Brief notes on statistics: Part 1 Histograms, averages, measures of spread, probability and the normal distribution

Michael Wood (Michael.wood@port.ac.uk)

11 March 2013

Introduction and links to electronic versions of this document and the other parts at http://woodm.myweb.port.ac.uk/stats. The data in the tables, and the figures, are in the spreadsheet, http://woodm.myweb.port.ac.uk/stats/StatNotes.xls . For a rough, but still useful, understanding, you can ignore the numbered endnotes, which provide extra detail.

Analyzing data and seeing patterns

On a simple level the purpose of a statistical analysis is normally to look at some **data** and see if there are any **patterns** which might be interesting or useful. Does data on defects or errors suggest anything about the causes of defects or errors and how to prevent them? Does health data suggest any methods of reducing a patient's risk of heart disease. Does the data suggest that men get paid more than women? Does risk information suggest a financial portfolio is in difficulty? And so on.

The first thing to check is always the **source of the data**. Is the sample likely to give a reasonably accurate indication of the general situation? Or is it likely to be biased or inaccurate in some way? You will need to check ideas about sampling methods here – which you will find in most textbooks on statistics or research methods (and a brief note in http://woodm.myweb.port.ac.uk/rm/u3critic.pdf). In practice, a **random sample** is usually ideal because each member of the population has the same chance of being selected so there should no systematic bias.

Then, it is often possible to use **diagrams or tables** to show any patterns. There are many possibilities, most of them very obvious. You should be familiar with bar charts, pie charts, line graphs, tables of various kinds, and so on. There are many books which describe some of the main possibilities – e.g. Saunders et al (2003, pp. 338-351), Wood (2003, chapter 3). If in doubt, use your common sense, or ask a friend for feedback – if your friend can't understand what your diagram or table represents, then it obviously needs clarifying. There are a few more notes on tables and diagrams in http://woodm.myweb.port.ac.uk/stats/StatNotes0.ppt.

Some of the things you can do are a little more technical. I'll start with how you can analyze a number variable.

Analyzing one number variable

A *variable* is something that varies between different *units of analysis* or *cases*. For example, in Table 1 we have the earnings of a group of 50 people. Here the **units of analysis**, or **cases**, are the individual people, and the **variable** is how much they earn.

Table 1: Earnings of a sample of 50 people (in thousands of pounds per year)

98	60	8	21	49
19	21	6	16	11
35	22	25	16	15
69	2	17	16	35
17	18	106	76	27
24	9	22	24	17
16	40	156	159	26
3	42	42	29	54
39	60	26	15	14
110	53	11	27	39

Earnings is a **number variable** for obvious reasons. Other variables – e.g. sex, or hair colour – are **category variables** because the possibilities are separate categories rather than a number (e.g. Sex in Table 1 in Part 3). The cases may not be people – they may, for example, be bank accounts, in which case a number variable might be the balance in the account and a category variable might be the type of account. We'll focus on number variables here because you can do a bit more with them.

It is difficult to see the pattern in Table 1. There are a number of ways of analyzing this information to **clarify the pattern** …

Histograms

These are **simply bar charts showing the frequency of numbers in different ranges** (known as "class intervals"). They are very useful for clarifying patterns and informally analyzing the data – see the examples attached to the hard copy of this document, and Figure 1 which shows the pattern of earnings in Table 1.

