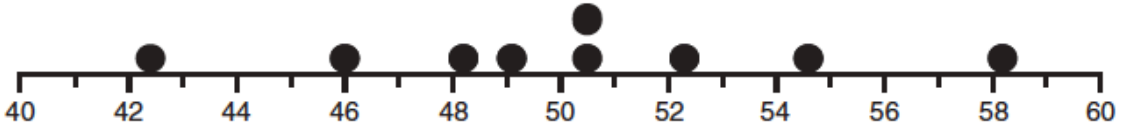
Presenting data

The Deepdown Mining Company have a long-term contract to supply coal to Bordnahome Power Station. The coal is supplied by the train-load. Nine samples are taken by Bordnahome from each train, analysed for parameters such as sulfur and ash content and tested for calorific value. We shall only concern ourselves with calorific values.

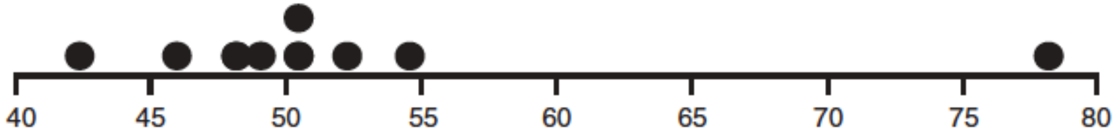
The latest train-load to arrive at the power station contains 1100 tonnes of coal and has been sampled to give the following calorific values (cal/g):

46.0 54.6 58.2 50.4 42.4 50.6 48.2 52.3 49.1

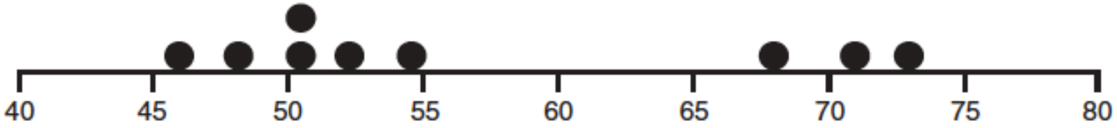
The blob chart is a simple but powerful way of expressing data. It is part of ‘descriptive statistics’ in many statistical packages (perhaps as a ‘dot plot’) but is so simple to draw. Mark out an axis, indicate the scale, and place a blob to indicate the value of each data point. The pattern of the data is easy to see.



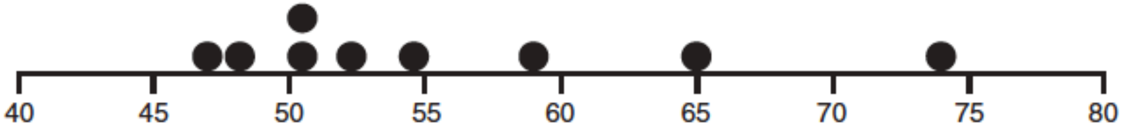
Blob diagram of the coal data



Hypothetical example containing a rogue result (outlier)



Hypothetical example containing two populations



Hypothetical example with more data on the left-hand side than on the right-hand side (skewed distribution)

Averages

There is little doubt that the coal data are straightforward compared with the hypothetical examples, and we can bear this in mind when choosing summary statistics. There are two widely used averages: the mean and the median.

The **sample mean** is obtained by summing the observations and dividing by the sample size. If we need to use a formula:

