

Frequency, Distribution and Averages

1 Averages

It is a mistake common to many researchers new to data analysis that the “best representative” or “most typical” of a set of numerical data values (volume of sales per month, variation per quarter in interest base rate, amount of wastage per week in a supply chain) is the “average” of the data set (more properly, the “arithmetic mean”). For example: if your product sales in the past quarter have been: Month 1: 107, Month 2: 36, Month 3: 34 – then your most **typical** monthly sales volume is: $(107 + 36 + 34) / 3 = 59$ unit sales – WRONG!

In this session we revisit the three types of average, the “arithmetic” mean, the median and the mode, and consider some other mathematical “means” (often ignored) which stake an equal claim to being the “most typical” value of a set of data values – the “geometric” and “harmonic” means.

We also consider the range of a set of data values and its usefulness in discovering a “most typical” value

Arithmetic Mean =	$\frac{\text{sum of the values}}{\text{number of values}}$
Median =	middle value (when the data is arranged in order); [where there are two central values, the median is their mean].
Mode =	most commonly occurring value.
Range =	difference between largest and smallest values.
Geometric Mean =	the n^{th} root of the product of n data values.
Harmonic Mean =	the reciprocal of the arithmetic mean of the reciprocals of the data values.



Example 1: Here is a list of customer complaints received in the past 9 weeks:

1 7 3 8 2 3 6 5 10

For this sample,

- (a) calculate the mean,
- (b) determine the median,
- (c) state the mode,
- (d) calculate the range.



Solution

$$\begin{aligned}
 \text{(a) Mean} &= \frac{1 + 7 + 3 + 8 + 2 + 3 + 6 + 5 + 10}{9} = \\
 &= 45 / 9 = \\
 &= 5 \text{ (the arithmetic mean)}
 \end{aligned}$$

(b) To find the median, first write the numbers in order.

1 2 3 3 **5** 6 7 8 10

As the number of data items is odd, the median (the middle number) is “5” in this case, so:

Median = **5**

(c) The mode is the most common value, which is **3** for this set of values.

Mode = **3**

(d) Range = $10 - 1 = 9$

Comment:

For this simple example, the arithmetic mean (5) has some claim to being the most typical value in the data set (because it also agrees with the median value). Yet the range (9) is large, relative to the data values – so the “average” value may not be a reliable indicator of weekly customer complaints.



Example 2: A list of shifts when staff were absent from work for a manufacturing company (the ICC semi-finals and finals were played on day 3 and days 6, respectively):

Determine the median of the following set of values (in staff/days absence):

44 32 88 19 33 74 62 31



Solution

First write the numbers in order:

19 31 32 33 44 62 74 88

In this case, there are 2 middle numbers, 33 and 44. The median will be the mean of these.

Median = $(33 + 44) / 2 = 38.5$

Comment:

For this example, the median (38.5) has little claim to being the most typical value in the data set (because there are extreme outliers (62 and 88) on days when higher absence rates might be expected). If one eliminates the outliers (extremes), then the median value is $(32 + 33) / 2 = 32.5$. This may be a better representative of the most typical number of staff/absence days.



Example 3: A survey collected data on the number of people living in each home in a given residential area, which is shown in the following table:

Number of people living in home	Frequency
2	3
3	9
4	10
5	2
6	3
7	1
8	1
9	0
10	1

- (a) Calculate the mean number of people living in each home.
- (b) Determine the median of the data.
- (c) State the mode of the data.



Solution

- (a) The first step is to complete the table below:

Number of people living in home	Frequency	Cumulative total
2	3	6
3	9	27
4	10	40
5	2	10
6	3	18
7	1	7
8	1	8
9	0	0
10	1	10
Totals	30	126
Mean	(126 / 30) =	4.2

Mean residents per home= 4.2 people.

- (b) As there are 30 values, the median is the mean of the 15th and 16th values:

2,2,2,3,3,3,3,3,3,3,3,3,4,4,4,4,4,4,4,4,5,5,6,6,7,8,10

From the first table we can see that both the 15th and 16th values are 4, so the median is 4 people per home.

- (c) The most common value is 4 so the mode is 4 people per home.

Comment:

For this example, the median (4), the mode (4) and the arithmetic mean (4.2) are very close to each other. The range is 9 – which places the mid-range value at 4.5. Thus, there is good justification for taking the mean value (4.2) as close to the “average” number of residents per home in the sample



Exercises

1. Calculate the mean and the range of each of the following sets of data:

(a) 3 17 5 6 12

(b) 30 42 19 33 62

(c) 7 8 3 14 9 13 22

(d) 114 115 110 119 118 123 133

In each case, formulate a short comment on why the (arithmetic) mean may / may not be the most representative value.

