**A2Wd Measures of average and dispersion for a frequency distribution**

In this section, we shall extend the calculation of the three-data descriptors (mean, median and mode) to the situation where we are dealing with calculating **averages from a frequency distribution**.

**Example 1**

A **frequency distribution** is a simple data table that shows how many times entities, or frequencies, fall into every category. To illustrate this, let’s assume that 10 students achieved the following score in their exams as illustrated in Table 2.18.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 75 | 56 | 75 | 60 | 80 | 91 | 80 | 75 | 80 | 75 |

Table 1

If we group the data per the frequency of occurrence, we get the frequency distribution. In this case the frequency distribution would look like this (Score becomes X and Number of students achieving every individual score becomes a frequency, f):

|  |  |
| --- | --- |
| Score, X | Number of students, f |
| 56 | 1 |
| 60 | 1 |
| 75 | 4 |
| 80 | 3 |
| 91 | 1 |

Table 2

Knowing the value of X implies that the data values are predefined and given as a single number (not as a member of a group).

**Example 2**

Consider Example 2, which explores the how we can calculate **averages for a grouped frequency distribution** for the miles travelled by 98 trucks.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 403 | 407 | 407 | 408 | 410 | 412 | 413 |
| 423 | 424 | 424 | 425 | 426 | 428 | 430 |
| 435 | 435 | 436 | 436 | 436 | 438 | 438 |
| 444 | 444 | 445 | 446 | 447 | 447 | 447 |
| 452 | 453 | 453 | 453 | 454 | 455 | 455 |
| 462 | 462 | 462 | 463 | 464 | 465 | 466 |
| 474 | 474 | 475 | 476 | 477 | 478 | 479 |
| 490 | 493 | 494 | 495 | 497 | 498 | 498 |
| 415 | 430 | 439 | 449 | 457 | 468 | 482 |
| 416 | 431 | 440 | 450 | 457 | 469 | 482 |
| 418 | 432 | 440 | 450 | 458 | 470 | 483 |
| 419 | 432 | 441 | 451 | 459 | 471 | 485 |
| 420 | 433 | 442 | 451 | 459 | 471 | 486 |
| 421 | 433 | 442 | 451 | 460 | 472 | 488 |

Table 3

The group frequency distribution for this data set is provided in Table 4.

|  |  |
| --- | --- |
| Mileage | Frequency  f |
| 400 – 419 | 11 |
| 420 – 439 | 23 |
| 440 – 459 | 30 |
| 460 – 479 | 21 |
| 480 – 499 | 13 |

Table 4

The layout is similar to Example 1, except that ‘X’ is the **class mid-point** value that has to be calculated from the class intervals.

To illustrate the calculation process we shall use the Example 2 data set to calculate the following descriptive statistics:

1. Value of the mean.
2. Value of the variance and standard deviation.
3. Value of the mode.
4. Value of the median and interquartile range.
5. Value of any percentile.

Unfortunately, SPSS Statistics does not handle frequency distributions well so to keep the explanation as simple as possible we will use Excel only to aid the calculation process for this section.

### Calculating the mean for a frequency distribution

The **mean value for a frequency distribution** is given by equation (1):

(1)

**Example 3**

Consider Example 2 data set and calculate the mean and standard deviation. The group frequency distribution for this data set provided in Table 4. The value of ‘X’ is the **class mid-point** value that has to be calculated from the class intervals as illustrated in Table 5.

|  |  |  |  |
| --- | --- | --- | --- |
| Mileage | Class Mid-Point, X | Frequency  f | f X |
| 400 – 419 | 409.5 | 11 | 4504.5 |
| 420 – 439 | 429.5 | 23 | 9878.5 |
| 440 – 459 | 449.5 | 30 | 13485 |
| 460 – 479 | 469.5 | 21 | 9859.5 |
| 480 – 499 | 489.5 | 13 | 6363.5 |
|  |  | ∑ f = 98 | ∑ fX = 44091 |

Table 5

The **mean for a grouped frequency distribution** is calculated using equation (1):

Mean

The mean miles travelled by the 98 tucks is 450 miles.

Note that the value of X is the class mid-point which is computed using the true limits of each class. This assumes that the data values within each class vary uniformly between the lowest and highest data values within the class. Also note that even if a distribution has unequal class widths, the same procedure is followed.