$$\bar{x} = \frac{\sum x}{n}$$

where

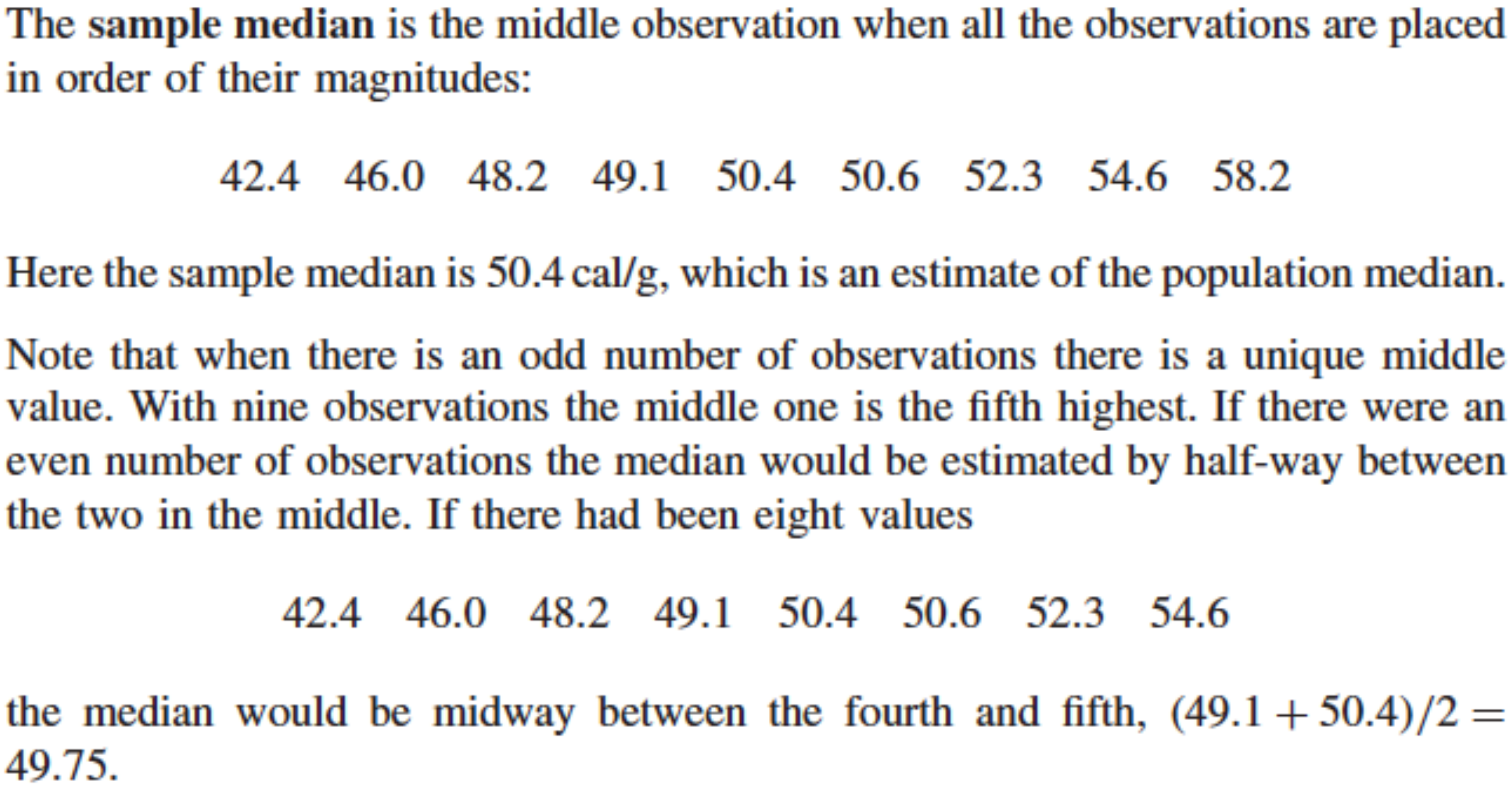
$\bar{x}$ is the sample mean,
Σx is the total of the observations, and
n is the sample size (or number of observations).

In the example

$$\sum x = 451.8, \;\; n = 9$$

so

$$\bar{x} = \frac{451.8}{9} = 50.2\,\text{cal/g}$$



The **sample median** is the middle observation when all the observations are placed in order of their magnitudes:

$$42.4 \quad 46.0 \quad 48.2 \quad 49.1 \quad 50.4 \quad 50.6 \quad 52.3 \quad 54.6 \quad 58.2$$

Here the sample median is 50.4 cal/g, which is an estimate of the population median.

Note that when there is an odd number of observations there is a unique middle value. With nine observations the middle one is the fifth highest. If there were an even number of observations the median would be estimated by half-way between the two in the middle. If there had been eight values

$$42.4 \quad 46.0 \quad 48.2 \quad 49.1 \quad 50.4 \quad 50.6 \quad 52.3 \quad 54.6$$

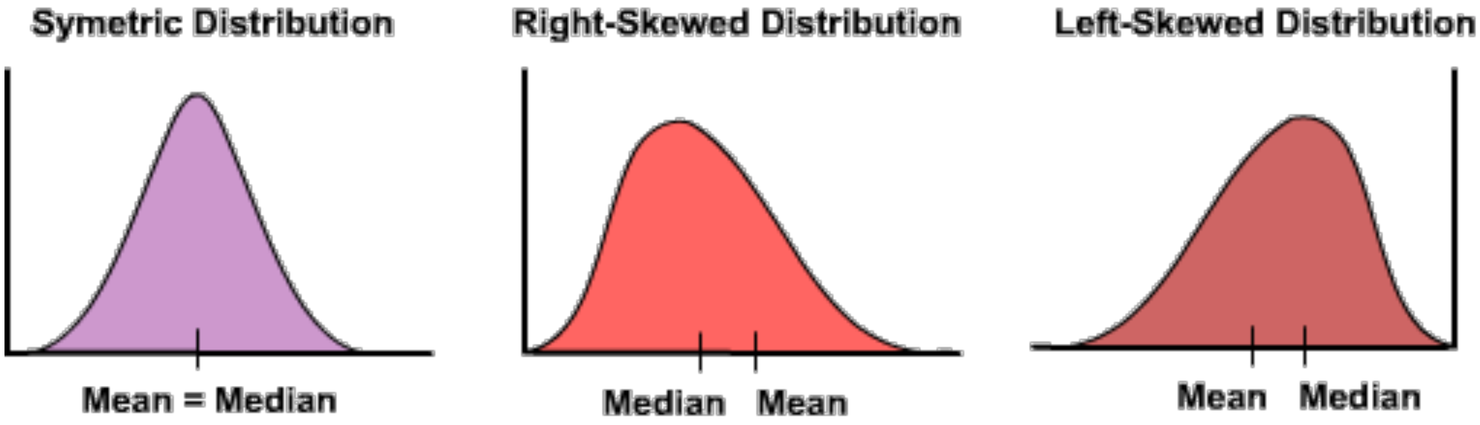
the median would be midway between the fourth and fifth, $(49.1 + 50.4)/2 = 49.75$.