2. Determine the median and the mode of each of the following sets of data:

a) 8 5 19 32

b) 33 14 16 19 22

c) 5 9 19 3 14 21 7

d) 11 21 19 13 16 11 19 22 20

In each case, formulate a short comment on why the (arithmetic) mean may / may not be the most representative value.

3. In which of the following data sets is the **mean** the same as the **median**?:

a) 34 6 19 17

b) 29 12 17 18 44 13 40

c) 101 107 183 51 57 77 100 92

d) 27 92 56 83 45

4. Which of the following data sets has the largest range:

a) 14 27 88 73 56 61

b) 374 521 628 314 729

c) 888 912 897 907 887 893

5. The following table gives the results of a survey question asking people how many television sets they had in their home.

Number of Televisions	Frequency
0	3
1	18
2	64
3	73
4	22
5	14
6	2

For this data,

- calculate the mean,
- determine the median,
- state the mode.
- comment briefly upon which may represent the most “typical” value for this data set.

6. A shop manager recorded the number of customers entering her shop each hour for three days. The data she collected is listed below.

Day1: 16 22 17 6 5 8 32 15 9 7 14 33

Day2: 21 11 6 5 11 14 12 22 19 11 14 3

Day3: 14 7 23 41 32 16 5 19 14 33 7 12

For this data:

- calculate the mean,
- determine the median,
- determine the mode,
- calculate the range.

Comment upon which of the 3 averages the manager should use to determine in which hour / day she needs more shop assistants available

8. For the set of data given below, calculate the mean and determine the median.

4 7 3 9 5 6 142 7 3 11

Describe the advantages of using the median, rather than the mean in this case.

1.1 Other Means



The *arithmetic mean* is often not the most suitable for types of data when the data values are not measured in the same units. In these situations, the arithmetic mean will usually give misleading results.

For example: Suppose, as a potential investor, you wish to compare two companies (A and B) in relation to two criteria measured over 2 years: each company's *environmental sustainability* and each company's *financial viability*. Each criterion is measured and reported in a different way:

Environmental sustainability is measured on a scale of 0 to 5, where 0 represents the weakest and 5 the strongest value for sustainability.

Financial viability is reported as a percentage indicating the relative strength of capital against liabilities, year by year.

Companies A and B report the following data values for years 1 and 2:

Company A	Year 1	Year 2
<i>Environmental sustainability (0-5)</i>	3	3
<i>Financial viability (%)</i>	80%	90%
Company B	Year 1	Year 2
<i>Environmental sustainability (0-5)</i>	2	5
<i>Financial viability (%)</i>	80%	85%

If an arithmetic mean were used, the financial viability of each company is given more weight because its numeric range is larger - so a small percentage change in the financial rating (e.g. going from 80 to 90 (10%)) makes a much larger difference in the arithmetic mean than a large percentage change in environmental sustainability (e.g. going from 2 to 5 (150%))

The average (arithmetic mean) increase for Company A is 88% for both criteria (ES and FV), whereas the mean increase for Company B is 84% for both ES and FV. This suggests Company A should be preferred due to its greater average yearly increase – but this is entirely **wrong**. Company A has made a 10% increase in FV but no increase in ES. Company B has made a 5% increase in FV but a 150% increase in ES. In terms of improvement across both criteria, B is obviously preferable to A, but the arithmetic mean increase is misleading us.

The solution is to use a different “mean” – the **geometric mean**, because this calculation has the effect of “normalizing” the ranges of the different data values. The use of a geometric mean “normalizes” the ranges being averaged, so that no range dominates the weighting, and a given percentage change in any of the properties has the same effect on the geometric mean. Calculating the geometric mean of groups of data values uses the following formula:

Geometric Mean (gm) = the n^{th} root of the product of the n values in the data set.

For example: If the data values for ES increase from 4 to 4.8 (a 20% increase), then this change has the same effect on gm of ES as a 20% increase in FV (from, say, 60 to 72).



Exercise

9. Using the data given for ES and FV for each company A and B, above, calculate the arithmetic mean increase from year 1 to year 2. For the same data values, calculate the geometric mean increase from year 1 to year 2. Comment upon how the use of the geometric mean provides a better estimate of average increase overall.