**Excel solution**

Figures 1 and 2 illustrate the Excel solution for the mean

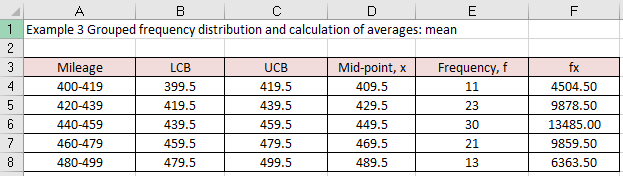


Figure 1

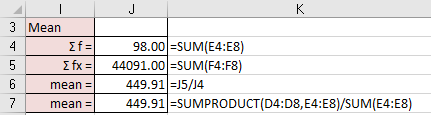


Figure 2

**Excel solution**

Mileage Cells A4:A8 Values

LCB Cells B4:B8 Values

UCB Cells C4:C8 Values

Mid-point x Cell D4 Formula: =(B4+C4)/2

Copy down formula D5:D8

Frequency, f Cell E4:E8 Values

fx Cell F4 Formula: =E4\*D4

Copy down formula F5:F8

Mean Σ f = Cell B11 Formula:=SUM(E4:E8)

Σ fx = Cell B12 Formula:=SUM(F4:F8)

mean = Cell F11 Formula:=J5/J4

mean = Cell F12 Formula:=SUMPRODUCT(D4:D8,E4:E8)/SUM(E4:E8)

### Calculating the variance and standard deviation for a frequency distribution

To calculate the **variance for a grouped frequency distribution** (and **standard deviation for a grouped frequency distribution**) we use equations (2) and (3):

(2)

(3)

**Example 4**

We will re-consider Example 2, a grouped frequency data set to calculate the following descriptive statistics (mean, variance, and standard deviation) as illustrated in Table 6.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Mileage | Class Mid-Point, X | Frequency  f | f X | F X2 |
| 400 – 419 | 409.5 | 11 | 4504.5 | 1844592.75 |
| 420 – 439 | 429.5 | 23 | 9878.5 | 4242815.75 |
| 440 – 459 | 449.5 | 30 | 13485 | 6061507.5 |
| 460 – 479 | 469.5 | 21 | 9859.5 | 4629035.25 |
| 480 – 499 | 489.5 | 13 | 6363.5 | 3114933.25 |
|  |  | ∑ f = 98 | ∑ f X = 44091 | ∑ f X2 = 19892884.50 |

Table 6

From Table 6:

∑ f = 98, ∑ f X = 44091, and ∑ f X2 = 19892884.50

The mean value is given by equation (1):

Mean

Mean miles travelled = 449.9 to 1 decimal place.

The variance is given by equation (2):

Therefore, the standard deviation is given by equation (3):

Standard deviation, SD(X) = 23.90 miles to 2 decimal places

**Excel Solution**

Figures 3 and 4 illustrate the Excel solution.

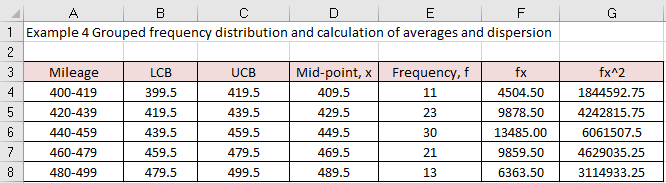


Figure 3

**Excel solution**

Mileage Cells A4:A8 Values

LCB Cells B4:B8 Values

UCB Cells C4:C8 Values

Mid-point x Cell D4 Formula: =(B4+C4)/2

Copy down formula D4:D8

Frequency, f Cell E4:E9 Values

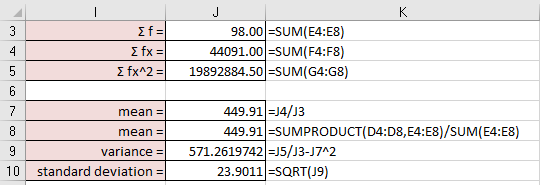


Figure 4

**Excel solution**

fx Cell F4 Formula: =E4\*D4

Copy down formula F4:F8

fx^2 Cell G4 Formula: =E4\*D4^2

Copy down formula G4:G8

Σ f = Cell B10 Formula:=SUM(E4:E8)