Median is a robust statistics!

Should we choose the median or the mean?

Provided the distribution is reasonably symmetric (and in particular if the data follow the ‘normal’ distribution, which is introduced in the next lessons) we tend to prefer the mean, giving a better estimate because it uses the magnitudes of all the data.

If, however, the distribution is skewed, the median is often preferred since it is a better indication of the centre.



Measures of variability

There are three measures in common use: the range, standard deviation and relative standard deviation.

The **range** is the difference between the largest and smallest observations. For the coal data the range is

$$58.2 - 42.4 = 15.8\,\text{cal/g}$$

The range is easy to calculate and understand but suffers from two main disadvantages:

(i) It tends to increase with sample size, which makes it difficult to compare ranges which have arisen from different sample sizes.
A new observation added to a sample will either be within the range of the sample, leaving the range unchanged, or will be outside the range of the sample, thus increasing the range. The range cannot decrease with more data, only stay the same or increase.

(ii) It depends upon only two observations and is therefore highly affected by rogue observations.
If an outlier is present, it will either be the highest or lowest value in the sample, and will contribute to the calculation of the range.

Despite these disadvantages the range is widely used in such applications as process control in which small samples of the same size are taken regularly.

The **standard deviation** is the most commonly used measure of variability.

Calculation of standard deviation

	Observation	Deviation from Mean	$(\text{Deviation})^2$
	46.0	–4.2	17.64
	54.6	4.4	19.36
	58.2	8.0	64.00
	50.4	0.2	0.04
	42.4	–7.8	60.84
	50.6	0.4	0.16
	48.2	–2.0	4.00
	52.3	2.1	4.41
	49.1	–1.1	1.21
Total	451.8	0.0	171.66 (sum of squares)
Mean	50.2		

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

$$\begin{aligned}\text{Sample standard deviation } (s) &= \sqrt{\frac{\text{Sum of squares}}{\text{Degrees of freedom}}}\\ &= \sqrt{\frac{171.66}{8}}\\ &= \sqrt{21.4575}\\ &= 4.63\,\text{cal/g}\end{aligned}$$

The standard deviation is an average (a strange sort of average) deviation from the mean. (It is actually the 'root mean square deviation'.) Our standard deviation is 4.63 cal/g. This means that some observations will be nearer the mean than 4.63 and some will be further away.

Relative standard deviation

The **relative standard deviation** (**RSD**) represents the standard deviation expressed as a percentage of the mean. It is also known as the **coefficient of variation** (**CV**) or **%SD**.

This is easily calculated from the standard deviation (s) and the mean ($\bar{x}$):

$$\text{RSD} = \frac{100s}{\bar{x}}$$

For the coal example,

$$\text{RSD} = \frac{100 \times 4.63}{50.2} = 9.2\%$$

The relative standard deviation is popular in many industries since it is easily understood.

It has one other advantage: if a number of sets of data have widely differing means and their standard deviations are proportional to their means, the RSD is a general measure of the variability which applies at any mean level. The converse is also true: there is no advantage in using relative standard deviation if the standard deviation is not proportional to the mean.

There are also two situations in which the relative standard deviation must not be used:

(a) With a measurement scale which does not have a true zero. For example, the Celsius scale of temperature does not have a true zero, so temperature variability must always be in absolute units, represented by the standard deviation. If we used RSD we would obtain a nonsense. For example, a negative mean temperature would give a negative RSD. There are many measurements with a true zero, including those using concentration, density, weight, and speed.

(b) When the measurement scale is a proportion between 0 and 1. For example with particle counting we can express the same result as 'proportion below 10 µm' or 'proportion above 10 µm'. If we then converted the standard deviation of the proportions into RSDs we would obtain different values depending on whether we used 'below' or 'above'.

Degrees of freedom

$$\text{Sample standard deviation } (s) = \sqrt{\frac{\text{Sum of squares}}{\text{Degrees of freedom}}}$$

Degrees of freedom is the number of observations which can be varied independently under a constraint.

With nine observations we can vary eight of them but the ninth must be such as to make the deviations sum to zero. So with Bordnahome's data the mean is 50.2, the standard deviation is 4.63 and the degrees of freedom associated with the standard deviation is 8.

