MEI Insights 3:

Statistics in Mathematics A level for Teaching from 2017

by **Stella Dudzic, Stephen Lee and Charlie Stripp**

Introduction

Mathematics in Education and Industry (MEI), through its charitable status, is committed to improving mathematics education and promotes teaching and learning through numerous strands of activity. We work to have a positive influence on national mathematics education policy and one aspect of this is through our support for curriculum development.

There will be new A levels in Mathematics and Further Mathematics for first teaching from 2017. As well as pure mathematics, both mechanics and statistics will be included in A level Mathematics; there will be no choice of content.

This, our third article in a series entitled 'MEI Insights', aims to help teachers to start thinking about the statistics in the new A level Mathematics and how they might teach it. The article includes information about the content*, links between different aspects of the content and some example ideas for teaching.

(*Please note that due to publication schedules this article was written in July 2015. Further information on the new A levels may have emerged between submission and actual publication of this article.)

Some Features of Statistics in the new Mathematics A Levels

The content of AS and A level Mathematics for teaching from 2017 is available online (DfE, 2014). Some of the features of this content which are dealt with in this article are listed below.

- The use of technology, such as mathematical and statistical graphing tools and spreadsheets, is expected throughout the study of the new AS and A level.
- Students are expected to have calculators which will work out summary statistics from data and also probabilities from statistical distributions.
- Students are expected to become familiar with at least one large data set before the assessment, using technology to investigate questions arising from real contexts.
- The statistics content includes the binomial and normal distributions and hypothesis testing.

Learning with Real Data

Starting to teach with real data, rather than starting with teaching techniques and considering applications later, helps students see the relevance of statistics and helps them understand how to interpret data. A few example contexts are considered below.

As we grow older, we become more interested in our blood pressure – this is a topic that is probably less directly relevant to most 16-year-olds, but they may have relatives who take tablets for high blood pressure.

There are a number of risk factors for developing high blood pressure, including age. The American National Health and Nutrition Examination Survey (NHANES) contains a wealth of health related data. Probably the easiest way for teachers to download a sample of the data is through the EEPS data zoo (see references). The data and scatter diagram (Table 1, Fig. 1) are for a random sample of 10 adults from NHANES. For students to be able to draw a scatter diagram by hand without it taking too long, sample sizes need to be small. Students will already have seen scatter diagrams like this at GCSE, and KS3, and will be able to see that there is positive correlation.

Table 1

Age (years)	68	83	75	41	38	17	17	32	19	79
Systolic blood pressure (mm Hg)	138	139	132	122	119	104	97	121	124	160

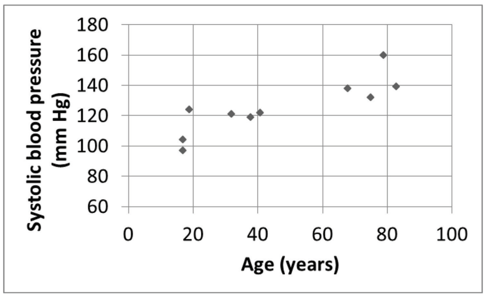


Fig. 1 Scatter diagram for Table 1

Most students are likely to be unfamiliar with the term 'systolic blood pressure' and the units it is measured in. Working with real data, like this, during the A level course will allow students to understand the wide applicability of statistical techniques.

It would be a nuisance to draw a scatter diagram by hand for a larger sample but it is straightforward with a spreadsheet. Here are two scatter diagrams (Fig. 2; Fig. 3) for a random sample of 250 adults, one showing their systolic blood pressure, the other their diastolic blood pressure, each plotted against their age.

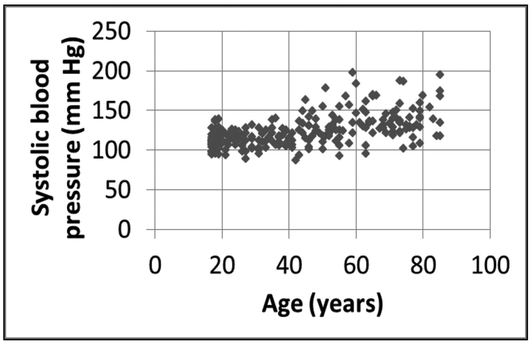


Fig. 2 Scatter diagram for the systolic blood pressure for 250 adults

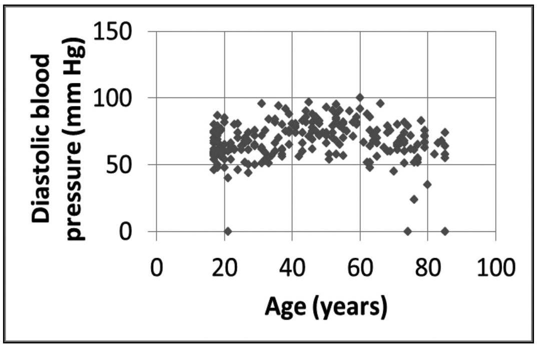


Fig. 3 Scatter diagram for the diastolic blood pressure for 250 adults

What do you notice from the scatter diagrams for the sample of size 250?