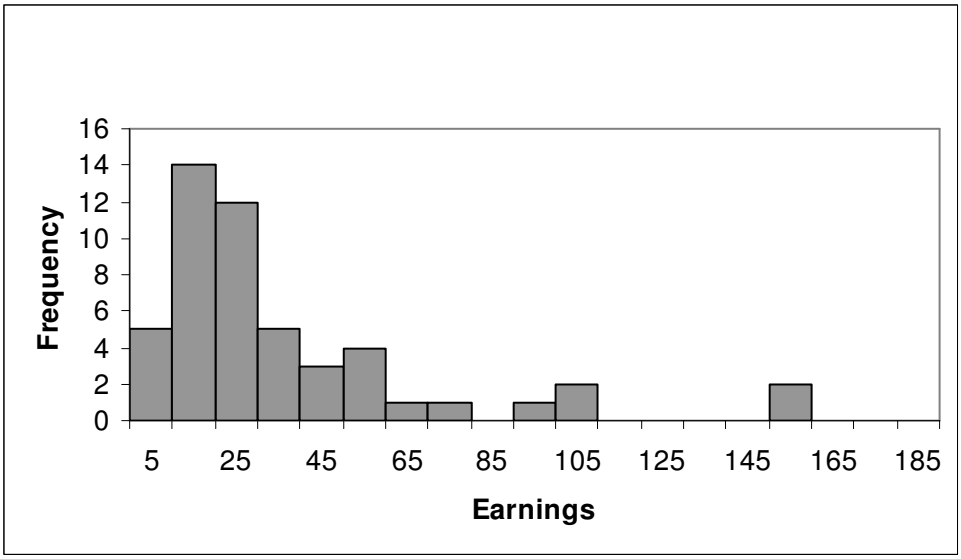
To draw a histogram, start by producing a frequency[1] table as in Table 2. To do this you need to decide on suitable class intervals. I have chosen 1-10, 11-20, etc but you could use 1-5, 6-10, 11-15, etc instead. It is important to make the **width of all class intervals the same** – e.g. don't use 0-20, 21-30, 31-35 … – otherwise a comparison of the heights of the bars will be misleading. (Sometimes it may be possible to have a single number in each class intervals – e.g. the data in the Aerosol sheet of http://woodm.myweb.port.ac.uk/stats/StatNotes.xls.)

Table 2: Beginning of frequency table based on Table 1

Class intervals	Frequency
1-10	5
11-20	14
21-30	12
and so on …	

And then draw a bar chart. Notice that each interval is labeled with the middle of the class interval: 5, 15, etc.

Figure 1: Histogram based on Tables 1 and 2 (doing this with Excel is not as easy as you might think, so if necessary use a paper and pencil for histograms) [2]



Quick question 1. Which part of a football match do you think goals are most likely to be scored? The beginning, the middle, the end, or when?

Now look at the histograms at http://woodm.myweb.port.ac.uk/stats/Histograms.pdf (apologies for the very poor quality of these copies – I will post better copies soon). The third histogram gives an answer based on real data – most people are not aware of this pattern without collecting data systematically and displaying it like this.

Measures of the average or central tendency

It is often helpful to summarise number data by means of an average – which gives an idea of the **"middle" or the "central tendency"**. There are two useful and commonly used averages: the *mean*[3] and the *median*. (Another one often mentioned is the mode, but this is a bit awkward and not much used, so I will ignore it here.)

The ***mean*** is obtained by adding up all the numbers and dividing by the number of numbers. In the example in table 1, if you add the numbers up you will get 1862, and if you divide by the number of numbers, 50, you get 37.24 – which is the mean of these numbers.

The ***median*** is the middle number when they are arranged in order of size. In Table 1, if you arrange the numbers starting from the lowest you get

2, 3, 6, 8, 9, 11, 11 …

The 25th number is 24, and the 26th number is 25. As there are an even number of numbers (50) there is no middle number, so the obvious number to take as the median is half way between the 25th and the 26th. In this case the median comes to 24.5.

So the mean of the data in Figure 1 is 37.24, and the median is 24.5. Both are somewhere in the middle, but obviously in a slightly different sense.

Which is the most useful? It obviously depends on what you want. The median is better as a "typical" value – half the sample earn more and half earn less. The mean is what everyone would get if all the money was shared out equally – which of course it isn't!

There are **Excel functions** for mean (=average) and median (=median), and you can also use Excel to sort the data (Data – Sort on the menu at the top), provided it is in a single column.

Quick question 2. What are the mean and median of the first row of data in Table 1?

Quick question 2a. What do you think are the mean and median incomes of UK taxpayers? Which would you expect to be the greater? (See http://www.hmrc.gov.uk/stats/income_distribution/3-2tabledec08.pdf for a slightly out of date answer.)

Quartiles and percentiles

Like the median these are also **based on putting the data in order**. The first quartile is 25% of the way up the list, the third quartile is 75% of the way up, the 80th percentile is 80% of the way up, and so on. The median, of course, the 50th percentile.

For example, the first quartile (25th percentile) of the data in table 1 is "number 12.5" in the list (as there are 50 of them). This is midway between the 12th and 13th number in the list; these are both 16, so the first quartile is 16. Similarly the third quartile is 42[4]. (When people talk about "quartiles" they mean the first and third quartiles. The second is, of course, the median, and the fourth is the maximum.)

