Exploratory data analysis using statistics

Dr Gavin M Abernethy

Contents

Today we will cover. . .

- Classifications of data.
- Group data and visualise with histograms.
- Analyse data using the mean, median, and mode.
- Determine the spread of data using standard deviation, variance, range, IQR.

Types of data

There are basically two kinds of data that can be collected:

- **Qualitative** - nonnumeric data such as “favourite colour”, “hairstyle”, “blood type”.
- **Quantitative** - data that can be represented by a number. For example, “height”, “number of family members”.

Quantitative data can be further classed in two subgroups:

- **Discrete** – a variable that can be counted or that has a fixed set of values. For example, the number of visitors to a park (you can’t have half or 0.2 of a person).
- **Continuous** – a variable that can be measured on a continuous scale. For example, “temperature” or “height”.

Grouping data

We can begin to analyse a small data set simply by **ordering** the data from smallest to largest.

Next, if there are only a few possible values it may be sensible to create a frequency table (counting how often each value occurs).

However, if there are many possible values that the data can take, it may be more helpful to group the data into classes (also called groups or "bins").

Unless specified, it is up to us to decide how many classes to use and how wide they should be.

Example: Grouping data

The number of hours worked per week by employees at a factory:

45 31 46 25 57 40 59 11 38 38 22 33 39
57 37 43 51 33 44 47 42 46 66 34 50

Create a tally table, and determine the frequency of each class:

Overtime Hours	Tally Marks	Frequency
10-19		
20-29		
30-39		
40-49		
50-59		
60-69		

Grouping data

The number of hours worked per week by employees at a factory:

45 31 46 25 57 40 59 11 38 38 22 33 39
57 37 43 51 33 44 47 42 46 66 34 50

Create a tally table, and determine the frequency of each class:

Overtime Hours	Tally Marks	Frequency
10-19	|	1
20-29	||	2
30-39	𝍸 |||	8
40-49	𝍸 |||	8
50-59	𝍸	5
60-69	|	1

Visualising frequency distributions using histograms

A histogram is a graphical representation of a frequency distribution, with vertical rectangular blocks such that:

- The centre of the base indicates the central value of the class.
- The area represents the class frequency.

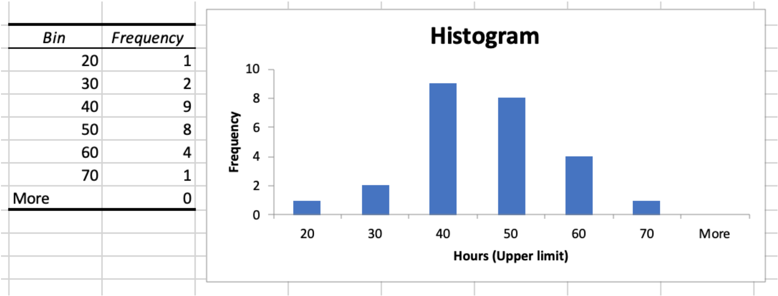
The height for each class is determined by:

$$\text{frequency distribution} = \frac{\text{frequency}}{\text{width}}$$

If each class width is **regular** (they all have the same width) then the frequency is often denoted by the height. This is the most common form of a histogram.

Example: histogram

Using Excel's built-in *Histogram* function, we can specify the upper limits of bins (another name for groups/classes) and obtain the following for the previous example:



Bin	*Frequency*
20	1
30	2
40	9
50	8
60	4
70	1
More	0

Measures of centrality

We can calculate some properties to give us an indication of where "most" or the middle of the data lies.

- Mean - the average.
- Median - the middle data value.
- Mode - the most common single data value.

Arithmetic mean

The arithmetic mean is defined as the sum of the data $\{x_i\}$ divided by the number N of data points:

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

Usually $\bar{x}$ denotes the mean of a sample, while μ ("mu") denotes the true mean of the whole population.

This is usually what we refer to as the "average". However, it can be skewed if there are extremely large or small data values present that are not typical of the rest of the set. In such cases, a more useful indicator of a "typical" value is the median...

Median

The median is the value that separates an ordered list into two. For an odd number of values this is the middle number, for example:

$$1, 4, \underbrace{5}_{\text{median}}, 8, 10$$

If there are an even number of values, the median is the halfway point between the middle two numbers, for example:

$$1, 4, \underbrace{5, 7}_{\text{middle numbers}}, 8, 10$$

The median in this case is then

$$\frac{5+7}{2} = 6$$

Example

Calculate the arithmetic mean and median of this data set:

29 32 26 40 12 20 35 190

Example

Calculate the arithmetic mean and median of this data set:

29 32 26 40 12 20 35 190

The arithmetic mean is:

$$\frac{12+20+26+29+32+35+40+190}{8} = 48$$

Example

Calculate the arithmetic mean and median of this data set:

29 32 26 40 12 20 35 190

The arithmetic mean is:

$$\frac{12 + 20 + 26 + 29 + 32 + 35 + 40 + 190}{8} = 48$$

To calculate the median we first order the data:

12 20 26 29 32 35 40 190

The median is then:

$$\frac{29 + 32}{2} = 30.5$$

Example: mean for grouped data

When using grouped data, we can calculate the mean by multiplying the midpoints and frequencies of each group, sum them, then divide by the total size of the sample.

Example:

Class	Frequency f	Midpoint x_i	$f \cdot x_i$
15-20	2	17.5	35
20-25	1	22.5	22.5
25-30	3	27.5	82.5
30-35	2	32.5	65
35-40	1	37.5	37.5
40-45	1	42.5	42.5

The mean is then:

$$\frac{\sum f \cdot x_i}{\sum f} = \frac{1}{10} \cdot 285 = 28.5$$

Example: median for grouped data

When using grouped data, we can calculate the **cumulative frequency** to help us locate the median.

