

# Getting the most out of NQF statistics: A guide for users

## Part 5 - Elementary concepts in data analysis

This section introduces a number of elementary concepts in statistical description and analysis. For the most part, it is for teachers and other data users who have little or no background knowledge of statistical concepts. However, some of the topics are more advanced, in particular the sections on confidence intervals and chi-square tests. These techniques are included to provide an opportunity for readers with a basic knowledge of statistics to learn some more advanced approaches to data analysis.

An understanding of the first five topics described below (up to and including *data aggregation*) is important to most analyses that might be undertaken. The concepts presented here are in a condensed form, and usually with specific reference to NQF/NCEA data. Some readers might find it useful to consult an elementary statistics textbook to supplement this information.

The concepts introduced are:

- 1 The difference between raw numerical data and *percentage* data, and how to calculate that percentage data from raw numerical data.
- 2 Plotting *histograms* to provide graphical displays of data.
- 3 The *mean* and the *median*, two measures of central tendency, to provide ways of determining a typical value in a distribution of values.
- 4 The *standard deviation*, which is a measure of how widely the data are spread around the typical value.
- 5 Aggregating data to provide more robust estimates of underlying trends and patterns.
- 6 Measurement error and confidence intervals which provide a way of determining the reliability of a statistical estimate, and
- 7 Use of *chi-square tests* to compare distributions of grades.

### Percentage data

A *percentage* is the rate at which a particular phenomenon or characteristic is observed out of every hundred cases<sup>1</sup>. The NZQA website presents data both as percentages and as simple numbers. Generally speaking, percentage data are more useful for statistical analyses than raw numerical data.

To calculate a percentage, divide the number of cases in which the phenomenon or characteristic is observed, by the total number of cases, and then multiply the result by 100. For example, in 2004, 680,838 NQF results out of a total of 1,072,837 results gaining credit, achieved by Year 13 secondary students, were for achievement standards. To determine the *percentage* of total results that were achievement

---

<sup>1</sup> It is not necessary to have as many as 100 cases to calculate a percentage. The way to think about a percentage is that it is the number of cases that *would* have a certain characteristic if the total number of cases were exactly 100. So if 25 out of 50 actual cases show a characteristic, then the percentage is 50%; 1 out of every 2 cases has the characteristic, which is the same as 50 out of 100.

standards, we divide 680,838 by 1,072,837, which gives 0.635 (to three decimal places). Multiplying this number by 100 gives 63.5%. So, 63.5% of all NQF results gaining credit, achieved by Year 13 secondary school students in 2004, were in achievement standards (rather than unit standards). This means that 63.5 out of every hundred results were in achievement standards.

The useful aspect of percentage data is that it allows a direct comparison of different numbers of cases; for example, achievement rates for a qualification at two different schools with different numbers of students. For example, if one school has 130 Year 13 students, 95 of whom achieve level 3 NCEA, and another school has 46 Year 13 students, 31 of whom achieve level 3 NCEA, which school has the highest achievement rate for this qualification? Ninety-five is 73.1% of 130 and 31 is 68.9% of 45. So, the two schools show quite similar achievement rates for level 3 NCEA amongst Year 13 students, although the first school shows a slightly higher rate than the second. The percentage data make this explicit, whereas it is not necessarily clear at first glance from the raw numbers.

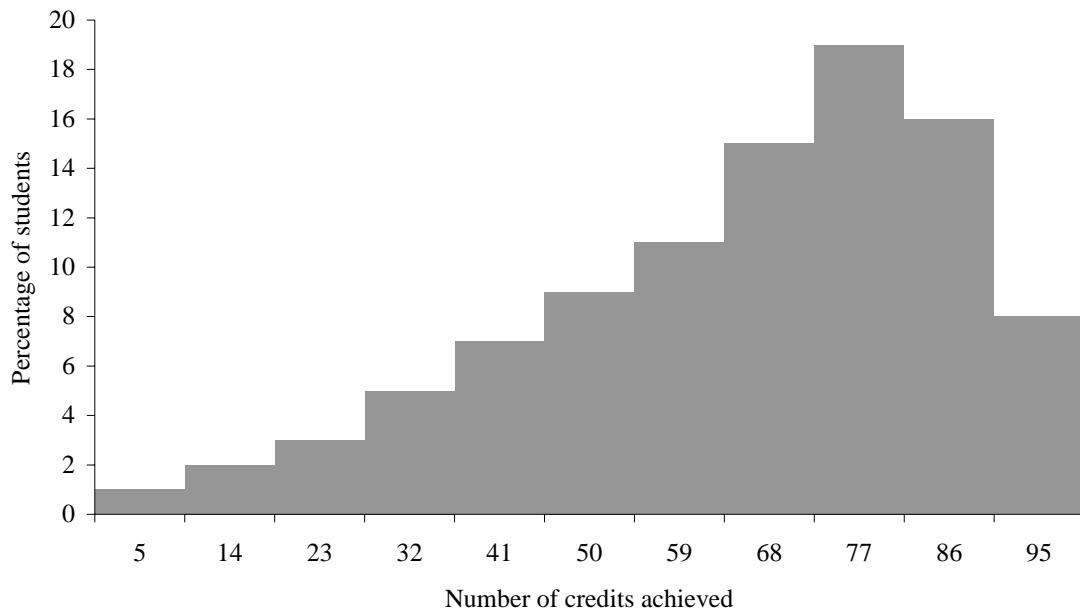
When reporting percentage data, it is important to make it clear what a percentage is a percentage *of*. It is not useful, for example, to say that a particular school has a success rate of 70% in Level 2 NCEA. Does this mean that 70% of *all students* at this school *leave* with Level 2 NCEA? Does it mean that 70% of *all Year 12 students* attained Level 2 NCEA *in a given year*? Or 70% of all *senior students*? Without this information, the meaning of percentage data may be unclear and may potentially result in misinterpretation.