Quick question 3. What are the 5th and 95th percentiles of the data in Table 1? How would you describe this information to someone who was not familiar with the term "percentile"?

Measures of spread or dispersion or variation

As well as the central tendency or average, it is also useful to know how spread out the data is. For example the earnings in Table 1 are very spread out ranging from 2 to 159 (thousand pounds). If, on the other hand, everyone earned 37 thousand pounds, the mean (average) earnings would be the same but the spread would be much less – in fact it would be zero. Spread can be measured by:

The ***range*** is simply the biggest number minus the smallest: 159 – 2 = 157 thousand pounds in the case of Table 1.

The range is very dependent on the two extremes, and will usually be larger with larger samples (see below for examples). This means it is a rather **erratic and unsatisfactory measurement**. A more stable measure is the ***interquartile range*** which is the difference between the first and third quartiles: 42 – 16 = 26 thousand pounds (quartiles are explained in the previous section).

The ***mean deviation from the mean*** is another way of measuring spread. The arithmetic is hard work here so I'll illustrate it with the last row of Table 1 only. The mean of the five numbers (110, 53, 11, 27, 39) is 48. The deviations (i.e. differences) of the five numbers from this mean are:

62, 5, 37, 21, 9

and the mean (average) of these deviation is 26.8. (The Excel function is =avedev .)

In practice the mean deviation from the mean is very rarely used. The statistic which is used is the ***standard deviation*** which is the *square root of the mean of the squares of the deviations from the mean*. The way to work this out starts off in the same way with the mean (48) and the deviations from the mean:

62, 5, 37, 21, 9

However, now we *square* the deviations (i.e. multiply each one by itself):

3844, 25, 1369, 441, 81

and find the mean of these by adding them up (5760) and dividing by 5 (1152). The standard deviation (*sd*) is defined as the *square root* of this, which is 33.9.

It should be obvious that both the sd and mean deviation from the mean provide a measure of the spread because **the more spread out the numbers are, the greater the deviations from the average will be**. The only difference is that the standard deviation involves squaring the deviations and then taking the square root at the end to get back to something in pounds rather than "square pounds". The sd is almost always bigger, but the two statistics are usually a similar size and measure much the same thing. The mean deviation is more intuitive, so it may be useful to give you a **rough idea** of what the sd is.

If you use the same method to work out the sd of *all* the numbers in Table 1, the answer is 34.8. This illustrates the principle that the **sd of a small sample is likely to be similar to the whole group from which the sample is taken**. For obvious reasons, this is not true of the range; the range of this last row of data is 99 (=110 – 11), and of the one above is only 45, both of which are a lot smaller than the overall range of 157.

The reason why the sd tends to be used rather than the mean deviation from the mean, despite the fact that it is more complicated, is that it **fits in with the theory of the normal distribution** (see below). You need to know the sd to use the normal distribution.

A further complication of the sd is that it comes in **two versions**[5]. The version we have worked out above is =stdevp in Excel, or σ_n on many calculators. The other version, =stdev in Excel or σ_{n-1} on many calculators, is always a bit bigger. This second version is recommended when you are using a *sample* of data to estimate the sd for a *wider population*, because it can be proved mathematically that the other formula will tend to produce estimates which are consistently slightly too small. However, for large samples (say, more than 20) the two are very similar, so the distinction is not worth worrying about!

Quick question 4. Work out the range, average deviation and standard deviation of the 8th row of data in Table 1:

3 42 42 29 54

Would it be sensible to work out the inter-quartile range from a sample this small? (And why have I chosen the 8th row in Table 1 rather than, for example, the first?)

Quick question 4a. Now do the same for a group of five numbers which are much less spread out: 44, 42, 41, 41, 42. Are the results as you expect? Check that you can see how all three statistics measure the spread, although in a slightly different sense.

Probability and the normal distribution

A probability is a number between 0 and 1 representing how likely something is to happen. If the something is impossible, the probability is 0; if it's certain to happen, the probability is 1.

If the probability is 0.5, this means that if we give something the chance to happen on 100 occasions, it is likely to happen on about half of them, or about 50 times. Similarly, the probability of obtaining two heads if you toss two coins is 0.25 (see below): this means that if you did this 1000 times, you would get two heads on about 250 occasions. Probably not exactly 250 times, but the ***expected*** number would be 250. The word "expected" in statistics is used in the technical sense of what we would expect in the long run, taking probabilities into account.