Σ fx = Cell B11 Formula:=SUM(F4:F8)

Σ fx^2 = Cell B12 Formula:=SUM(G4:G8)

Mean = Cell B14 Formula:=J4/J3

Mean = Cell B15 Formula:=SUMPRODUCT(D4:D8,E4:E8)/SUM(E4:E8)

Variance = Cell B16 Formula:=J5/J3-J7^2

Standard deviation = Cell B17 Formula:=SQRT(J9)

The interpretation of all the measures remains the same as for the previous example. The average number of miles travelled is 454 miles and the standard deviation is 27.4 miles. As we will learn shortly, if the data points were distributed in accordance with the normal distribution, then 68% of all the data would be within the range of = ± 1s. In our case, 454 ± 27.4 = 426.6 to 481.4. One way to express this is to say that 68% of the time, the number of miles travelled will be between 426.6 to 482.4 with an average value of 454 miles.

### Calculating the mode for a frequency distribution

As we now know, the mode is the most frequently occurring score. This value can be determined directly from a frequency distribution or a histogram.

**Example 5**

If we consider the distribution given in this Example 2, we can see that the most frequently occurring class is the class 440 - 459 miles. This class is known as the **modal class**. If we look at the histograms associated with this example (Figure 2.80), the mode is very apparent: it is the class with the highest rectangle. We can estimate the mode for a **grouped frequency distribution** using a formula or graphical method:

(i) Formula method

Having established which is the class interval with the highest frequency (the modal class), then the mode can now be estimated using equation (4):

Mode =  (4)

Where: L = lower class boundary of the modal class, f0 = frequency of the class below the modal class, f1 = frequency of the modal class, f2 = frequency of the class above the modal class, and c = modal class width. Note that this formula only works if the modal class and the two adjacent classes are of equal width. Therefore, in this example we have L = 439.5, f0 = 23, f1 = 30, f2 = 21 and c = 20.

Mode = = 448 miles to the nearest mile

The modal distance travelled is 448 miles.

(ii) Graphical method

We can estimate the mode graphically by constructing the histogram. In this example the data set class widths are of the same size (constant) and therefore the height of the column will represent the frequency of occurrence (f). We can see from the histogram that class 440 – 459 is the modal class and the mode will lie within this class. The frequencies of the two adjacent classes (420 – 439, 460 – 479) can be used to estimate the value of the mode:

1. Construct the two crossed diagonals (see dashed lines in histogram).
2. Drop a perpendicular from where these two lines meet to the horizontal axis.
3. Read from the horizontal axis the value estimate for the mode.

Figure 5 illustrates the graphical solution to estimate the modal value.

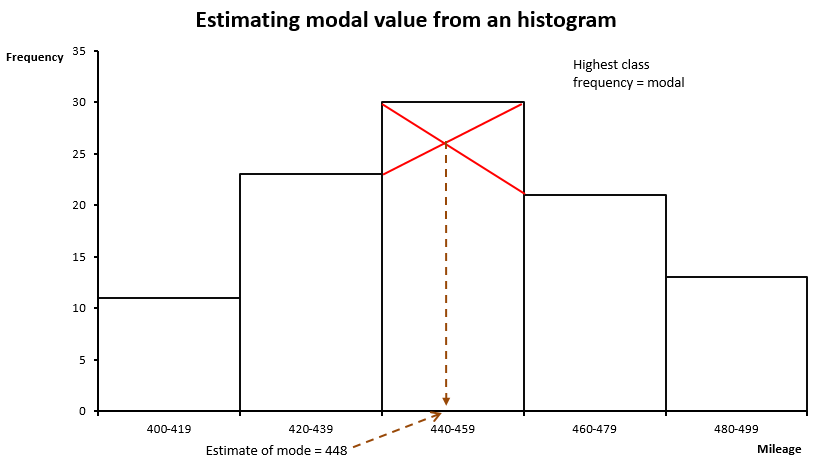


Figure 5

The modal value is estimated from the histogram to be 448 miles travelled.

