Statistics Sector 1: Numerical Measures

Aims

- To be able to calculate measures of average and measures of spread.
- To understand the difference between measure of average and measures of spread.
- To be able to calculate standard deviation and variance.
- To understand when different numerical measures are appropriate.
- To calculate estimates for the mean and standard deviation of grouped data.

Population a collection of all possible items (people, objects) about which we wish to gather information

Sample a portion, or part of the population of interest.

One of the major purposes of statistics is to examine samples of data and make inferences about the population from which the samples are drawn.

A **variable** – characteristic being considered

A **parameter** is a numerical property of a population and a **statistic** is a property of a sample.

Types of Data

Qualitative - when the variable is non numeric

Quantitative - when the variable can be reported numerically

Discrete - can only assume certain values, usually arise from counting something

Continuous - can assume any value within a specific range

Example 1

Packets of a particular type of sweet are known to have a mean of 100 grams. The number of sweets in a packet is approximately 30 and the sweets come in any one of five flavours. The weight of 50 packets are taken and the mean is found to be 98.3 grams.

From the above passage, identify;

a. a population
b. a parameter
c. a sample
d. a qualitative variable
e. a continuous variable
f. a discrete variable

Measures of Average

- The **mode** or modal value is the value that occurs most frequently. There can be more than one mode or it may not exist.
- The **median** is the middle value of **ordered** data, it is the $\left(\frac{n+1}{2}\right)$th term.
- The (arithmetic) **mean** or average is the sum of the values divided by the number of values ($\bar{x} = \frac{\sum x}{n}$). Remember the mean may be a value that cannot occur such as $\bar{x} = 2.43$ children. The sample mean is denoted $\bar{x}$ and the population mean is denoted by μ.

Measures of Spread

- The **range** is the difference between the highest and the lowest values.
- The **interquartile range** (IQR) is the difference between the upper and lower quartiles, $Q_3 - Q_1$.
    - The lower quartile, Q_1, is the median of the ordered values to the left of the median or the $\left(\frac{1}{4}(n+1)\right)$th term.
    - The upper quartile, Q_3, is the median of the ordered values to the right of the median or the $\left(\frac{3}{4}(n+1)\right)$th term.
    - The median is also the second quartile, Q_2.
- The **standard deviation** (where the sample standard deviation is denoted s and the population standard deviation is denoted by σ) and the **variance** (where the sample variance is denoted s^2 and the population variance is denoted by σ^2) are measures of the average deviation of the values from their mean.

Example 2

At a doctors surgery they record the number of patients who are late for appointments each week. The records for the first 12 weeks are recorded below.

14 23 18 37 a 21 16 b 32 28 19 26

Unfortunately on two of the weeks the number of lates was incorrectly recorded, however they do know that $a < 12$ and $b > 40$.

Calculate the median and IQR of the 12 values.

To do this on a graphical calculator you must enter any number less than 12 for a and any number bigger than 40 for b.

Example 3

The following table gives the number of complaints recorded by a telecoms company on each day for a period of 50 days.

Number of Complaints	0	1	2	3	4	5	6	7
Number of Days	1	5	7	14	9	10	3	1

a) Calculate the range, IQR and mode of the data.
b) Calculate the mean and standard deviation of the data.

Standard Deviation and Variance

The **standard deviation** is a measure of the average deviation of the values from their mean. The formula depends on whether the value represents a population or a sample. A sample is a set of values selected from the population. Standard deviation is based on the sum of squares so should never be negative.

The formula for the population standard deviation, σ, is given in the formula booklet:

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

The formula for the sample standard deviation, s, is not given in the formula booklet:

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

Remember:
$Standard\ deviation = \sqrt{variance}$
or
$variance = standard\ deviation^2$

Example 4

Find the variance and standard deviation of the number of children per family from the data below:

Number of Children	0	1	2	3	4	5	6
Number of Families	9	4	6	5	2	0	1

Example 5

The response time, X minutes, is the time it takes between an emergency call being answered and the ambulance arriving at the scene. The value of X was recorded on a random sample of 50 occasions. The results are summarised below, where $\bar{x}$ donates the sample mean.

$$\sum x = 321.4 \qquad \sum (x - \bar{x})^2 = 42.62$$

Find the values for the mean and standard deviation of this sample of 50 response times.

Example 6

The mean blood cholesterol level of the adult living on a small housing estate has been found to be 5.8 millimoles per litre.
Monica is a researcher who believes that daily consumption of yogurt can reduce blood cholesterol level. Each of the 80 residents consumed yogurt daily and she measured the blood cholesterol level, X, of each resident obtaining the following results

$\sum x = 452.8$ and $\sum x^2 = 2596.4$

Find the values for the mean and standard deviation of the residents. Comment on the values obtained.

Estimating the Mean and Standard Deviation

A grouped frequency distribution is a list of continuous classes together with their corresponding frequencies. As we do not know the exact values of all the data we can only estimate the mean and standard deviation using the **midpoints** of each group.

Example 7

A class of 35 pupils all complete a small puzzle as an aptitude test. Each pupil was timed in seconds and the times are recorded below.

Time to complete puzzle (s)	Frequency
$20 \leq x < 40$	6
$40 \leq x < 50$	8
$50 \leq x < 55$	7
$55 \leq x < 60$	5
$60 \leq x < 100$	9

Remember you need to enter the midpoints in the x column.

Calculate estimates for the mean and standard deviation for the time taken to complete the puzzle.