When planning for the future, we often have to acknowledge that we cannot be certain what is going to happen. **Probabilities are useful because** they allow us to quantify uncertainty so that

we can distinguish what is fairly likely to happen from what is very unlikely, and to estimate the impact of a train of uncertain events.

Quick question 5. Suppose the data in Table 1 comes from a random sample of adults in a particular town. Use this data to estimate the probability of a randomly chosen adult earning more than £50,000 per year.

Probability theory is a very complex branch of mathematics. One of the things it enables us to do is to calculate probabilities of various combinations of different events. For example, if we toss two coins, we may get no heads, one head or two heads. The probabilities of these three possibilities are 0.25, 0.5 and 0.25. (You may be able to work this one out in your head, but if you were tossing ten coins you would probably need some help from probability theory.) This is known as a ***probability distribution*** because it indicates how the probabilities are distributed (or allocated) across the possible outcomes – and so if we have a particular outcome in mind (say one head) it gives us the probability of this specific outcome occurring. Probability theory includes formulae to calculate probabilities for specific types of distribution like this[6].

There is one distribution which is of particularly importance because it appears all over the place: the ***normal distribution***. This applies when you have a number variable that ***depends on a large number of small independent factors[7].*** Starting from this assumption, mathematicians have derived a formula for the shape of the distribution – a symmetrical bell shape like Figure 2. (The formula is too complicated to use without the help of computer software or tables – if you are interested in seeing it, try a Google search.)

As an example, the heights of 18 year-old girls depend on a large number of genetic and environmental influences, so they should follow the normal distribution. If you got a very large number of such girls, measured their heights, and drew a very detailed histogram, the result would look very like Figure 2.

Figure 2: Normal distribution with mean = 163 cm, sd = 6 cm

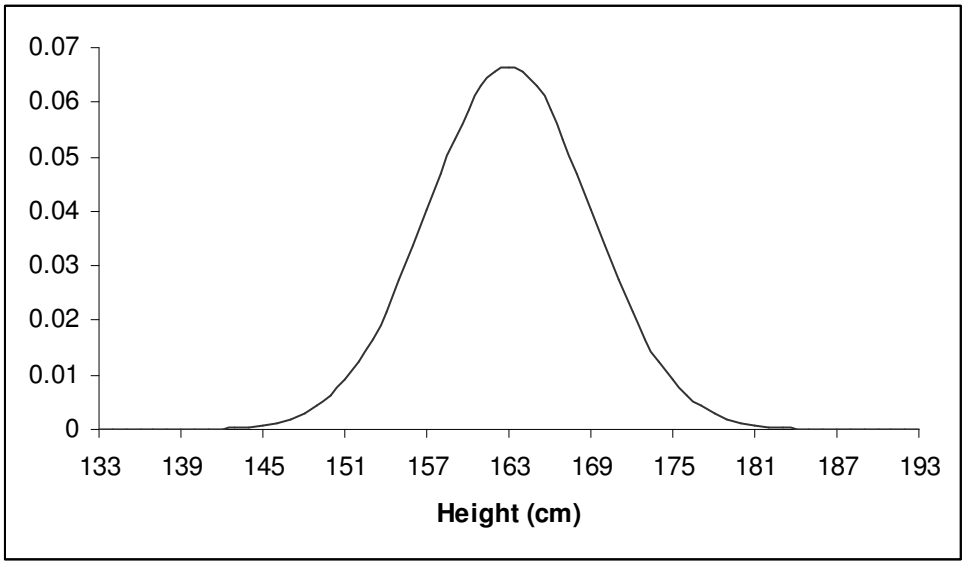


Figure 2 is based on the mathematical formulae for the normal distribution. To use this formula, or computer software based on it, **you need to know the mean and sd**: the values used (163 cm and 6cm) are based on the survey reported in Gregory and Low (2000).

Quick question 6. Imagine that in another country the mean height of 18 year old girls is the same, but the standard deviation is only 2 cm. Sketch this distribution on top of the one in Figure 2, and then check your answer by changing the sd in cell D2 of the Normal sheet of http://woodm.myweb.port.ac.uk/stats/StatNotes.xls. What would the graph look like if the sd were 12? What would happen if the mean were larger – say 170? In each case decide what you think the answer is, then use the spreadsheet to check.