**Excel solution**

Figures 6 and 7 illustrate the Excel solution for the mode

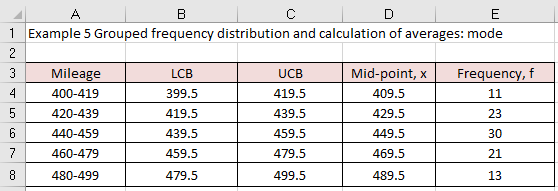


Figure 6

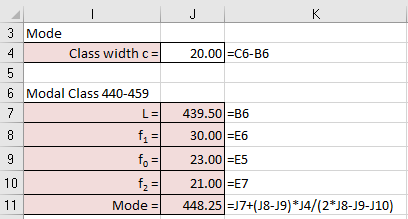


Figure 7

**Excel solution**

Mileage Cells A4:A8 Values

LCB Cells B4:B8 Values

UCB Cells C4:C8 Values

Mid-point x Cell D4 Formula: =(B4+C4)/2

Copy down formula D5:D8

Frequency, f Cell E4:E8 Values

Mode Class width c = Cell F15 Formula:=C6-B6

Modal Class 440-459

L = Cell F18 Formula:=B6

f1 = Cell F19 Formula:=E6

f0 = Cell F20 Formula:=E5

f2 = Cell F21 Formula:=E7

Mode = Cell F22 Formula:=J7+(J8-J9)\*J4/(2\*J8-J9-J10)

### Calculating the median for a frequency distribution

As with cases where X is known, finding the **median from a grouped frequency distribution** where X is class mid-point involves some further calculations. The cumulative frequency distribution and cumulative frequency polygon (or ogive) are used.

**Example 6**

If we consider the distribution given in Example 2, the median of the 120 values is given by the (98 + 1)/2th value (or 49½th value) and this value lies in the class (440 - 459) miles.

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Cumulative Frequency Distribution | |
| Mileage | Frequency | Upper Class Limit  UCL, XCF | Cumulative Frequency  CF |
| 400 – 419 | 11 | < 419.5 | 11 |
| 420 – 439 | 23 | < 439.5 | 34 |
| 440 – 459 | 30 | < 459.5 | 64 |
| 460 – 479 | 21 | < 479.5 | 85 |
| 480 – 499 | 13 | < 499.5 | 98 |

Table 7 Cumulative frequency distribution for Example 2

An estimate of the value of the median within that class can be determined either by calculation or by using a graphical method:

(i) Formula method

Equation (5) can be used to estimate the median:

Median =  (5)

Where: L = true lower-class boundary of the median class, C = median class width, F = cumulative frequency before the median class, f = frequency within the median class, N = total frequency.

Position of Median = 49 1/2th number value in the ordered list

Median Class = Between 420 – 439 and 440 – 459, because 49.5 is between 34 and 64, which are the cumulative frequencies for these two classes.

From Table 7:

L = 439.5, C = 459.5 - 439.5 = 20, F = 34, f = 30, and N = 98.

Median = = 450 miles

The median number of miles travelled is 450 miles. This is reasonably close to the value obtained for the mean (449 miles) and we would expect this given that the histogram for miles travelled looks quite symmetrical. The concept of symmetry, and a measure of how symmetrical a distribution is, will be explored when discussing the concept of skewness.

(ii) Graphical method

Figure 8 represents a cumulative frequency curve (or ogive) for the miles travelled example. We can use this curve to provide an estimate of the median. The median is then the value which corresponds to half of the total frequency.

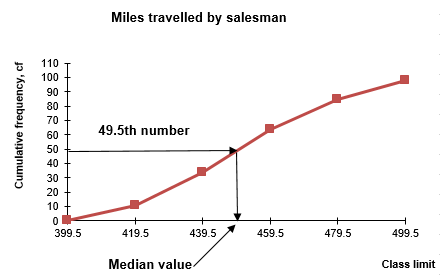


Figure 8

Therefore, the Position of the Median = . We can now use the cumulative frequency curve to estimate the value of the median. From the graph, the median is approximately 450 miles.