Finally, a less commonly used “mean” – the third of the (so-called) Pythagorean means – is the *Harmonic Mean*. The harmonic mean is the preferred method for averaging multiples, such as the price/earning ratio (the P/E ratio is typically defined as the market price per share divided by the annual earnings per share). If these ratios are averaged using an arithmetic mean (a common error), high data points are given greater weight than low data points. The harmonic mean, on the other hand, gives equal weight to each data point.

The *harmonic mean* is defined to be the reciprocal of the arithmetic mean of the reciprocals of the data values. For example, given the positive data values $\{1, 2, 4\}$, their arithmetic mean is obviously $7/3 = 2.3333$; whereas the harmonic mean is calculated as $n / ((1/x_1) + (1/x_2) + (1/x_3)) = 3 / ((1/1) + (1/2) + (1/4)) = 1 / (1/3) * ((1/1) + (1/2) + (1/4)) = 12 / 7 = 1.7143$.

Note that, typically, the harmonic mean will be lower than the arithmetic mean, because lower data values have equal weight to higher values. This is especially significant when comparing ratios or rates of change.

For example: if one is calculating the average speed for a number of fixed-distance trips, then, if the speed for going from point *A* to *B* was 60 km/h, and the speed for returning from *B* to *A* was 40 km/h, then the average speed (harmonic mean) is given by $2 / ((1/60) + (1/40)) = 48$, which is the mean **speed** for the trips. Compare this to the arithmetic mean $(60 + 40) / 2 = 50$. If a vehicle travels a certain distance at a speed of 60 kilometres per hour and then the same distance again at a speed of 40 kilometres per hour, then its average speed is the *harmonic mean* of 60 and 40 (48 kilometres per hour), and its total travel time is the same as if it had traveled the whole distance at that average speed. However, if the vehicle travels for a certain amount of *time* at a speed 60 and then the same amount of time at a speed 40, then its average speed is the arithmetic mean of 60 and 40, which in the above example is 50 kilometres per hour.

2 Grouped Data



Grouped data arises when data values are divided into sub-groups according to different ranges of possible values – for example, when subjects are grouped according to age-range (15-24, 25-34, . . . and so on) or, perhaps, by income band (£10K-£14K, £15K-£19K, and so on). When dealing with grouped data it is important to think about the type of data that is being processed. You also have to decide the range of values that each group may contain. When calculating a mean for grouped data, we assume that all the values lie at the midpoint of the group.

These ideas are illustrated in the following example.

Example:

The table below shows the times taken by a group of workers to complete an operational task. Their times have been recorded to the **nearest minute**.

Illustrate the data using a bar chart and a frequency polygon.

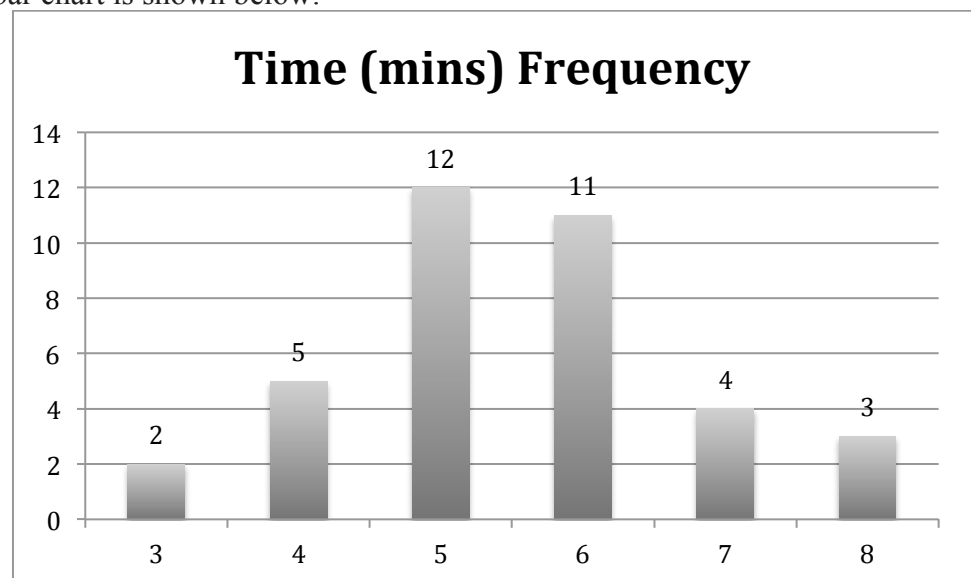
Time (mins)	3	4	5	6	7	8
Frequency	2	5	12	11	4	3



Solution

A time of 5 minutes actually means a time that is greater than or equal to 4.5 mins but less than 5.5 mins. Thus, the bar representing this time on the bar chart will begin at 4.5 and end at 5.4. Similarly, the bar for a time of 3 will begin at 2.5 and end at 3.4.

