Averages and spread

Check your knowledge of terminology

Match each of the terms below with its definition:

|  |  |  |
| --- | --- | --- |
| MEDIAN |  | The difference between the highest and lowest data values |
| RANGE |  | The number which has 25% of the data values below or equal to it and 75% of the data values above or equal to it |
| UPPER QUARTILE |  | The total of the data values, divided by the number of data values: $\overbar{x}=\frac{\sum\_{}^{}x}{n}$ |
| MODE |  | A data value that does not appear to be valid |
| VARIANCE |  | The difference between the upper and lower quartiles |
| LOWER QUARTILE |  | The data value with the highest frequency |
| STANDARD DEVIATION  |  | The number which has 50% of the data values below or equal to it and 50% of the data values above or equal to it |
| MEAN |  | The average of the **squared** deviations from the mean:$$s^{2}=\frac{\sum\_{}^{}\left(x-\overbar{x}\right)^{2}}{n-1}$$ |
| OUTLIER |  | The number which has 75% of the data values below or equal to it and 25% of the data values above or equal to it  |
| INTERQUARTILE RANGE  |  | A data value that is very large or very small compared to most other data values |
| ANOMALY |  | The square root of the variance: $$s=\sqrt{\frac{\sum\_{}^{}\left(x-\overbar{x}\right)^{2}}{n-1}}$$ |

Can you write better definitions for any of the terms?

Worked examples:

1. Raw data

Raw data refers to a data set where each data value is known individually. All of the different measures of location and dispersion can be calculated from raw data.

**Example**

In an experiment, scientists from CERN were tested for their level of radiation exposure (millisievert, mSv), over a year. The results were as follows:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 12.3 | 21.2 | 19.0 | 13.1 | 17.1 | 18.1 | 24.0 | 15.1 | 15.4 | 21.7 | 18.2 |

1. Calculate the following numerical measures
2. Mean
3. Median
4. Mode
5. Interquartile range
6. Range
7. Variance
8. Standard deviation
9. Interpretation:

The safe level for one year’s exposure is 20.0 mSv. Explain if the following statement is correct, using the data you have just calculated.

‘The scientists at CERN are working within the safe levels of radioactive exposure.’

Worked examples:

1. Frequency table

**Example**

Isabella went up and down the street to find out how many parking spaces each house has. Here are her results:

|  |  |
| --- | --- |
| **Number of parking spaces** | **Frequency** |
| 1 | 15 |
| 2 | 27 |
| 3 | 8 |
| 4 | 3 |

1. Calculate the mean, variance and standard deviation
2. Calculate the median and interquartile range
3. Calculate the mode and the range

Worked examples:

1. Grouped data

**Example**

A survey is conducted to look into the amount of money the average customer spends at a supermarket checkout. This was done with a sample of 100 people. The information was then grouped into the following intervals:

|  |  |
| --- | --- |
| **Amount spent (£)** | **Frequency** |
| 5 ≤ x < 25 | 10 |
| 25 ≤ x < 40 | 13 |
| 40 ≤ x < 70 | 12 |
| 70 ≤ x < 100 | 29 |
| 100 ≤ x < 150 | 23 |
| 150 ≤ x < 200 | 13 |

1. Estimate the mean and standard deviation of these data.
2. Which groups contain the median and interquartile range of these data?

Averages and spread 3

1. Fifty drivers were asked to note exactly how many miles they drive in the next year. One year later the results were given as follows.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 8345 | 25846 | 17543 | 1273 | 5342 |
| 18543 | 7004 | 12735 | 11008 | 24735 |
| 6354 | 8645 | 11899 | 13421 | 7735 |
| 9087 | 6453 | 25065 | 1983 | 22543 |
| 6729 | 8222 | 6483 | 14635 | 7654 |
| 8673 | 8600 | 1365 | 11846 | 21218 |
| 7573 | 9732 | 15846 | 16382 | 17502 |
| 7309 | 8991 | 27635 | 7435 | 9732 |
| 17637 | 18509 | 16735 | 7163 | 9217 |
| 7364 | 8725 | 24447 | 19922 | 2005 |

1. Construct a grouped frequency table of this data, setting your own class boundaries.
2. Calculate the mean using your grouped frequency table.
3. Comment on why you might have a different value for the mean from someone else in the class.
4. Comment on how you could make your value for the mean more accurate.
5. The times, in seconds, taken by 20 people to solve a simple numerical puzzle were:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 17 | 19 | 22 | 26 | 28 | 31 | 34 | 36 | 38 | 39 |
| 41 | 45 | 43 | 47 | 50 | 51 | 53 | 55 | 57 | 58 |

1. Calculate the mean and standard deviation of these times.
2. In fact, 23 people solved the puzzle. However, three of them failed to solve it within the allotted time of 60 seconds.
Calculate the median and interquartile range of the times taken by all 23 people.
3. For the times taken by all 23 people, explain why:
4. the mode is not an appropriate numerical measure
5. the range is not an appropriate numerical measure
6. Katrina receives email messages. The table below shows, for a random sample of 40 weekdays, the number of email messages received by Katrina:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Number of email messages | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| Number of weekdays | 2 | 3 | 5 | 6 | 11 | 7 | 3 | 2 | 1 |

1. Calculate estimates for the mean and standard deviation of the number of email messages received per weekday by Katrina.
2. The number of emails per day for Katrina’s friend Narinder has a mean of 3.5 and a standard deviation of 3.6. Does this mean that Katrina receives more emails each day than Narinder? Explain.
3. In two Year 9 classes, all the students’ heights were recorded as follows:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Height (9A)** | **Frequency** |  | **Height (9B)** | **Frequency** |
| 130 < *h* ≤ 140 | 1 |  | 130 < *h* ≤ 140 | 2 |
| 140 < *h* ≤ 150 | 3 |  | 140 < *h* ≤ 150 | 5 |
| 150 < *h* ≤ 160 | 3 |  | 150 < *h* ≤ 160 | 6 |
| 160 < *h* ≤ 170 | 7 |  | 160 < *h* ≤ 170 | 5 |
| 170 < *h* ≤ 180 | 4 |  | 170 < *h* ≤ 180 | 6 |
| 180 < *h* ≤ 190 | 5 |  | 180 < *h* ≤ 190 | 1 |
| 190 < *h* ≤ 200 | 2 |  | 190 < *h* ≤ 200 | 0 |

Compare the two distributions, with reference to suitable measures of location and dispersion.

1. A sample of people who commute regularly from a town in Surrey into London was asked for an estimate of the time taken on their most recent journey. Their replies are summarised below:

|  |  |
| --- | --- |
| **Time (minutes)** | **Frequency** |
| 35 – | 12 |
| 45 – | 54 |
| 55 – | 68 |
| 65 – | 41 |
| 85 – 105 | 23 |

1. Calculate estimates of the mean and standard deviation of these times.
2. A sample of people who commute regularly from a town in Essex into London were also asked for an estimate of the time taken on their most recent journey. Their replies have a mean of 64 minutes and a standard deviation of 21 minutes.
Compare, briefly, the times estimated by commuters from the two towns.
3. Give two reasons why the data presented in parts a. and b. may not adequately represent typical commuting times from the two towns.