Calculation of standard deviation to illustrate degrees of freedom

	Observation	Deviation from Mean	$(\text{Deviation})^2$
	46.0	–4.3	18.49
	54.6	+4.3	18.49
Total	100.6	0.0	36.98
Mean	50.3		

Exploratory data analysis

Introduction

Before plunging headlong into a statistical analysis, it is essential to examine the data.

Methods for exploring data will allow us to gain a feel for the distribution of the data, for example:

(i) Are the values distributed in a symmetrical fashion about the centre?
(ii) Are they skewed, with a greater concentration at one end than at the other?
(iii) Are there extreme values present?

We have already seen the advantages of visually examining data using blob diagrams. These are extremely useful with small data sets but are not so useful with large data sets.

More useful with larger quantities of data is a **histogram** in which we can look at the shape of the distribution of data and make appropriate inferences.

An alternative to the histogram is the **box plot**, which is extremely useful when looking at trends in several data sets or, for example, when comparing several machines. The box plot gives a clear graphical display of the spread of the data, and in particular the median, the magnitude of the variability and any outliers.

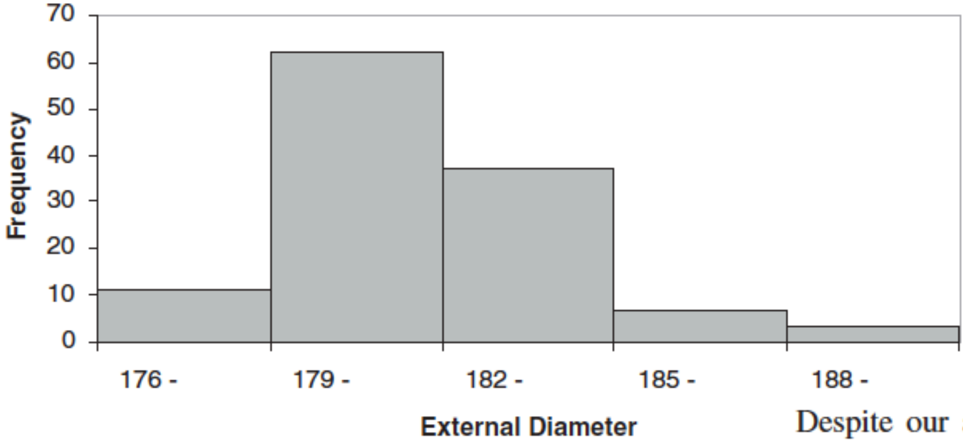
External diameters of 120 bearings (mm)

179.6	183.9	181.3	183.9	182.6	176.1
181.0	180.7	182.0	179.0	182.8	178.2
181.6	179.9	181.5	179.9	183.6	181.7
180.9	184.0	183.0	177.8	181.7	182.7
179.0	187.5	180.5	180.9	187.4	180.5
183.1	179.5	182.9	183.5	179.1	181.9
182.8	179.7	183.2	181.7	180.1	178.5
183.4	180.1	181.6	181.9	182.7	182.8
181.8	187.3	179.3	181.9	182.8	179.0
186.1	184.7	180.7	182.5	182.9	179.6
181.7	181.9	181.8	182.2	178.9	181.5
182.3	184.1	182.5	182.0	178.5	187.2
183.3	180.8	182.3	183.0	179.0	180.1
181.7	184.5	182.0	182.4	179.4	180.3
183.3	181.0	181.8	184.1	181.3	181.0
188.4	178.4	181.4	180.9	180.7	178.6
179.6	181.4	180.2	189.2	182.1	181.9
180.8	177.9	180.0	180.9	180.0	182.6
182.9	179.9	189.8	181.9	181.6	181.6
177.3	180.4	177.8	187.8	181.0	186.7

Table 3.2 Counting the frequencies

Class Interval	Tally	Frequency
176.0 to 178.9	𝍸 𝍸 |	11
179.0 to 181.9	𝍸 𝍸 𝍸 𝍸 etc.	62
182.0 to 184.9	𝍸 𝍸 𝍸 𝍸 etc.	37
185.0 to 187.9	𝍸 ||	7
188.0 to 190.9	|||	3

Table shows the tally for each interval and the frequency. They can now be shown graphically in a histogram in which each frequency is represented by the height of each bar. This is shown in Figure