**Excel solution**

Figures 9 and 10 illustrate the Excel solution for the median

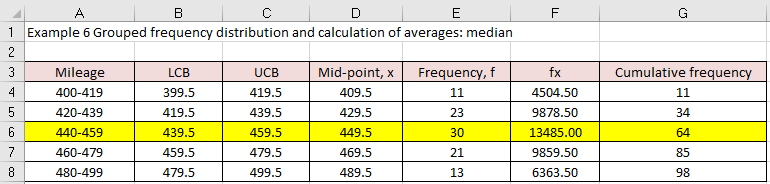


Figure 9

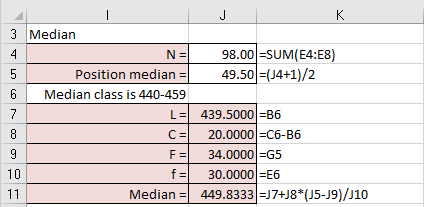


Figure 10

**Excel solution**

Mileage Cells A4:A8 Values

LCB Cells B4:B8 Values

UCB Cells C4:C8 Values

Mid-point x Cell D4 Formula: =(B4+C4)/2

Copy down formula D5:D8

Frequency, f Cell E4:E8 Values

fx Cell F4 Formula: =E4\*D4

Copy down formula F5:F8

CF Cell G4 =E4

Cell G5 =G4+E5

Copy down formula G6:G8

Median N = Cell B15 Formula:=SUM(E4:E8)

Position median = Cell B16 Formula:=(J4+1)/2

Median class is 440-459

L = Cell B18 Formula:=B6

C = Cell B19 Formula:=C6-B6

F = Cell B20 Formula:=G5

f = Cell B21 Formula:=E6

Median = Cell B22 Formula:= J7+J8\*(J5-J9)/J10

From Excel, the average number of miles travelled are mean = 450, median = 450, and mode = 448.

### Calculating the percentile for a frequency distribution

Equation (6) can be used to calculate the value of any **percentile for a frequency distribution:**

Percentile value P =  (6)

Where: L = true lower-class boundary of the percentile class, C = percentile class width, F = cumulative frequency before the percentile class, f = frequency within the percentile class, N = total frequency.

**Example 7**

In example 6, we calculated the median value as 450 miles. The median is the 50th percentile value in the grouped frequency distribution. Equation (6) can be used to calculate the percentile value for any grouped frequency distribution. For example, if we required the 25th percentile value then:

Position of 25th percentile = 25/100 (98 + 1) = 24.75th number value. From table 2.24, the 25th percentile class is 420 – 439 which contains the 24.75th number value. Therefore, L = 419.5, C = 439.5 – 419.5 = 20, F = 11, f = 23, N = 98. Substituting these values into equation (2.25) gives the value of the percentile:

25th percentile value =

25th percentile value = 431.46

The 25th percentile value is equal to 431 to the nearest whole number.

Excel solution

Figures 11 and 12 illustrate the Excel solution.

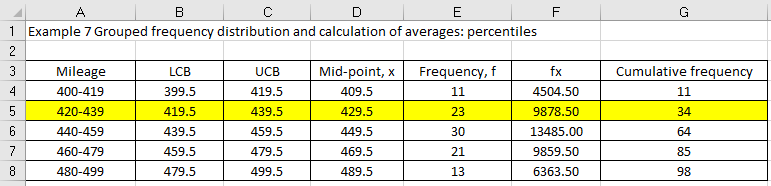


Figure 11

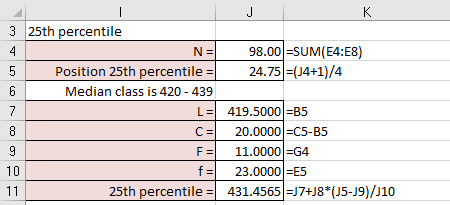


Figure 12

**Excel solution**

Mileage Cells A4:A8 Values

LCB Cells B4:B8 Values

UCB Cells C4:C8 Values

Mid-point x Cell D4 Formula: =(B4+C4)/2

Copy down formula D5:D8

Frequency, f Cell E4:E8 Values

fx Cell F4 Formula: =E4\*D4

Copy down formula F5:F8

CF Cell G4 =E4

Cell G5 =G4+E5

Copy down formula G6:G8

Median N = Cell B15 Formula:=SUM(E4:E8)

Position median = Cell B16 Formula:=(J4+1)/2

Median class is 440-459

L = Cell B18 Formula:=B5

C = Cell B19 Formula:=C5-B5

F = Cell B20 Formula:=G4

f = Cell B21 Formula:=E5

Median = Cell B22 Formula:= J7+J8\*(J5-J9)/J10

From Excel, 25th percentile value = 431.5.