Figure 2 was produced using the normdist function in Excel. It suggests that the most likely height is 163 cm, and that almost no women are less than 145 cm or more than 181 cm. (It is best not to take any notice of the vertical scale on this diagram. It can be used to estimate probabilities, but using Excel directly – as described below – is much easier.)

Quick question 7. Use the graph in Figure 2 to estimate (roughly) the percentage of girls who are taller than 170 cm. Does this seem about right?

To get more detailed estimates of probabilities you need to use Excel. The **normdist function** should be self-explanatory (with the aid of Help), with the possible exception of the *cumulative* parameter. If you set this to True, the function will give you the cumulative probability *up to* the value in question: for example, for x=163 it will give you 0.5, and for x=151 it will give 0.023 (rounded off to three figures).

Even without a computer, you can use the normal distribution to make some rough estimates of probabilities. The numbers on the horizontal axis in Figure 1 are chosen to show how this is

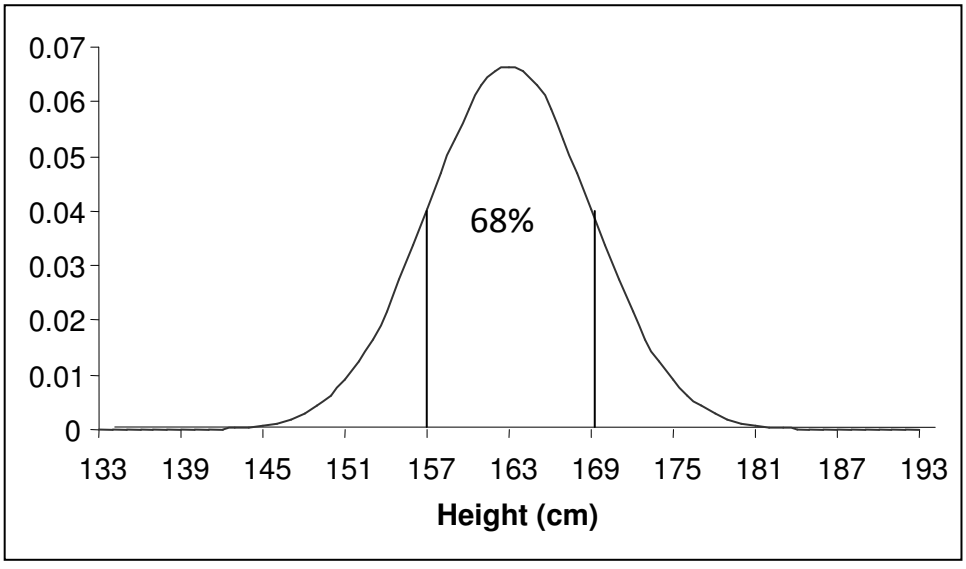
done. The mean is 163 and the sd is 6, so 169 is one sd above the mean, 175 is two sds above the mean, 151 is two sds below the mean, and so on.

It is worth remembering (or writing down somewhere) the following approximate facts about the normal distribution:

About 68% of individuals lie within one standard deviation from the mean
About 95% of individuals lie within two standard deviations from the mean
About 99.7% of individuals lie within three standard deviations from the mean

I'll take the 95% fact as an example. Two standard deviations below the mean is 151, and we saw above that Excel gives a probability of 0.023 or 2.3% for the probability of an individual woman being less than 151 cm. As the distribution is symmetrical there will be another 2.3% who are more than 175. This means that the rest (100% - 2.3% - 2.3% = 95.4%) will be within two standard deviations of the mean. Similarly, the Excel function shows that the one standard deviation range encompasses 68.3%. Figure 3 illustrates this 68%. (Strictly it should be 68.2689 to four decimal places, but it's usual to round it off to 68%.)

Figure 3: Normal distribution with the middle 68% marked



These facts give you a convenient way to make sense of the normal distribution and the standard deviation. If a distribution is normal, "almost all" individuals will lie within three sds of the mean, and two thirds (68%) will lie within one sd.

Quick question 8. A psychometric test is designed so that the mean score is expected to be 100, and the sd is expected to be 15. Suppose that 600 people take the test. How many would you expect to score:

(a) between 85 and 115

(b) between 70 and 130
(c) more than 130
(d) more than 150?

Would you expect your expectations to correspond exactly to the actual numbers?