Here are some teaching points which arise from use of this real data, these would not occur naturally for students working with small artificial data sets to prepare for an examination.

- Real data sets often have missing values or errors – it is important to be able to deal with these. Unusual data values are sometimes called outliers and they need to be investigated to check whether they are real data values or errors.
- It can be difficult to tell whether there is any correlation or not – a measure of correlation is needed; the correlation coefficient measures how close the data points are to a straight line.
- There are two measurements of blood pressure taken; only one of them tends to increase with age; this can be reported on some medical websites as "blood pressure tends to increase with age". Using the correct data is important when investigating a statistical hypothesis.

Binomial and Normal distributions as Models

The histogram in Figure 4 shows the heights of a sample of 122 women from NHANES. A normal distribution with the same mean and standard deviation as the data set is also shown. The normal distribution seems to fit these data fairly well.

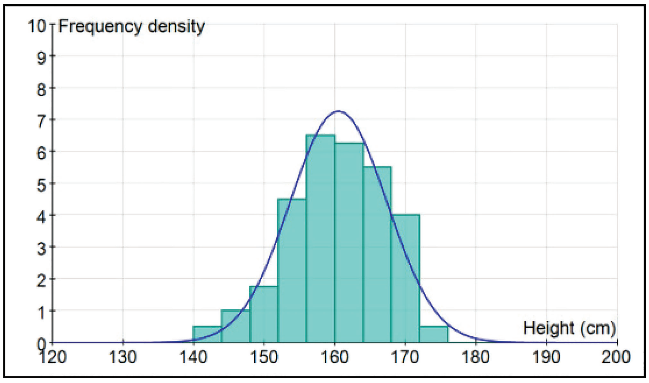


Fig. 4 A histogram showing the heights of a sample of 122 women

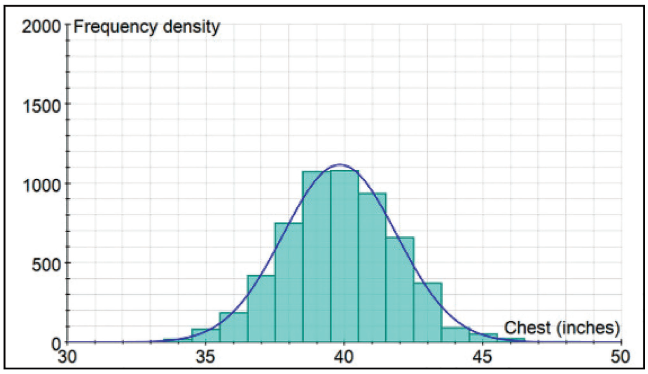


Fig. 5 A histogram showing Quetelet's data set of chest measurements of 5738 Scottish militiamen

The histogram in Figure 5 shows Quetelet's data set of chest measurements of 5738 Scottish militiamen; the normal distribution seems to fit these data well too.

However, looking at other real data such as blood pressure or weight will often reveal that the data are skewed – symmetry in data sets is not as common as we might suppose. As well as possible lack of symmetry, the detailed shapes of the 'tails' of data sets also quite often depart from those of a normal distribution.

Nevertheless, the normal distribution is a common and useful model for data and so the new A level includes it. An intuitive understanding of the normal distribution enables us to make predictions about what we expect to happen. For example, we might reasonably expect that school value added data are normally distributed and we would not be surprised if half the schools were below average. These data are freely available online and can be downloaded as a spreadsheet.

The graph in Figure 6 shows the KS4 data for England for 2014 – can you explain what you see?