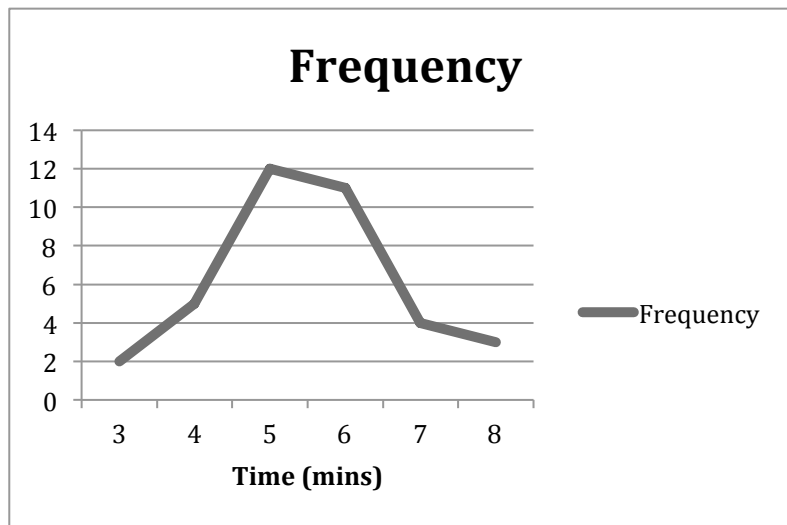
The bar chart is shown below:



The frequency polygon is shown overleaf. We obtain it by joining the mid-points of the tops of the bars from the previous graph. The purpose of plotting a frequency polygon is to provide a visual representation of the distribution of values (for each group). In this case, the polygon approximates to a “bell curve” showing a near “normal distribution” of frequency values for the numbers of employees clustering around a time of 4.5 – 6.4 minutes for the task. [We discuss “normal” distributions later in these workshops].



Frequency polygon:



Since the frequency polygon shows the mid-point of each bar (group of values), it can be used to provide an estimate of the (arithmetic) mean. If we tabulate the intervals for each (time) group, read the mid-point for each group, then multiply each point with the frequency for each group, then the estimated mean can be calculated as follows:

Group	Mid-point	Frequency	Frequency * Mid-point
$2.5 \leq x < 3.4$	3.0	2	6.0
$3.5 \leq x < 4.4$	3.9	5	19.5
$4.5 \leq x < 5.4$	5.0	12	60.0
$5.5 \leq x < 6.4$	6.1	11	67.1
$6.5 \leq x < 7.4$	6.9	4	27.6
$7.5 \leq x < 8.4$	8.2	3	24.6
Totals		37	204.8

Thus, the mean task time can be estimated as $204.8 / 37 \approx 5.54$ minutes.



Exercise:

10. A market gardener records the height after 4 weeks of some tomato plants grown from seed. Her results are recorded in the following table:

Height (cm)	11 – 15	16 – 20	21 – 25	26 -30	31 – 35	36 - 40
Frequency	3	7	19	20	11	2

- Draw a frequency polygon for the data. (*Use the back of this sheet for this*)
- Calculate an estimate of the mean height of the plants.



3

Cumulative Frequency

Cumulative frequencies are easy to calculate from a frequency table. Cumulative frequency graphs can then be used to estimate the median of a set of data. In this session we also look at the idea of quartiles, the interquartile range and the semi-interquartile range. These lead naturally into our investigation of “distributions” which are critical to understanding samples, sampling means, dispersions and variances from the mean (especially standard deviation).

When you have a set of n data values, in order (from lowest to highest):

$$\text{Lower quartile} = \frac{n+1}{4}^{\text{th}} \text{ value}$$

$$\text{Median} = \frac{n+1}{2}^{\text{th}} \text{ value}$$

$$\text{Upper quartile} = \frac{3n+1}{4}^{\text{th}} \text{ value}$$

$$\text{Interquartile range} = (\text{upper quartile} - \text{lower quartile})$$

$$\text{Semi-interquartile range} = (\text{interquartile range} / 2)$$

If the data values are arranged in an ordered list, and the number of values, n , is odd, then the median $((n+1)/2^{\text{th}})$ will be a single value from the list. For example, if there are 95 values ($n = 95$), then the median will be $((95+1)/2) =$ the 48th value. If n is even, however, then $((n+1)/2)$ will fall between two median values, so we take their mean value as the median. For example, if $n = 156$, then $((156+1)/2) = 78.5$, which tells us that, to get the median, we take the average of the 78th and 79th values. For very large data sets (thousands of data values), we would usually estimate the lower quartile, median and upper quartile using the $(n/4)^{\text{th}}$, $(n/2)^{\text{th}}$ and $(3n/4)^{\text{th}}$ values.