Histogram of external diameter

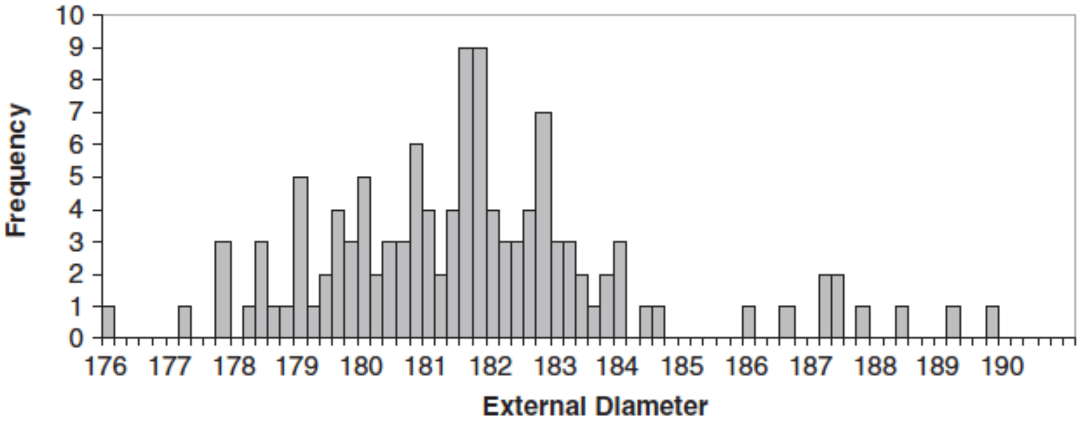
Despite our attempts, the histogram displays little information because we have used too few groups. Clearly the size of the class interval is important. An interval of 3.0 mm is too wide. Let us repeat the procedure with a class interval of 0.2 mm.

The histogram with a class interval of 0.2 mm has 70 classes. The data are very sparse, with many classes having frequencies of zero. This makes it difficult to make a judgement about the process. We need something between the class intervals of 3 and 0.2 mm.

In general, a histogram should have approximately eight classes when there are 50 data values and around fifteen when there are 100 or more data, with the class intervals being sensibly chosen as rounded numbers. As we have seen, it is not helpful to have too few or too many classes.

Frequencies with a class interval of 0.2 mm

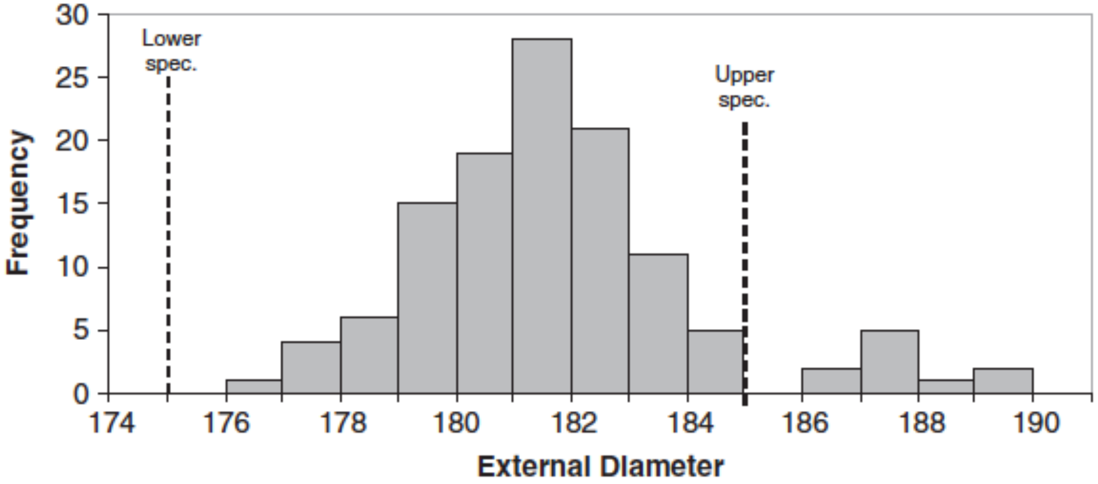
Class Interval	Frequency
176.0 to 176.1	1
176.2 to 176.3	0
176.4 to 176.5	0
176.6 to 176.7	0
176.8 to 176.9	0
177.0 to 177.1	0
177.2 to 177.3	1
177.4 to 177.5	0
177.6 to 177.7	0
177.8 to 177.9	3
etc.	



Histogram with class interval of 0.2 mm

Frequencies with a class interval of 1.0 mm

Class Interval	Frequency
176.0 to 176.9	1
177.0 to 177.9	4
178.0 to 178.9	6
179.0 to 179.9	15
180.0 to 180.9	19
181.0 to 181.9	28
182.0 to 182.9	21
183.0 to 183.9	11
184.0 to 184.9	5
185.0 to 185.9	0
186.0 to 186.9	2
187.0 to 187.9	5
188.0 to 188.9	1
189.0 to 189.9	2