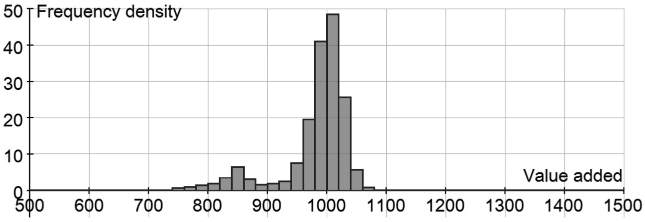


Fig. 6 Histogram showing the value added scores for KS4 data for England in 2014

The value added scores are based on results in the best 8 GCSEs; the graph shows a bimodal distribution; it looks as though it could be composed of two overlapping normal distributions – the schools in the smaller 'hump' are nearly all special schools with small numbers of students taking GCSEs.

Another commonly used distribution is the binomial distribution. Like the normal distribution, it can be used to model real data (Dudzic, 2012) but it can also be derived using mathematics which students already know from GCSE and it connects to the binomial expansion in pure mathematics.

Imagine a 10-question true/false quiz where people might just guess the answers. It is easy to see that, on average, they would get 5 questions right just by guessing. Working out the probabilities of getting different numbers of questions correct can be built up from simpler cases using tree diagrams (Fig. 7).

For a 2-question test, the probabilities are shown in Table 2. Once students have worked out the probabilities for 3 and 4 questions, they should be able to use the pattern to obtain further probabilities. Once they understand the pattern, that is a suitable time to introduce the usual formula for binomial probabilities and explore the connections with the binomial expansion.

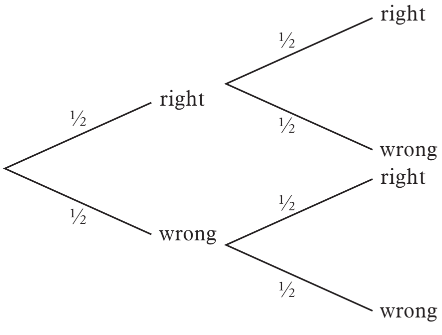


Fig. 7 Tree diagram for a simple true/false quiz

Table 2

No. correct	0	1	2
Probability	$\left(\frac{1}{2}\right)^2$	$2\times\left(\frac{1}{2}\right)^2$	$\left(\frac{1}{2}\right)^2$

The probabilities can also be displayed using software such as *Geogebra* or *Autograph*.

Introducing Hypothesis Testing

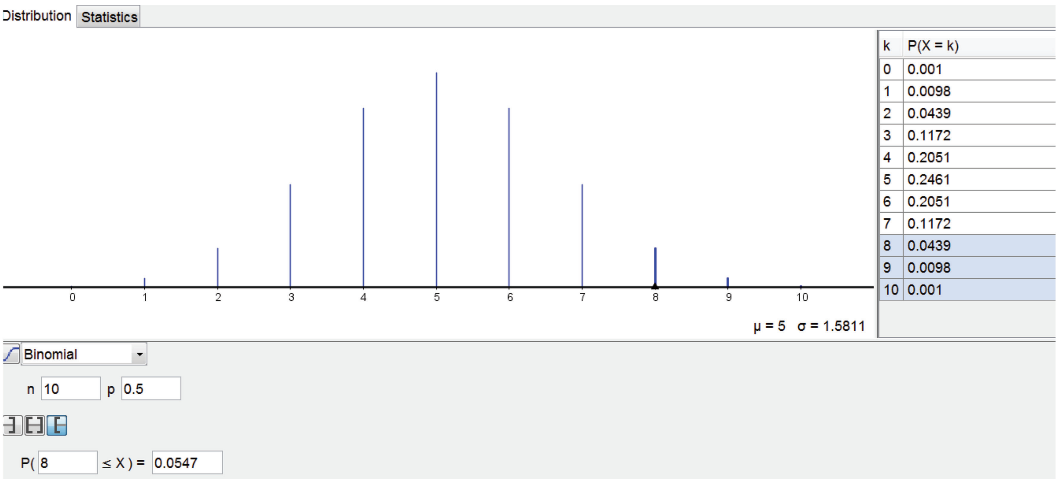
Imagine that a large number of people have applied for a well-paid job. The recruiters want to get the right person so they decide that the first stage in recruitment will be to administer a simple test to all applicants. The test will have 10 true/false questions; the questions will be chosen at random from a large bank of questions which reflect the knowledge needed for the job. The recruiters have devised special software to administer the test; the software requires immediate answers so there is no time to find out an unknown answer – but applicants could guess. Where should they set the pass mark to be fairly sure that people who guess all the answers do not pass the test?

Assuming that an applicant guesses all the time leads to the probabilities shown in Figure 8.