The normal distribution is, as its name implies, **pretty normal** – it applies in a great many very diverse situations. Look at a few histograms: you will be surprised how often you get the characteristic symmetrical shape of the normal distribution. (The simulations you will meet when we look at hypothesis testing and confidence intervals usually show this pattern, for example.) **However, there are exceptions**: Figure 1 above is obviously not normal.

Quick question 9. Which of the histograms at http://woodm.myweb.port.ac.uk/stats/Histograms.pdf follow the normal distribution, and which don't?

Exercises

1 Click on the Exam worksheet of http://woodm.myweb.port.ac.uk/stats/StatNotes.xls . (This data is genuine.)

(a) Work out the mean and standard deviation of the marks for the exam and the assignment. (Use the statistical functions in Excel by clicking on the f_x *symbol.)* Which had the higher average mark? Which set of marks was more spread out?

(b) Draw a histogram of the exam marks. You will need to decide on class intervals – I would suggest 0-9, 10-19, 20-29 and so on. Now count up the number in each interval and draw the histogram. (If you are not familiar with drawing graphs using Excel it may be easier to sketch it on a piece of paper.)

(c) Work out the quartiles and the inter-quartile range of the exam marks. (There is an Excel function for quartiles, but you should be able to work them out without this, and then use the Excel function to check.)

(d) Does this distribution look statistically normal? Work out the the proportion of students with marks between one standard deviation below the mean and one standard deviation above? Is this close to the normal prediction?

(e) Now work out the mean and standard deviation of the exam marks of the *first 10 students*. Are the results as you would expect? (In the notes above, we found that the standard deviation of one row of data in Table 1 was similar to the standard deviation of all the data. Why isn't the sd of the first 10 students' marks similar to the sd of all the students?)

2 The data on flow rates in the Aerosol worksheet in the spreadsheet http://woodm.myweb.port.ac.uk/stats/StatNotes.xls are quality control data taken during the

night shift at a factory in Portsmouth making aerosol nozzles. A critical feature of these nozzles is the size of the hole or the flow rate – the rate at which gas passes through the nozzle. If this is too high or too low the aerosol will not work properly. In this particular case the requirement is that the flow rate through the nozzle should be between 28 and 39. In order to monitor the process the flow rates of a sample of 20 nozzles was measured each hour – as in the spreadsheet. (This data is genuine except for (i) below.)

(a) Sort the data in flow rate order so that you have the low numbers at the top of the list (select the data, then click on Data – Sort.). Now use the sorted data to read off the median flow rate, and the quartiles. Then work out the range and inter-quartile range.

(b) *Estimate* the mean and standard deviation. You should be able to make a rough guess by looking at the sorted data.

(c) Use the functions in Excel to check your estimates to the mean and standard deviation.

(d) Use both Excel functions for the sd (stdevp and stdev) and check that the second is a bit larger than the first, but only very slightly.

(e) What do you consider the most useful average (mean or median) to use? What do you consider the most useful measure of spread? Why?

(f) Produce a histogram of the original data. **(**The easiest way of doing this is to use the sorted data to count up the number of 29s, 30s etc, then use this to make a frequency table and then a bar chart.) Does it look roughly normal? How many nozzles are outside the tolerance range of 28-39?

(g) Do you think that the process is operating in a satisfactory manner (bearing in mind that the flow rates *must* be between 28 and 39)?

(h) Use the normal distribution to *predict the proportion of the total production during the shift* that will be outside the tolerance range of 28-39. **(**You should be able to get a rough estimate by working out how many standard deviations from the mean these numbers are. To get a more precise answer you will need to use the Excel function normdist.**)**

(i) The night after, the mean of the sample was 32, but the standard deviation had increased to 6. Make up a sample of 20 flow rates with this mean and standard deviation. Is the process operating in a satisfactory manner now?

(j) Now use the normal distribution to predict the percentage of nozzles outside the tolerance range if the mean is 32 and the sd is 6..

(k) In (i) and (j) your normal predictions probably won't correspond exactly to the actual proportion in the sample which are outside the tolerance range. Why is there a difference?

3 In a firm with 2000 employees the upper quartile of the salary distribution is £25,000, the 95th percentile is £40,000, the mean salary is £19,000 and the standard deviation of the salaries is £8,000. Answer as many of the following questions as you can (you may not have the information to do them all):

(a) How many employees earn more than £40,000?