Example:

18.3 20.2 24.0 26.9 27.1 28.0 32.4 34.0 39.3 41.7

Class	Frequency	Cumulative freq.
15-20	2	2
20-25	1	3
25-30	3	6
30-35	2	8
35-40	1	9
40-45	1	10

Example: median for grouped data

When using grouped data, we can calculate the **cumulative frequency** to help us locate the median.

Example:

18.3 20.2 24.0 26.9 27.1 28.0 32.4 34.0 39.3 41.7

Class	Frequency	Cumulative freq.
15-20	2	2
20-25	1	3
25-30	3	6
30-35	2	8
35-40	1	9
40-45	1	10

The median is then the $\frac{10+1}{2} = 5.5^{th}$ value, which is 27.5.

Mode

The mode is the value of the sample that occurs the most.

This may be useful for, say, a discrete set of integer data.

Example:

The age (in years) of people working in an office:

$$18 \quad 19 \quad 19 \quad 22 \quad 27 \quad 33 \quad 40$$

The mode is then 19.

Example: modal group

With a continuous variable it is possible that no data values repeat. Here, we can use a tally chart and find the **modal group**.

Example:

18.3 20.2 24.0 26.9 27.1 28.0 32.4 34.0 39.3 41.7

Class	Frequency
15-20	2
20-25	1
25-30	3
30-35	2
35-40	1
40-45	1

Example: modal group

With a continuous variable it is possible that no data values repeat. Here, we can use a tally chart and find the **modal group**.

Example:

18.3 20.2 24.0 26.9 27.1 28.0 32.4 34.0 39.3 41.7

Class	Frequency
15-20	2
20-25	1
25-30	3
30-35	2
35-40	1
40-45	1

The modal group is then 25-30.

Dispersion or Spread of Data

We often want to know how spread a given data set is so that we know how far values stray from the mean. When repeating experimental measurements, the spread of data allows us to estimate the error in our measurement.

We will consider the following:

- Range
- Interquartile range - the range spanned by the central 50% of the data.
- Standard deviation and variance - are points close to the mean, or far from it?

Range and Quartiles

The range of the data is simply the difference between the largest and smallest values.

Range

$$\text{Range} = \text{maximum value} - \text{minimum value}$$

Quartiles and IQR

Data can also be characterised by the upper and lower quartiles. Arrange the data values in increasing order, then . . .

Quartiles

The lower quartile (denoted L_{25} or Q_1) is the median of the lower half of the data.

The upper quartile (U_{25} or Q_3) is the median of the upper half.

Range and Quartiles

There are actually several competing methods for calculating the upper and lower quartiles, and statisticians are *not* agreed on any one method to use!

If it is not clear how to divide the data into an “upper” and “lower” half (e.g. if there is an odd number of values, do you include the median in either half, or not?), one method is the following:

If there are N values, the lower quartile is located at the $\frac{1}{4}(N+1)^{th}$ value and the upper quartile is the $\frac{3}{4}(N+1)^{th}$ value.

In general, the exact method used should not have a large impact on large data sets. We will typically use EXCEL’s built-in `=QUARTILE` function when calculating the IQR.

Quartiles and IQR

The difference between the upper and lower quartiles is the **interquartile range** (IQR):

IQR

$$IQR = U_{25} - L_{25}$$

Example: Range and Quartiles

Example:

1 1 2 3 3 3 4 5 5 7 7 8 10 19

The range is the largest value minus the lowest value:

$$19 - 1 = 18$$

There are 14 data points, so the median of the first 7 is the 4th value. Thus,

$$L_{25} = 3$$

The median of points 8-14 is the 11th value. Thus,

$$U_{25} = 7$$

And so we have:

$$IQR = U_{25} - L_{25} = 7 - 3 = 4$$

Variance

The variance (also known as “mean squared deviation” or “mean squared displacement”) is the average of the square of the deviation of each data point from the arithmetic mean. There are two variations depending on the data set we are working with:

Variance

Variance if the data set is the entire population:

$$\sigma^2 \; = \; \overline{\Delta x^2} \; = \; \frac{1}{N}\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^2$$

Variance of the population based on a sample:

$$\sigma^2 \; = \; \overline{\Delta x^2} \; = \; \frac{1}{N-1}\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^2$$

Variance

Why do we have to square the deviations from the mean?

If we did not do so, the deviations *above* and *below* the mean would cancel out, and so:

$$\overline{x_i - \bar{x}} \;=\; \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x}) = 0$$

would be true for any data set.

Expand the formula above and remember the definition of the mean to see why this must be true.

Standard Deviation

The standard deviation is then the square root of the variance, so it also depends on whether we have the population or just a sample:

Standard deviation

S.D. if the data set is the entire population:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\bar{x} - x_i)^2}$$

S.D. of the population based on a sample:

$$\sigma = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(\bar{x} - x_i)^2}$$

σ has the same units as x_i or $\bar{x}$.

Standard Error

Standard error is most often used when discussing experimental data where repeated measurements are taken.

It is defined as:

Standard error

$$SE_{\bar{x}} = \frac{\sigma}{\sqrt{N}}$$

where N is the number of measurements taken.

This is essentially an estimate of the confidence of how accurately we have measured the mean.

Standard deviation versus the standard error

Let's say we are measuring the temperature of an object. Even though the *average* temperature in the room remains constant over a long time, there are constant random fluctuations that lead to small changes in the temperature.

Even if we take measurements until the end of time, the standard deviation will remain the same, because the size of the fluctuations is (on average) the same.

However, our estimate of the mean becomes more and more accurate and our "confidence" in our measurement of the temperature of the bar improves, so the *error* is reduced.