### **Plotting histograms**

A histogram (which is essentially a modification of a standard bar chart) provides an excellent way of graphically depicting a *distribution* of data. Although data are depicted as bars in discrete intervals, with the height of each bar representing either the number or the percentage of data in that interval, the horizontal scale on a histogram is continuous, and typically labelled according to the mid-point of each interval. In Figure 5a, the width of each interval is 9 (credits) so, for example, the first interval is labelled 5, being the midpoint between 1 and 9.

A normal distribution is one that, when plotted on a histogram, has a classic bell-shaped curve, meaning that most values are close to the centre of the distribution, with fewer and fewer values towards the extremes. A great many naturally occurring phenomena, including many human abilities, vary according to a normal distribution. Examples of idealised normal curves are shown in Figure 5c.

Figure 5a Histogram showing hypothetical data for the percentage of students in a school achieving numbers of credits in intervals of nine credits



Note that Figure 5a makes the shape of the depicted distribution visually clear, which it would not be if the data were presented as a list of numbers. The distribution illustrated in Figure 5a is what is known as a skewed (asymmetrical) normal distribution. This means that although it has a clear peak in the interval labelled 77, the tail of the distribution in the negative direction (to the left of the peak) is longer than that in the positive direction (to the right of the peak). For this reason, this distribution is labelled as *negatively skewed*; if the tail were longer in the positive direction it would be *positively skewed*.

### Measures of central tendency: the mean and the median

A large set of numbers is usually difficult to make sense of, and one of the questions frequently asked is, '*what value is most representative or typical of the set of data as a whole?*' One would expect that a typical value would be one that is close to the centre of the distribution. For this reason, statistics that represent typical values in a distribution are called measures of *central tendency*. But which value should be chosen? The two most commonly used measures of central tendency are the *mean* and the *median*. (There is also the *mode*, defined as the most common value in a distribution.)

The *mean* is another name for the average. It is calculated by totaling all values in the distribution and dividing by the number of data. For example, the mean of the distribution {2, 3, 3, 5, 9}, is the sum of these numbers, which is 22, divided by the number of data, which is 5. So the mean is 22 divided by 5; that is, 4.4.

The *median* is the value that is exactly in the middle of a distribution. To find the median, all of the values in the distribution are placed in rank order, and the value that is exactly half way through the ranking is identified as the median. (If there is an even

number of values in the distribution, the median is the average of the two central values.) In the distribution {2, 3, 3, 5, 9}, the median is 3; the third value in the list when the five values are placed in rank order. In the distribution {1, 2, 2, 5, 7, 9}, the median is 3.5, the average of 2 and 5, the third and fourth values in the rank order.

When a distribution is exactly symmetrical, the mean and the median are the same. A symmetrical distribution is one in which the values are evenly distributed around the median. For example, the distribution {1, 4, 7, 10, 13} is symmetrical. The median of this distribution is 7. There is one value that is 3 greater than the median (10) and one value that is 3 less than the median (4). Similarly there is one value that is 6 greater than the median (13) and one value that is 6 less than the median. On the other hand, the distribution {1, 3, 6, 10, 13} is not symmetrical. The median is 6, and there is a value 4 greater (10) and a value 3 less (3), as well as a value 7 greater (13), and a value 5 less (1). Note that in the first (symmetrical) distribution the mean is 7, the same value as the median. On the other hand, in the second (asymmetrical) distribution, the mean is 6.6 – slightly greater than the median of 6.

The fact that the median is less than the mean for the second distribution illustrates that it is positively skewed, which means that there is a concentration of data at the low end of the distribution, with the data more spread out at the upper end. Similarly, if the median were greater than the mean, it would show that the distribution was negatively skewed (see Figure 5b for an illustrative example).