(b) How many employees earn £25,000 or less?

(c) How many employees earn between £25,000 and £40,000?

(d) How many employees earn less than £11,000?

Additional exercises

4 The spreadsheet at http://woodm.myweb.port.ac.uk/nms/drink.xls has data on the drinking habits of a sample of students – see the Notes sheet in the spreadsheet for a brief explanation.

(a) Find the mean and the median of the number of units drunk on Saturday. You should find one is substantially more than the other. Why are they different? Which do you think is the better measure of the average alcohol consumption of this group of students on Saturday?

(b) Which of the three courses has the greatest variation in the age of the students? Choose a suitable measure of variation (spread) and work it out for each of the three courses. (The easiest way to do this is to start by sorting the data by course.)

(c) Draw three histograms to show the age distributions of each of the three courses. Are these distributions normal?

5 Each class interval in Figure 1 is 10 units wide. Do you think this is a good choice. Produce a histogram from the same data with a class interval width of 20, and another with a width of 1. Which do you think is most useful?

6 The CEO of a large automobile company is concerned about the design of a new version of a well-known model. There is a risk of fuel-led fire, and:

* The plan is to produce 40 000 000 cars.

* The probability of explosion is 1/100 000.

* Maximum compensation of 200 000 euros per death

* The cost of re-design is 9 euros per car

What do you think he should do?

(From a presentation by Marc Menestrel who got it from Mark Ashbar: "The Corporation".)

7 Look at the data and graphs in sharesan.xls (at http://woodm.myweb.port.ac.uk/nms/sharesan.xls). Much of the theory of stock markets assumes that share returns are normally distributed. Is this assumption justified for the three shares here? You can check this by looking at the shape of the histogram, and also by comparing the percentiles of the distribution with the normal figures. For example, according to the normal distribution, the 2.5th percentile should be about two standard deviations below the mean, and the 97.5th percentile should be about two standard deviations above the mean. Are these relationships right?

References

Gregory, J, and Lowe S. (2000). *National diet and nutrition survey: young people aged 4 to 18 years, Vol. 1.* London: Stationary Office.

Saunders, M., Lewis, P., & Thornhill, A. (2003). *Research methods for business students (3rd edition)*. Harlow: Pearson Education Ltd.

Wood, M. *Making sense of statistics: a non-mathematical approach*. Basingstoke: Palgrave, 2003.

Answers to Quick questions and notes on a few of the exercises

2 Mean is 47.2, median is 49.

2a The mean will definitely be greater than the median.

3 As there are 50 numbers, the 5th percentile will be "number 2.5" in the list when arranged in order of size. Arranging them in order of size, the list is 2, 3, 6 .. so number 2.5 is midway between 3 and 6, so the 5th percentile, using this method is 4.5. Similarly, the 95th percentile is number 47.5 in the list which is midway between number 47 (106) and 48 (110), which is 108.

However, using the percentile function in Excel, I got 6.9 and 108.2. Excel uses a slightly different, more sophisticated method. With larger samples, the two methods will tend to give very close answers. From the point of view of a basic understanding of what percentiles are, I don't think it is worth bothering about the more sophisticated method, but if you are interested see this note[8].

Taking the Excel answers, we can say that 5% earn less than 6.9 thousand pound a year, and 95% earn less than 108.2.

4 51, 14.4, 17.4 (or 19.5 using the other sd function.) This sample is too small to estimate the quartiles accurately.

4a 3, 0.8, 1.1. As you should expect these are all far less than the answers to the previous question.

5 11/50 or 22%.

6 If the sd=2 the distribution will be thinner indicating that the variation in heights is smaller. If the sd=12, the distribution will be fatter, and if the mean is larger the whole distribution will shift to the right.

7 The answer according to the Excel function is 12%. You should have something between 5% and 20% by eye.

8 (a) 408. (b) 570. (c) 15. (d) 0. No, these expectations will not correspond exactly to what happens. This means you should not worry about the exact answers. For example, I generally think of the "one standard deviation range" as encompassing two thirds of individuals in the population, which gives an answer of 400 for (a) – which is reasonable. The exact probability from Excel is actually 68.2689% which gives an answer of 410.