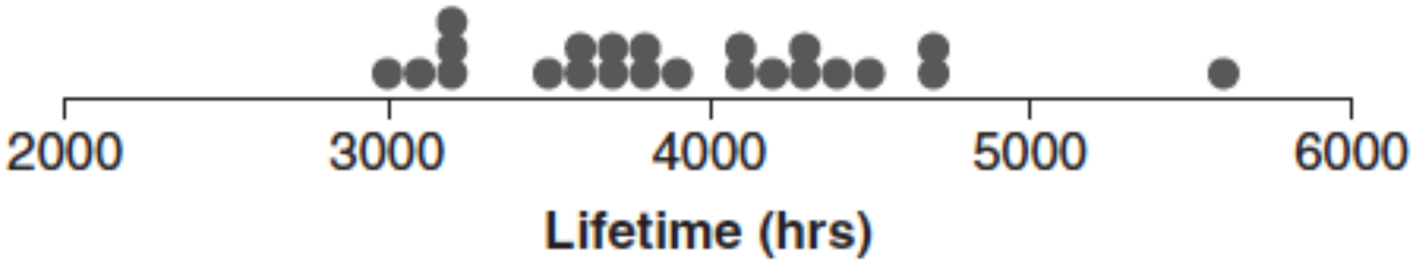


Histogram with class interval of 1.0 mm

The difficult procedure is to find the cause of the large diameters…

The histogram shows that the distribution of diameters falls into two groups. The main group is within specification of 180 ± 5 mm. The specification limits are shown on the histogram. There is, however, a small group, divorced from the main group, which is outside the upper specification limit.

Box plots: how long before the lights go out?

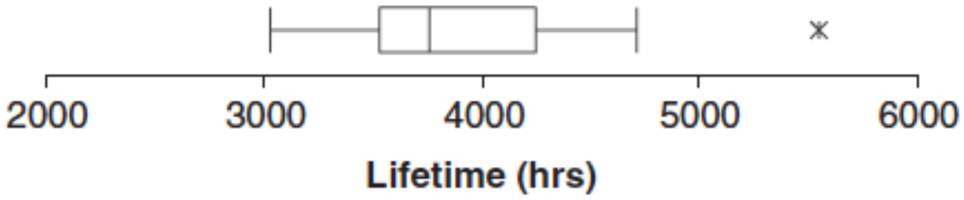


Blob diagram

environmental variables (units = …)

4710 3760 4050 3460 3690 3210 3240 3100 3750 3180 4720 3610
5550 4195 3930 4370 4050 3630 3680 3030 4540 4250 4260

The box plot is a highly graphical presentation for the distribution of a large set of data. The box plot for the Glowhite data is shown in Figure 3.5.



Box plot of lifetimes for Glowhite

environmental variables (units…)

The essential features of the box plot are as follows:

(i) The **median** – the value such that half the lifetimes are less than it, and half are greater than it.

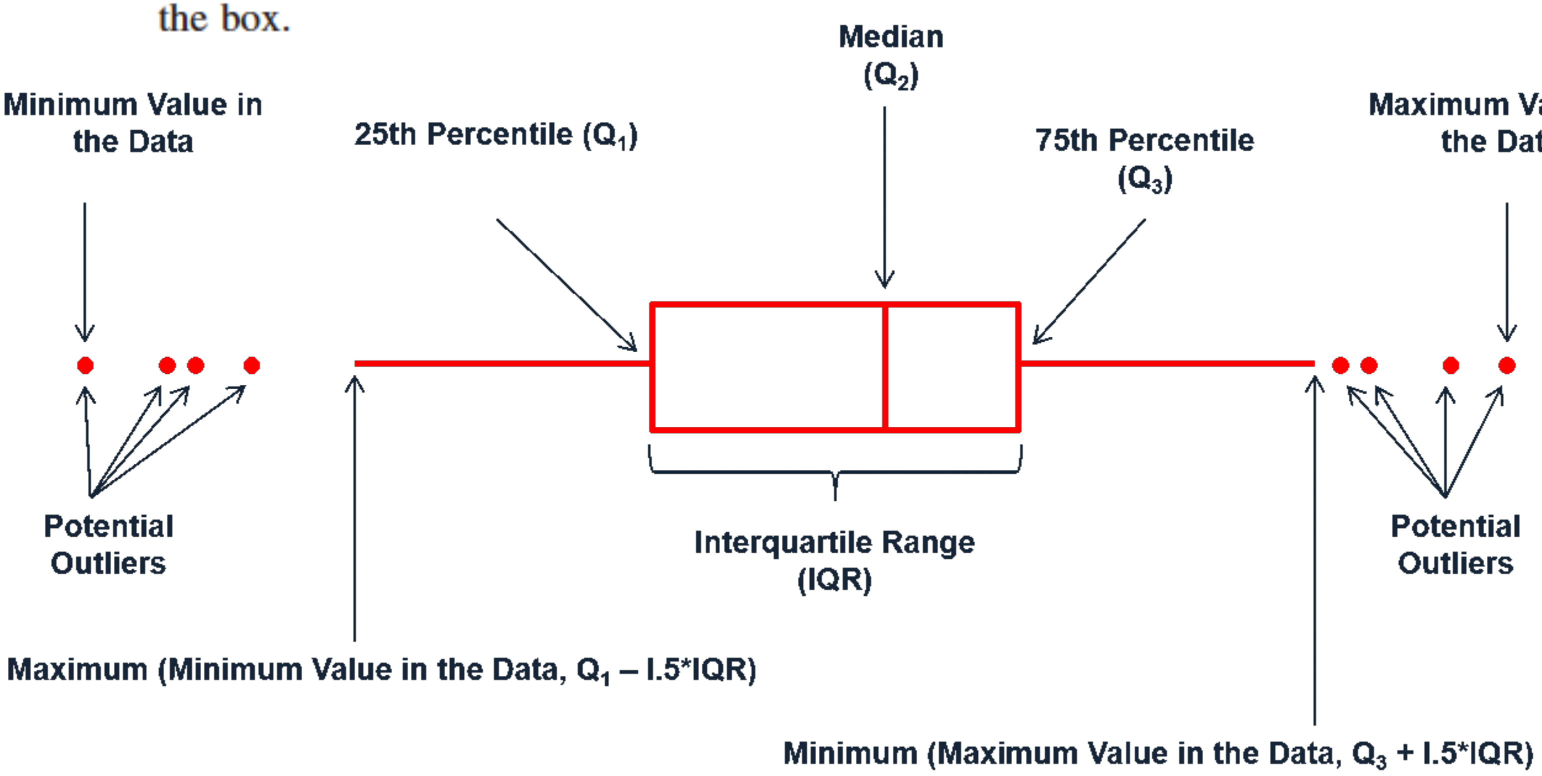
(ii) The ends of the box represent the lower and upper **quartiles** – along with the median, they divide the population into four parts with equal numbers of observations.