The reason that the mean and median of an asymmetrical distribution are different is that in determining the median, only the rank order of the values is taken into account, whereas in determining the mean, the actual values themselves are used. When extreme values are exactly balanced, as they are in a symmetrical distribution (i.e., for each value greater than the median, there is another value less than the median by the same amount), they cancel one another in the averaging process, so that the mean ends up being exactly equal to the median. However if the distribution is negatively skewed (i.e., the values deviate from the median in the positive direction by greater amounts than they do in the negative direction), the mean will end up being positively weighted relative to the median i.e., it will be greater than the median.

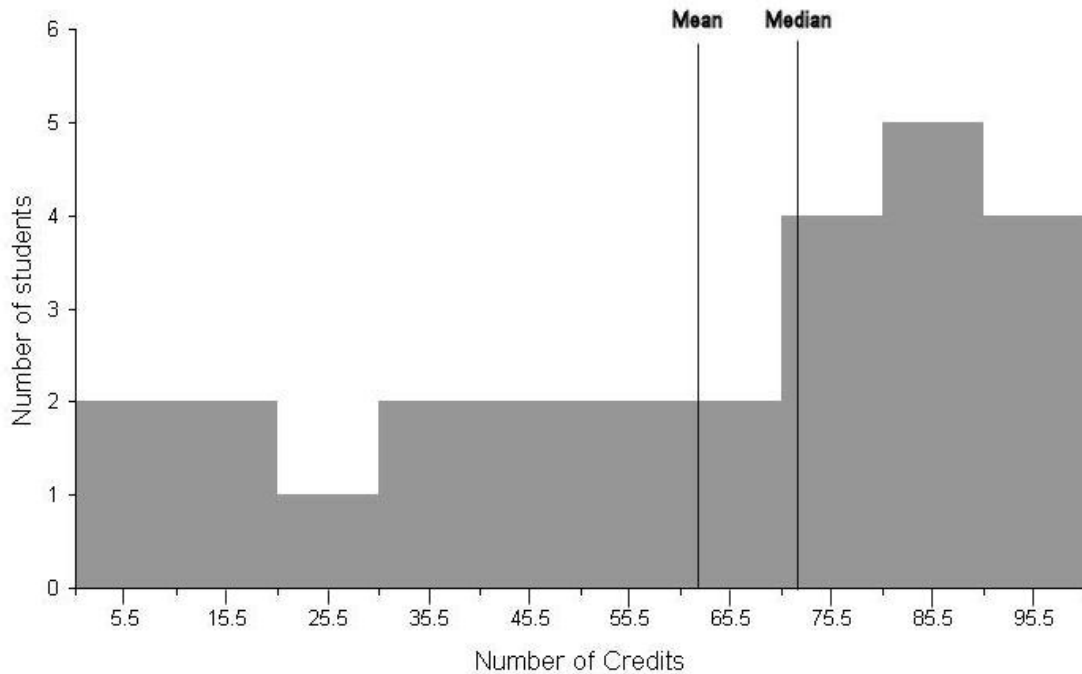
Whether it is better to use the median or the mean depends on the purpose of an analysis. If a distribution is symmetrical, it doesn't matter, because the median and mean will be equal, but if the distribution is asymmetrical, this is not so. If it is important to take into account extreme values in a distribution, then the mean provides a better measure. The mean has an added advantage in that the standard deviation can be used to determine the proportion of the data that lie within a specified distance from the mean on a normal distribution. If all that is wanted is the most central value, then the median is the appropriate statistic to use. As shown in Figure 5b, on an asymmetrical, *unimodal* distribution (i.e., one with a single peak), the median is always closer to the peak of the distribution than the mean; that is, it is closer to the part of the distribution where the data are most densely represented.

#### Mean and median example

The following distribution in Figure 5b represents the number of credits gained in a year by Year 12 students at a small (hypothetical) school. There are 26 data in the

distribution plotted in a histogram {3, 7, 12, 20, 26, 33, 37, 41, 47, 51, 60, 63, 70, 73, 75, 78, 80, 83, 87, 88, 88, 90, 95, 96, 97, 99}.

Figure 5b Histogram showing the number of credits gained in a year by a Year 12 cohort at a hypothetical school



The mean of this distribution is 61.7, and the median is 71.5. It shows an asymmetrical, unimodal distribution (i.e., one with a single peak), and a long tail in the negative direction (negatively skewed). It peaks between 81 and 90 credits (in the interval labelled 85.5), and the median is closer to the centre of the peak bar than the mean, as it will always be for negatively skewed data.