9 The only normal one is the heights of the army recruits. The marital problems and the goals in the football matches definitely aren't normal. The project marks looks roughly normal, but with a chunk taken out: there are no marks between 20 and 49 whereas the normal pattern at the top of the distribution might lead one to expect some marks here. The pass mark was 50%, and the likely explanation for this pattern was the reluctance to fail candidates! This is a good example of the way that a histogram can suggest a story about the data.

Exercise 2.

a, b, c, d. You should be able to use the Excel functions to check the median and quartiles. The range should be 5, and both standard deviations and the inter-quartile range should be greater than 0 but less than 1.5.

g, h, i. The process is not really operating in a satisfactory manner because the bottom of the distribution is close to the lower specification limit. It is true that none of the nozzles in the sample is outside the tolerance range of 28-39, but when you work this out using the normal distribution you should get a small proportion, but one that is definitely more than 0. The normal distribution predicts what would happen if you had a lot more nozzles – as would be produced from the actual production process – and the pattern does suggest that you would get the occasional nozzle with a flow rate below 28. (The all assumes, of course, that the distribution follows the normal pattern with the mean and standard deviation worked out from the sample.) A histogram is a very good way of showing the general pattern so as to see if there are any potential problems like this.

j, k. If the standard deviation increased to 6, this would be very bad news for the process, and the number falling outside the tolerance range would be much higher. The smaller the standard deviation the less the variability of the process and the more consistent the output.

Endnotes – these include more detail which can be ignored for a rough understanding

[1] There is a frequency function in Excel but it is an *array formula* – you will need to check the Help to see how to use it.

[2] **S**trictly, there are two minor problems with Figure 1. First, the midpoint of the first class interval should be 5.5, the second should be 10.5, and so on. Second, the second figure in Table 1 is 60: if this means £60,000 rounded off to the nearest thousand, then it might correspond to any income from £59,501 to £60,499. This means that the boundary between the bars labeled 55 and 65 in Figure 1 should be at 60.5 (thousand pounds) and not at 60. However, as the idea is just to show the pattern, it is usual not to bother about details like these. Excel will produce histograms if you install the Analysis Toolpak (click the Office button on the top left of the screen, then Excel options, then Add-ins) and click Data then Data analysis, but they are not very good – unfortunately the bars are labeled by the top of each class interval rather than the middle.

[3] Strictly, this is the *arithmetic* mean to distinguish it from the *geometric* mean. This works on a similar principle, but instead of adding the numbers and then dividing by the sample size, *n*, you multiply the numbers together and then take the *n*th root, so that you end up with a sort of multiplicative average.

[4] If you use this method for the median you will take the 25^{th} number rather than midway between the 25^{th} and 26^{th}, so you may wonder if you should make a similar adjustment for other percentiles. On the whole this is not worth the bother, as adjacent numbers are likely to be very close with large samples, and you need large samples to estimate percentiles accurately. The percentile function in Excel does make such an adjustment (a rather more complicated one, in fact), which is why the answers produced by this function may not correspond exactly to the answers given by the method here. This is not worth worrying about. With large samples all methods will almost always produce very close answers. Anything close is good enough!

[5] The difference between the two versions lies in the method used to work out the mean of the square deviations. For the second, bigger, version we divide by one less than the size of the sample. In the example above, this means we divide 5760 by 4 (instead of 5) and get 1440. The square root of this is 37.9 which is the value of second

type of standard deviation. It is possible to prove mathematically that this second type of standard deviation will give better estimates of the standard deviation of a wider population. For larger samples, the two are closer together, so the distinction matters less.

[6] There is a brief explanation of the basic rules of probability at http://woodm.myweb.port.ac.uk/stats/ProbRules.pdf. Probability questions may be more complicated than they appear – one classic puzzle which has fooled many experts is the Monty Hall Problem (see, for example, Wood, 2003 or http://youtube.com/watch?v=mhlc7peGlGg.)

[7] Strictly the variable should be the result of *adding* up all these factors. If the factors are multiplicative the distribution will be *lognormal*.

[8] To see one problem with the crude method, try counting from the top down. Then the 95th percentile should be number 2.5 in the list which is 133 (midway between 110 and 156). Similarly the 5th percentile comes to 7. Both answers are different from before! The Excel method gets round this problem, and also makes a correction for the fact that, because the numbers are likely to be bunched in the middle, the midway point between two numbers, percentile-wise, is likely to be closer to the middle of the whole group than the numerical midway point. However, this all gets rather complicated, and in real problems, with large samples, it makes very little difference to the answers.