### Weighted averages

For a simple mean we calculated the value using equation (2.1) which assumes that each value of X is of equal importance to all other values of X. In many cases we are faced with a situation where this is not true, for example, module grades are often computed using a **weighted average** since a different weighting is applied to different assessments. However, from the calculation point of view, this is no different to the method of calculating averages from frequency distributions. The weighted average is calculated using equation (7) (which is identical to equation (1).

 (7)

Where w is the level of importance placed on each assessment element and X is the actual mark associated with this weight. Note that equation (7) is identical to equation (1), except that the word and symbol for frequency has been replaced with the word and symbol for weight.

**Example 8**

Suppose that a statistics module is assessed via a series of assessments: multiple choice questions (mcq), in-course assignment (ica), end assignment (ea). The weighting given to every type of assessment is 20%, 30%, and 50% respectively. The actual marks awarded to one student were 74, 66, and 88. Calculate the weighted average. Figure 13 illustrates the Excel solution.

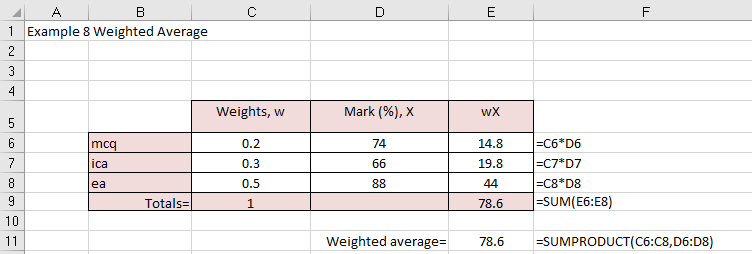


Figure 13

**Excel solution**

mcq, w Cell C6 Value

mcq, X Cell D6 Value

ica, w Cell C7 Value

ica, X Cell D7 Value

ea, w Cell C8 Value

ea, X Cell D8 Value

wX Cell E6 Formula:=C6\*D6

Copy formula down E6:E8

Total Cell E9 Formula:=SUM(E6:E8)

weighted average = Cell E11 Formula = SUMPRODUCT (C6:C8, D6:D8)

This student’s weighted average and, therefore module grade, would be 79%.

Note that if all the weights are equal, then the weighted mean is the same as the arithmetic mean. As emphasised before, Excel does not contain a built-in function to calculate a weighted average. Again, the =SUMPRODUCT() function is used. If the weights are given in percentages then the formula would be modified to =SUMPRODUCT(C6:C8,D6:D8)/SUM(C6:C8).

### Check your understanding:

X1 Cameos Ltd is employed by a leading market research organisation based in Berlin. The company is discussing with the firm contracted to expand the catering facilities provided to its employees whether to include a greater range of products. The initial research by Cameos has identified the following set of spend (€) per week by individual employees (Table 8).

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 22 | 16 | 26 | 33 | 33 | 37 | 9 | 23 | 32 | 17 |
| 20 | 13 | 12 | 18 | 19 | 10 | 21 | 22 | 25 | 22 |
| 22 | 22 | 34 | 24 | 23 | 21 | 38 | 31 | 41 | 20 |

Table 8

(a) Plot the histogram and visually comment on the shape of the weekly expenditure. Hint: use class width of 5.

(b) Calculate the values of the mean and median.

(c) Use descriptive statistics in conjunction with the histogram to comment on weekly expenditure.

X2 Form a frequency distribution of the following data given in Table 9 with intervals centred at 10, 15, 20, 25, 30, 35, 40, and estimate the mean value.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 9 | 26 | 33 | 24 | 41 | 24 | 37 | 39 | 30 | 28 |
| 24 | 42 | 17 | 26 | 18 | 33 | 40 | 28 | 31 | 20 |
| 32 | 21 | 39 | 25 | 16 | 17 | 26 | 11 | 30 | 28 |
| 34 | 24 | 19 | 23 | 27 | 18 | 32 | 21 | 40 |  |

Table 9

X3 The frequency distribution of the length of a sample of 98 nails is presented in Table 10 (measured to the nearest 0.1 mm).

|  |  |
| --- | --- |
| Length | Frequency |
| 4.0 - 4.2 | 4 |
| 4.3 - 4.5 | 9 |
| 4.6 - 4.8 | 13 |
| 4.9 - 5.1 | 20 |
| 5.2 - 5.4 | 34 |
| 5.5 - 5.7 | 18 |

Table 10

(a) Find the mean length of this sample by hand and by using a spreadsheet.