(iii) The box contains the middle 50% of the population.

(iv) The length of the box is the **interquartile range**, the difference between the two quartiles.

(v) The **whiskers** continue outwards to the highest and lowest values, provided they are not 'outliers'.

(vi) **Outliers**, indicated by stars, are defined in connection with the box plot as values which are more than 1 1/2 times the box length beyond either end of the box.


Median
(Q_2)
Minimum Value in the Data
25th Percentile (Q_1)
75th Percentile (Q_3)
Maximum Value in the Data
Potential Outliers
Interquartile Range (IQR)
Potential Outliers
Maximum (Minimum Value in the Data, Q_1 – I.5*IQR)
Minimum (Maximum Value in the Data, Q_3 + I.5*IQR)

rearranged in ascending order

Order	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
Observation	3030	3100	3180	3210	3240	3460	3610	3630	3680	3690	3750	3760	3930	4050	4050	4195	4250	4260	4370	4540	4710	4720	5550

Median. For a set of 23 observations, the median of the population is estimated by the middle observation, the 12th in order, 3760 hours. If there had been an even number, say 24 observations, it would have been mid-way between the 12th and 13th observations.

Lower quartile. The lower quartile is the lifetime for which there are three times as many observations above it as below it. It is estimated using the observation whose order is

$$\left(\frac{n+1}{4}\right)$$

i.e. the 6th observation, which is 3460 hours. If there had been an equal number, say 24 observations, the calculation would have been $6\,{}^{1}\!/_{4}$ and the lifetimes would have been ${}^{1}\!/_{4}$ of the distance between the 6th and 7th observations.

Upper quartile. This is obtained using order

$$3 \times \left(\frac{n+1}{4}\right)$$

i.e. the 18th observation, which is 4260 hours.