Example:

For the following set of data values:

4 7 18 3 9 5 10

- determine the median.
- determine the interquartile range.
- calculate the semi-interquartile range.



Solution:

First list the values in order:

3 4 5 7 9 10 18

- Since there are 7 values, the median will be the $((7+1)/2)^{\text{th}}$ value = 4th value
So the median = 7.
- The lower quartile will be the $((7+1)/4)^{\text{th}}$ value = 2nd value. So the lower quartile = 4.
- The upper quartile will be the $(3(7+1)/4)^{\text{th}}$ value = $24/4 = 6^{\text{th}}$ value. So the upper quartile = 10.
Since the interquartile range = (upper quartile – lower quartile), this is $(10 - 4) = 6$, and the semi-interquartile range will be $(6/2) = 3$. [Note that the IQ range and semi-IQ range are not values from the list – they represent the spread of values between the highest first quartile and the lowest fourth quartile sets].



Exercise:

11. For the following table of data showing frequencies of height ranges for mature tomato plants (cf. our market gardener example, above), use the lower half of the next page to draw a *cumulative frequency graph*; then estimate the *median* from the graph and determine the *interquartile range*.

Research Methods Workshops: S2_Frequency

<i>Height (cm)</i>	$150 \leq h < 155$	$155 \leq h < 160$	$160 \leq h < 165$	$165 \leq h < 170$	$170 \leq h < 175$
<i>Frequency</i>	4	22	56	32	5

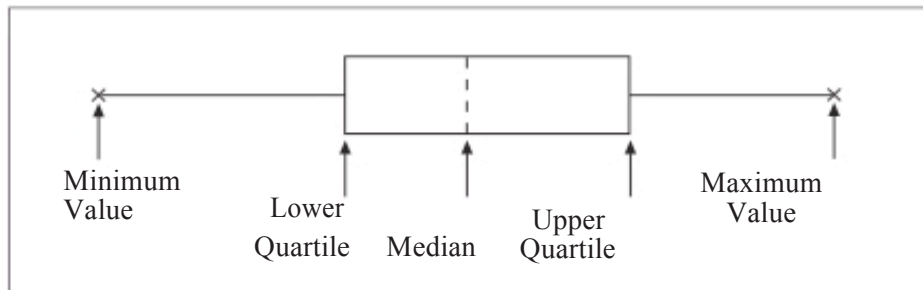
Note: To plot a cumulative frequency graph for this sample of frequency values, you will need, first, to tabulate the accumulating heights of tomato plants from 150cm (there are no heights below 150cm) up to less than 175cm – using equal intervals for each group:

<i>Height (cm)</i>	<i>Under 150</i>	<i>Under 155</i>	<i>Under 160</i>	<i>Under 165</i>	<i>Under 170</i>	<i>Under 175</i>
<i>Frequency</i>	0	$0+4 = 4$	$4+22 = 26$	$26+56 = 82$	$82+32 = 114$	$114+5 = 119$

So the y-axis of your graph (cumulative frequency) will go from 0 to 120; the x-axis (heights) will go from 150 to 175 (cms).

4 Box and Whisker Plots

A box and whisker plot is based on the minimum and maximum values, the upper and lower quartiles and the median. This type of plot provides a good way to compare two or more samples.



Note: Box and whisker plots must always be drawn accurately to scale.



Example 1

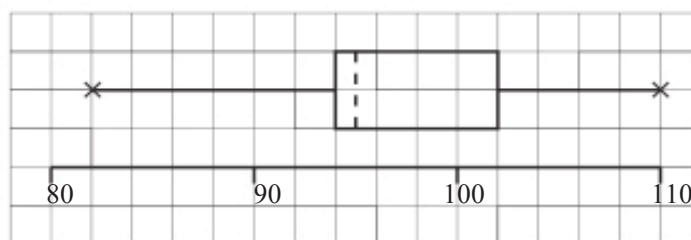
Given the information below, draw a box and whisker plot.

Minimum	82
Lower quartile	94
Median	95
Upper quartile	102
Maximum	110



Solution

The box and whisker plot is shown below.



Example 2 / Exercise

Show the lower quartile, median and upper quartile then draw a box and whisker plot for this sample:

5 7 1 9 11 22 15



Solution

First list the sample in order, to determine the median and the quartiles.