(b) Construct the cumulative frequency graph and use this to estimate the median.

(c) Check the value of the median using the formula method.

X4 The distribution of marks of 400 candidates in an A-level examination is presented in Table 11.

|  |  |
| --- | --- |
| Marks | Frequency, f |
| 0 - 10 | 6 |
| 11 - 20 | 15 |
| 21 - 30 | 31 |
| 31 - 40 | 80 |
| 41 - 50 | 93 |
| 51 - 60 | 69 |
| 61 - 70 | 54 |
| 71 - 80 | 33 |
| 81 - 90 | 12 |
| 91 - 100 | 7 |

Table 11

(a) Calculate the mean value.

(b) Construct the cumulative frequency curve and estimate the median, first and third quartile values.

X5 The data in Table 12 represents the life expectancy for males and females in 10 different European Union countries in 2014 when they are 65 years old in 2014.

|  |  |  |
| --- | --- | --- |
| EU country life expectancy from the age of 65 in 2014 | Men | Women |
| Denmark | 18.1 | 20.8 |
| Germany | 18.2 | 21.4 |
| Ireland | 18.4 | 21.1 |
| Spain | 19.3 | 23.5 |
| France | 19.7 | 24.0 |
| Italy | 19.2 | 22.8 |
| Netherlands | 18.6 | 21.4 |
| Sweden | 18.9 | 21.6 |
| United Kingdom | 18.8 | 21.3 |
| Norway | 18.8 | 21.6 |

Table 12

1. Calculate the median male life expectancy age when male is 65 in 2014.
2. Calculate the median female life expectancy age when female is 65 in 2014.
3. What do your answers to part (a) and (b) tell you about life expectancy across these ten EU countries?

X6 Over a one-month period the number of vacant beds in a West Yorkshire hospital was surveyed. The following frequency distribution resulted

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Beds Vacant | 0 | 2 | 3 | 5 | 6 | 8 |
| Frequency | 4 | 8 | 12 | 4 | 2 | 1 |

Table 13

Determine the mean and standard deviation.

X7 In several towns the distance of 1 sample of 122 supermarkets from the towns’ high street was measured to the nearest metre

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Distance | 57-59 | 60-62 | 63-65 | 66-68 | 69-71 | 72-74 | 75-77 |
| Frequency | 9 | 10 | 18 | 42 | 27 | 11 | 5 |

Table 14

Determine the Range, Mean and Standard Deviation.

X8 In a debate on the alteration of a traffic system in the city centre, measurements of the number of cars per minute were taken at two junctions, the results were as follows

|  |  |  |
| --- | --- | --- |
| Number of Cars per minute | Junction A | Junction B |
| 10 - 14 | 0 | 5 |
| 15 - 19 | 3 | 8 |
| 20 - 24 | 13 | 10 |
| 25 - 29 | 24 | 12 |
| 30 - 34 | 17 | 14 |
| 35 - 39 | 3 | 5 |
| 40 - 44 | 0 | 3 |
| 45 - 49 | 0 | 3 |
| Totals | 60 | 60 |

Table 15

Compare the two distributions by plotting out their frequency polygons and determine the means and standard deviations.

X9 An investor acquires 300 shares of a company in year 1 at €25 and 150 shares of the same company in year 2 at €42. Calculate the weighted average price paid for the shares.

X10 A module assessment consists of assessed quizzes, homework, and three examinations. The assessment weighting and results for a student are given in Table 16. Calculate the weighted average mark awarded to the student for this module.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Assessment type | Quiz | Homework | Exam 1 | Exam 2 | Exam 3 |
| Weighting (%) | 10 | 10 | 20 | 20 | 40 |
| Awarded mark (%) | 80 | 89 | 79 | 84 | 87 |

Table 16