Outliers. The box length, or interquartile range, is

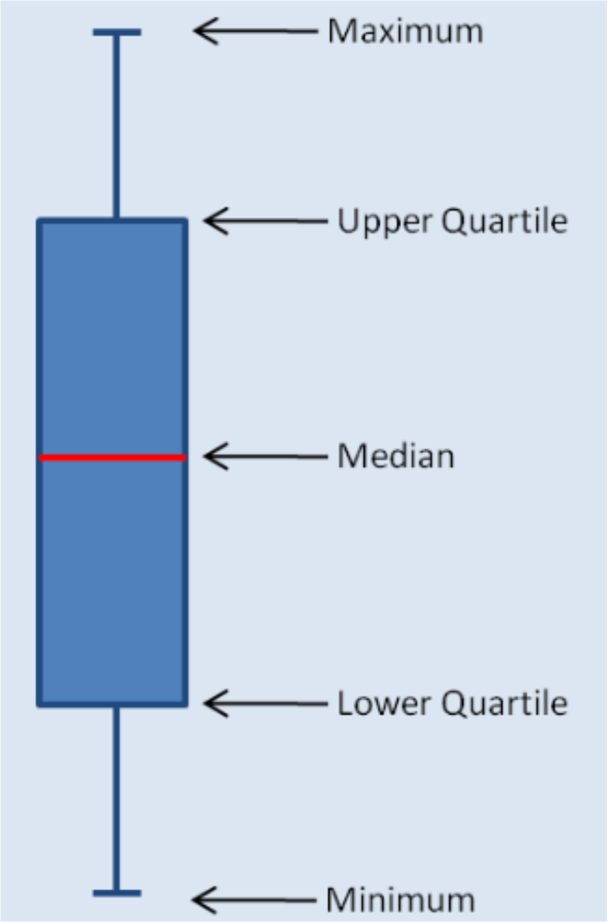
$$4260 - 3460 = 800 \text{ hours}$$

The criterion for an outlier is

$$1.5 \times 800 = 1200 \text{ hours}$$

beyond the ends of the box,

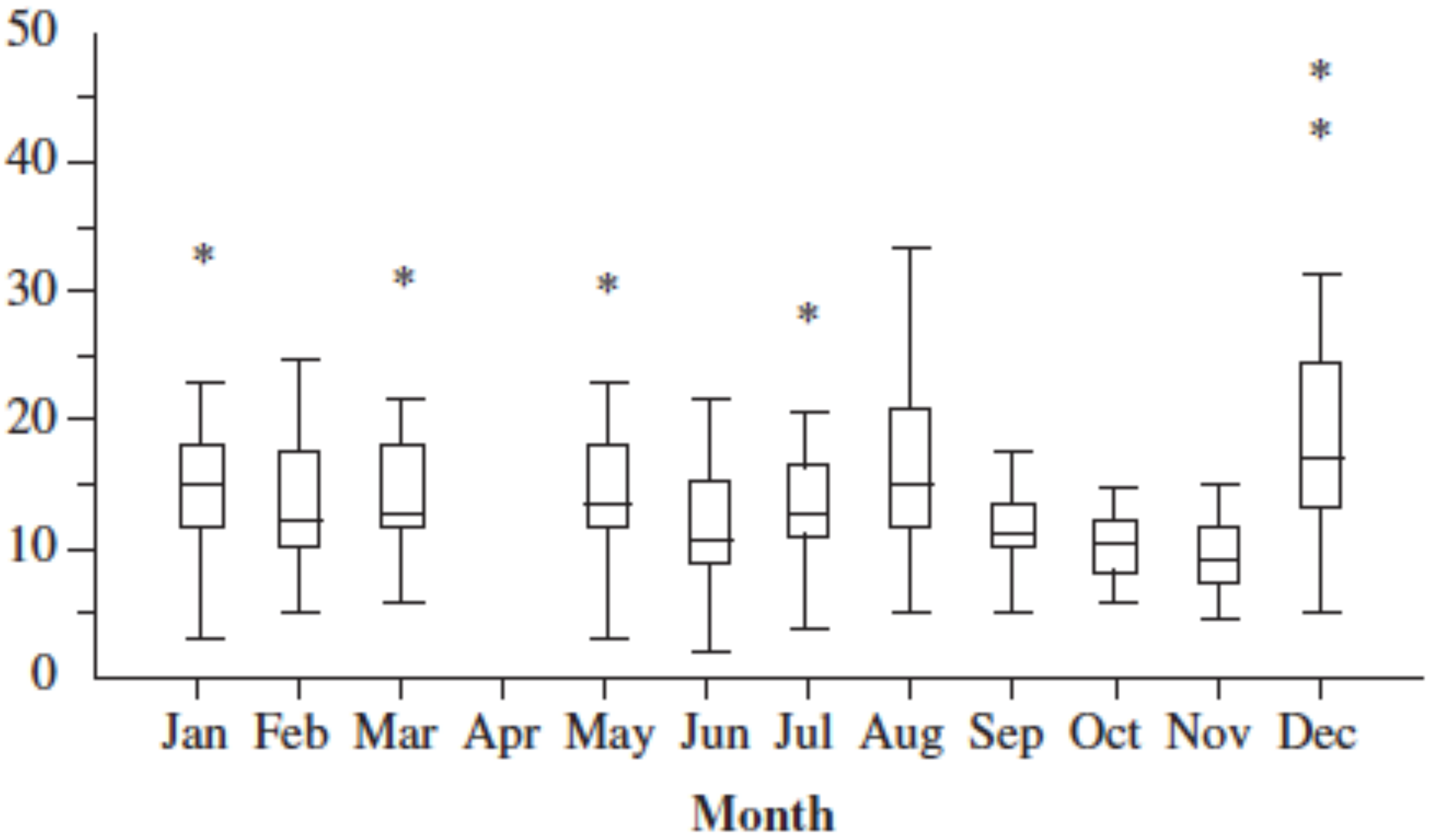
less than	$3460 - 1200 = 2260$ hours
or greater than	$4260 + 1200 = 5460$ hours



There are no outliers among the low values; the only outlier is the value of 5550 hours.

Whiskers. The whisker at the lower end goes to the lowest observation, 3030 hours.
The value of 5550 hours is an outlier, so the whisker is drawn at the highest remaining observation of 4720 hours.

The box plot in practice


Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Month