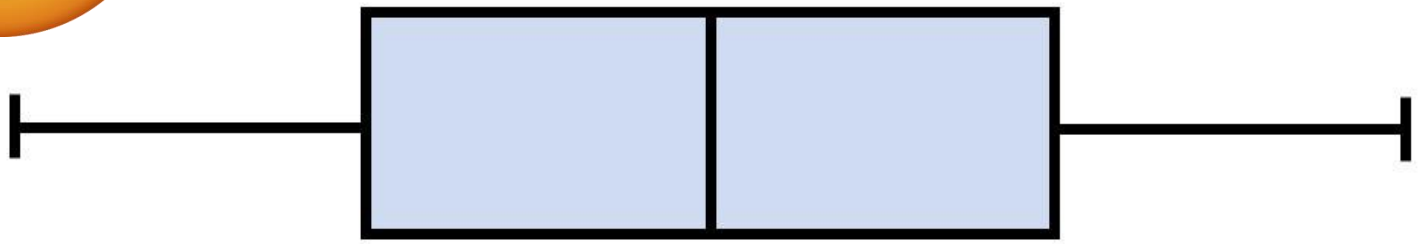
Approximately 99.7% or 0.997 x 1000 = 997 of the students will score between 30 and 90


BOX & WHISKER DIAGRAMS

A box and whisker diagram uses the 5 number summary to draw a simple diagram.

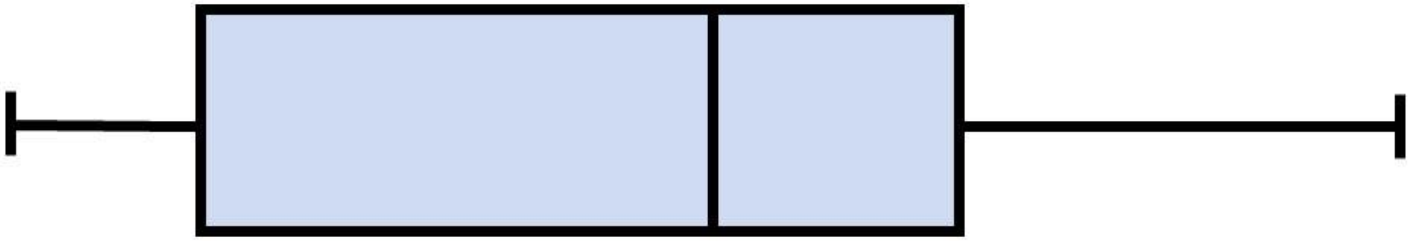


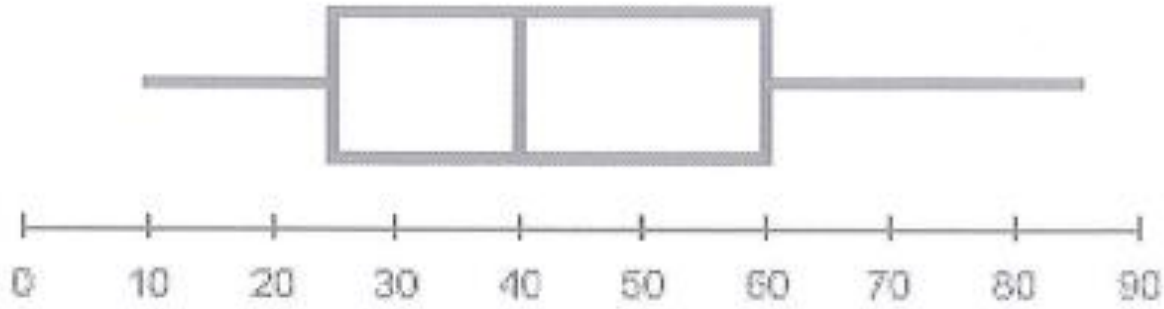
Normal Distribution



Positive Skew

Negative Skew





a) Estimate the median
b) Estimate the 1st and 3rd quartiles
c) Determine the interquartile range
d) Is the distribution symmetrical or skewed to the right/left?

A pizza place offers free delivery of its pizzas within 10km's.

For a sample of 20 deliveries, the owner determines the following information on the time in minutes, it takes the driver to deliver the pizza's:

Minimum value, $X_L = 12$

First quartile, $Q_1 = 14$

Median, $M_d = 18$

Third quartile, $Q_3 = 21$

Maximum value, $X_H = 30$

a) Draw a boxplot

b) How long does it take to deliver the middle 50% of deliveries?

c) Is the distribution of delivery times symmetrical or skewed to the right/left?