Example 8

Ted is a farmer and during early spring he planted his crop of tomatoes. He takes a random sample of 65 plants and counted the number of tomatoes on each plant. He recorded his results as follows.

Number of Tomatoes	Less than 7	7-11	12	13	14	15	16	17-19	20-26
Frequency	1	4	8	12	16	11	8	3	2

a) Determine values for the median and interquartile range.

b) Given that the smallest number of tomatoes was in fact 4, calculate estimates of the mean and the standard deviation.

Example 9

Katy works as a clerical assistant for a small company. Each morning, she collects the company's post from a secure box in the nearby Royal Mail sorting office.

Katy's supervisor asks her to keep a daily records of the number of letters that she collects (x). Her records for a period of 175 days are summarised in the table

x	0-9	10-19	20	21	22	23	24	25-29	30-34	35-39	40-49	50 or more	**Total**
f	5	16	23	27	31	34	16	10	5	3	4	1	**175**

(a) For these data:

(i) state the modal value; *(1 mark)*

(ii) determine values for the median and the interquartile range. *(3 marks)*

(b) The most letters that Katy collected on any of the 175 days was 54. Calculate estimates of the mean and the standard deviation of the daily number of letters collected by Katy. *(4 marks)*

(c) During the same period, a total of 280 letters was also delivered to the company by private courier firms.

Calculate an estimate of the mean daily number of **all** letters received by the company during the 175 days. *(2 marks)*

Choice of Numerical Measures

Mode

Advantages

- Easy to find
- Can be used with numerical or non-numerical data

Disadvantages

- May not be unique or may not exist. For example if there are two modes such as 2 and 17 it would be inappropriate to use this as a measure of average
- Difficult to estimate for grouped data
- The value may be unrepresentative especially if it is zero or the data has a wide range

Hint: You are often asked to explain why the mode is NOT appropriate in exam questions.

Median

Advantages

- Can sometimes be found when some of the data is missing.
- Useful when there are outliers (unusually small/large values).

Disadvantages

- Difficult to estimate for grouped data

Mean

Advantages

- Takes into account all the values.
- Provides a basis for further analysis.

Disadvantages

- If the sample is small it may be unduly affected by outlier or incorrect values.
- Can only be calculated if you know all the data

Hint: As a general rule statisticians favour the use of the mean as a **measure of average** but if it is not appropriate will use the median. Very rarely will they use the mode; this is nearly always the least appropriate measure of average in exam questions.

Range

Advantages

- Easy to calculate

Disadvantages

- Depends only on the extreme values
- Not appropriate for large data sets.
- Cannot be calculated if either of the extreme values are unknown

Interquartile Range

Advantages

- Can sometimes be found when some of the data is missing.
- Useful when there are outliers (unusually small/large values).

Disadvantages

- Difficult to estimate from grouped data.

Standard Deviation

Advantages

- Takes into account all the values.
- Provides a basis for further analysis.

Disadvantages

- If the sample is small it may be unduly affected by outlier or incorrect values.
- Can only be calculated if you know all the data
- Difficult to calculate from large data sets.

Hint: As a general rule statisticians favour the use of the standard deviation as a **measure of spread** but if it is not appropriate they use the IQR or for small sets ($n < 10$) with no outliers, the range.

Example 10

The length of time that customers are put on hold, in minutes, is recorded by a call centre for one hour, the results were as follows:

10 9 0 3 5 5 3 0 2 0 a 1 0 12 8

The value of a is unknown.
Give a reason why, for these values;

a) The mode is **not** an appropriate measure of average

b) The standard deviation is **not** an appropriate measure of spread.

Example 11

Below is an extract from the large data set, the random sample of 10 cars registered in 2016.

a) Calculate the mean and standard deviation of the mass of the car.

Make	GovRegion	EngineSize	YearRegistered	Mass kg
FORD	North West	998	2016	1176
FORD	North West	1242	2016	989
FORD	North West	998	2016	1177
FORD	North West	999	2016	1382
FORD	North West	1596	2016	1211
FORD	North West	1242	2016	1095
FORD	North West	999	2016	1333
FORD	North West	998	2016	1335
FORD	North West	1242	2016	1095
FORD	North West	999	2016	1329

A further sample is taken of 10 cars registered ion 2002.

Make	GovRegion	EngineSize	YearRegistered	Mass kg
FORD	North West	1596	2002	1167
FORD	North West	1596	2002	1189
FORD	North West	1242	2002	1017
FORD	North West	1988	2002	1261
FORD	North West	1388	2002	1163
FORD	North West	1388	2002	1097
FORD	North West	1299	2002	965
FORD	North West	1388	2002	1163
FORD	North West	2495	2002	1467
FORD	North West	1388	2002	1163

b) Sarah claims that as technology has improved cars have a lower mass in 2016 compared to 2002. Use the data to comment on her claim.

Exam Questions

Helen is studying the daily mean wind speed for Camborne using the large data set from 1987. The data for one month are summarised in Table 1 below.

Windspeed	n/a	6	7	8	9	11	12	13	14	16
Frequency	13	2	3	2	2	3	1	2	1	2

Table 1

(a) Calculate the mean for these data.

(1)

(b) Calculate the standard deviation for these data and state the units.

(2)

A survey of 120 adults found that the volume, X litres per person, of carbonated drinks they consumed in a week had the following results:

$$\sum x = 165.6 \qquad \sum x^2 = 261.8$$

(a) (i) Calculate the mean of X.

[1 mark]

(a) (ii) Calculate the standard deviation of X.

[2 marks]