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.

Introduction

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

hair colour, gender, ethnic group

Quantitative - when the variable can be reported numerically

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

shoesize, number of cars passing a point

Continuous can assume any value within a specific range

height
weight
temperature

Where does age fall? continuous or discrete?

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 — All packets of this particular type of sweet
b. a parameter — Population mean of 100 grams
c. a sample — 50 packets
d. a qualitative variable — flavour
e. a continuous variable — weight
f. a discrete variable — number of sweets

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

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

Q_1 Q_2 Q_3

Median = 22

$Q_3 = 30$ $Q_1 = 17$

$IQR = Q_3 - Q = 30 - 17 = 13$

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
cf	1	6	13	27	36	46	49	50

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

a) Range = 7 − 0 = 7 complaints
Mode = 3 complaints
$Q_1 = \frac{n+1}{4}^{th} = \frac{50+1}{4} = 12.75^{th}$ value = 2
$Q_3 = 3\frac{(n+1)}{4}^{th} = 3\frac{(50+1)}{4} = 38.25^{th}$ value = 5
IQR = 5 − 2 = 3

b) Mean $= \bar{x} = 3.44$
$S = Sd = 1.55$

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.

Standard Deviation and Variance of a Sample

For random sample of n values, $X_1, X_2, \ldots\ldots X_n$, the **standard deviation**, denoted by S, is given as

$$S = \sqrt{\frac{\left(\sum X^2 - \frac{(\sum X)^2}{n}\right)}{n-1}} = \sqrt{\frac{\sum (X-\bar{X})^2}{n-1}} \text{ where } \bar{X} = \frac{\sum X}{n}$$

The sample variance, S^2, is the square of the sample standard deviation. S^2 is an unbiased estimator of σ^2.

This is given in the formulae booklet as:

For a random sample $X_1, X_2, \ldots\ldots X_n$ of n independent observations from a distribution having mean μ and variance σ^2

S^2 is an unbiased estimator of σ^2, where $S^2 = \frac{\sum (X_i - \bar{X})^2}{n-1}$

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

Standard dev = $S = 1.59325501 = 1.59$ (3sf)
variance = $S^2 = (1.59326)^2 = 2.54$ (3sf)

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.

$\bar{x} = \frac{\Sigma x}{n} = \frac{321.4}{50} = 6.428$

$s^2 = \frac{\Sigma (x-\bar{x})^2}{n-1}$ $s = \sqrt{\frac{42.62}{49}} = 0.933$ (3sf)

Example 6

The mean blood cholesterol level of the adult residents of a particular country 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. She selected a sample of 80 residents who consumed yogurt daily and 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 this sample of 80 residents. Comment on the values obtained.

$\bar{x} = \frac{\Sigma x}{n} = \frac{452.8}{80} =$

$s = \sqrt{\frac{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}{n-1}} = \sqrt{\frac{2596.4 - \frac{(452.8)^2}{50}}{49}}$

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 to the nearest second and the times are recorded below.

Midpoint
30
45
52.5
57.5
80

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.

$\bar{x} = 54.7$

$s = 17.4$

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.

Midpart

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
	1	5	13	25	41	52	60	63	65

a) Determine values for the median and interquartile range.

Median $= \frac{1}{2}(65+1)^{th} = 33^{rd}$ term $= 14$

$Q_1 = \frac{1}{4}(65+1)^{th} = 16.5^{th}$ term $= 13$

$Q_3 = \frac{3}{4}(65+1)^{th} = 49.5^{th}$ term $= 15$

$IQR = 15 - 13 = 2$

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

No of Toms	4	9	12	13	14	15	16	18	23
f	1	4	8	12	16	11	8	3	2

$\bar{x} = 14.6$ $\quad s = 2.79$

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	65	16	23	27	31	34	16	10	5	3	4	1	**175**

(a) For these data:

(i) state the modal value: 23 *(1 mark)*

(ii) determine values for the median and the interquartile range. *(3 marks)* median 22 $Q_1 = 20$ $Q_3 = 23$

(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. $\bar{x} = 22.3$ $\quad s = 6.39$ *(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)*

Total private 280

Total royal mail $= \sum x = 3902.5$

$$\text{mean} = \frac{280 + 3902.5}{175} = 23.9$$

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

Mode is zero so unrepresentative of the data

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

There is an unknown value so standard deviation can't be calculated.

Example 11

The table below shows an extract from the Purchased quantities of household food & drink survey by Government Office Region and Country published by DEFRA in 2015

Purchased quantities of household food & drink by Government Office Region and Country

Averages per person per week

Region	Description	Units	2006	2007	2008	2009	2010	2011	2012	2013	2014
South East	Bread	g	667	664	632	636	584	569	575	565	529
West Midlands	Bread	g	796	738	705	713	684	674	688	648	610

Comment on the average consumption of bread over time and by region.

The average consumption of bread per person per week has decreased over time for both the south east and the West Mids

Other areas in the survey include London, East Midlands, North East, North West, Yorkshire and Humber, East, South West. What results would you expect from these areas? Have a look at the data set in Excel and see if your predictions are correct.

Linear Scaling

Linear scaling is most often used when the units change or if a question gives you the information involving differences.

$$\text{mean}(aX + b) = a\,\text{mean}(X) + b$$
$$\text{SD}(aX + b) = a\,\text{SD}(X)$$
$$\text{Var}(aX + b) = a^2\,\text{Var}(X)$$

If a number, b, is added to each piece of data then the mean increases by b but the standard deviation and variance remains unchanged. If each piece of data is multiplied by a, the mean and standard deviation are both multiplied by a (the variance is multiplied by a^2).

Example 12

The time, x seconds, in excess of 30 seconds, which a sample of buses waits at the bus stop, have a mean 18 and standard deviation 15. Find the mean and standard deviation of the times that this sample of buses actually waits at the stop.

mean = 18
sd = 15

Actual
Mean = 18 + 30 = 48
sd = 15

Past Exam Question

2 Before leaving for a tour of the UK during the summer of 2008, Eduardo was told that the UK price of a 1.5-litre bottle of spring water was about 50p.

Whilst on his tour, Eduardo noted the prices, x pence, which he paid for 1.5-litre bottles of spring water from 12 retail outlets.

He then subtracted 50p from each price and his resulting differences, in pence, were

−18 −11 1 15 7 −1 17 −16 18 −3 0 9

(a) (i) Calculate the mean and the standard deviation of these differences. *(2 marks)*

$\bar{x} = 1.5$ $s = 12.26$

(ii) Hence calculate the mean and the standard deviation of the prices, x pence, paid by Eduardo. *(2 marks)*

$\bar{x} = 51.5p$ $s = 12.26p$

(b) Based on an exchange rate of €1.22 to £1, calculate, in euros, the mean and the standard deviation of the prices paid by Eduardo. *(3 marks)*

£1 = €1.22

$\bar{x}$ = £0.515 s = £0.1226

× 1.22

$\bar{x}$ = £0.63 s = €0.15