If the pass mark was set at 8, those who got 8 or more correct would pass the test. For those guessing, the software shows that this has a probability of nearly 5.5% so, on average about 5.5% of those guessing would pass the test. The organizers might decide that this is too many and choose to have a pass mark of 9 instead. Having a pass mark of 10 would be even more sure to weed out the guessers but would also eliminate some candidates who know nearly everything. Let's assume the recruiters choose 8 as the pass mark.

Hypothesis testing is about using a sample to make a decision about a population. In the example above, a sample of 10 questions was used to make a judgement about the knowledge which an applicant possesses.

The default belief, or *null hypothesis*, is that the applicant is guessing (probability of correct answer = ½). The *alternative hypothesis* – which will only be accepted if there is enough evidence – is that the applicant is doing better than guessing (probability of correct answer >½). The pass mark or *critical value* is 8. Some candidates will score 6 or 7 marks, so they are doing better than the average guesser but there is not enough evidence that they aren't just lucky in their guessing.



k	P(X = k)
0	0.001
1	0.0098
2	0.0439
3	0.1172
4	0.2051
5	0.2461
6	0.2051
7	0.1172
8	0.0439
9	0.0098
10	0.001

μ = 5 σ = 1.5811

Binomial

n 10 p 0.5

P(8 ≤ X) = 0.0547

Fig. 8 Probabilities for correct guesses to 10 true/false questions

Many students who take A level Mathematics will go on to study subjects and work in professions which make use of hypothesis testing – they will often use software to conduct hypothesis tests themselves or read about tests conducted by others. The new A level Mathematics offers students the opportunity to develop understanding of how hypothesis testing works and what it is for before going to university.

Learning hypothesis testing using a binomial distribution allows students to develop understanding of what the process involves – there are examples of its use which students can understand such as historically for Zener tests of ESP, and currently for triangle taste tests. Once students understand the process of hypothesis testing using the binomial distribution, they should be able to understand different hypothesis tests more readily.

Extending Hypothesis Testing

Look back at the first scatter diagram in this article showing age and systolic blood pressure for a sample of 10 people. It is fairly clear that there is some correlation in the sample but we really want to know whether there is any correlation in the population. For small samples, there is often some correlation in the sample, even if there is no correlation in the population that the sample data came from (you can demonstrate this by selecting small samples from a large data set with no correlation and seeing how the correlation of the samples varies).

A larger sample would be less likely to give a correlation that is unrepresentative of that in the populuation but this isn't always possible. Suppose the sample of size 10 is the only information available. A hypothesis test can be used to decide whether the sample provides evidence of correlation in the population. The null hypothesis would be that there is no correlation in the population – we would only stop believing this if the evidence was strong enough. In this case, we want to know whether there is positive correlation in the population so we need to know how big the measure of correlation, the correlation coefficient, would need to be for the sample in order to convince us that there is positive correlation in the population. From statistical tables, in this case the critical value is 0.5494. There would only be a 5% chance of getting a correlation coefficient as high as (or higher than) 0.5954 for a sample of 10 items from a population with no correlation. Using a spreadsheet, or statistical functions on a calculator, the correlation coefficient for the sample is 0.85627; this is more than the critical value so the alternative hypothesis is accepted; there is evidence of positive correlation in the population between age and systolic blood pressure. Hypothesis testing cannot give us certainty – a high number correct in a multiple choice test or a high correlation in a sample can happen purely by chance.

Want to learn more about using data for statistical insight?

Professor Chris Wild has made the videos from the Data to Insight MOOC available on YouTube. Links to sources of data sets are available on the MEI website **http://www.mei.org.uk/data-sets**

References and Links

DfE 2014 *Mathematics AS and A level Content*.

Dudzic, S. 2012 'Would you like to take a Later Flight?', *Mathematics in School*, **41**, 3 pp. 33–35.

Information about the NHANES survey can be found at: **www.cdc.gov/nchs/nhanes.htm**

The EEPS data zoo can be found at: **www.eeps.com/zoo/**

Quetelet's data is available on DASL (Data and Story Library): **http://lib.stat.cmu.edu/DASL/DataArchive.html**

School and college performance tables can be found at: **www.education.gov.uk/schools/performance/**

A selection of true/false quizzes where students might just guess can be found at: **http://reverent.org/quizzes.html**

Wild about Statistics YouTube channel: **www.youtube.com/channel/UCElKp33-h_Yw0o8XATHlICg** .

Keywords: A level; Statistics; Real data.

Authors MEI, Monckton House, Epsom Centre, White Horse
Business Park, Trowbridge, Wiltshire BA14 0XG.