### The standard deviation: a measure of spread

Knowing a typical or central value of a distribution is an important piece of information when it comes to characterising that distribution. It is also important to know how spread out the distribution is around that central value. There are various ways of measuring the spread of a distribution. The one described here which is probably the most useful and commonly used is the *standard deviation*.

The standard deviation can be thought of as the average *distance* from the mean of the values on a distribution. It also has the property that for any normal distribution (i.e., any distribution that forms a bell-shaped curve), the same proportion of the data fall within a given number of standard deviations from the mean. Around 68% of the data will always fall within one standard deviation, around 95% within two standard deviations, and more than 99% within three standard deviations. This point is illustrated in Figure 3 below.

Formula 1 describes how to calculate the standard deviation, however, it is more important to understand what it is, rather than the details for calculating it.



Formula 1. The standard deviation

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

$x$  refers to each datum (number) taken in turn

$\bar{x}$  is the mean value

$n$  is the total number of data

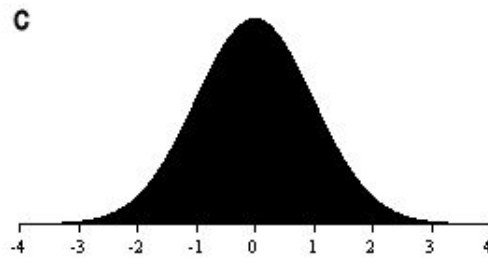
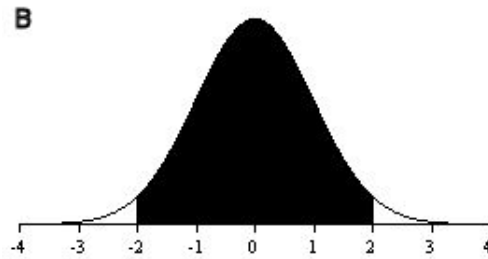
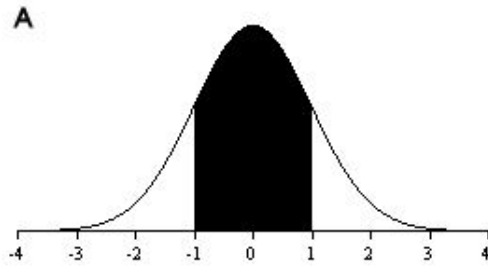
The Greek letter sigma ( $\Sigma$ ) means 'add up across all data'

In plain language, the formula above means the following: To calculate the standard deviation, follow these steps:

- Calculate the mean (see previous section)
- Subtract the mean from each datum (number) and square the result (i.e., multiply it by itself)
- Add up all of these squared differences
- Divide the result by one less than the number of observations
- Take the square root

The shaded areas in Figure 5c above show the proportions of data expected to fall within one standard deviation of the mean (Figure 5c - A: 68% of data), two standard deviations of the mean (Figure 5c - B: 95% of data), and three standard deviations of the mean (Figure 5c - C: more than 99% of data).

Figure 5c. Idealised normal curves divided into standard deviation units



### Aggregating data

When possible (given the question to be answered) it is a good idea to use data that are appropriately *aggregated*. When data are aggregated or *averaged*, the aggregated number is a better estimate of the true, underlying parameter than are the disaggregated data. This is because aggregation tends to cancel out random variability and suppress the effect of variability caused by deviation in one of the components of the aggregate. For these reasons, aggregated data are much more stable and reliable than disaggregated data.

Note that when aggregating data, the validity of the aggregate is likely to be compromised if there is a factor that differs systematically at different levels of the aggregating variable. For example, if results data showing student performance in a particular department are to be aggregated over several years, it is important that there are no (or only very minor) influences on student performance that are systematically stronger in some years than others e.g., a high teacher turnover or a major revision of the teaching programme.

#### Aggregating data example

An English teacher wishes to determine in which standard or standards her students show the weakest performance, so that she can reflect on her teaching methods in these areas. Consider the three standards presented over three consecutive years in Table 5.1. Note that the strongest performance was for a different standard in each

year: in 2002, standard 90056 showed the highest percentage of students gaining credit. In 2003 it was standard 90054, whilst in 2004 it was standard 90055. The weakest performance was in standard 90054 in 2002, but in both 2003 and 2004, it was standard 90056. The rightmost column below shows the data aggregated across the three years. This has the effect of canceling out some of the variability that caused the numbers to change from year to year, and the teacher can be confident that the aggregate will provide a better basis on which to determine the standards that show the strongest and weakest student performance. Note that standards 90054 and 90055 are just two percentage points apart in the aggregate, whereas differences of between 4 and 6 percentage points are evident in any given year. Standard 90056 shows the weakest performance in the aggregate (despite showing the best results in 2002). Aggregated data do not get rid of *all* unwanted variability, although the more instances (i.e., years) over which the aggregation can be performed, the less variability there will be due to measurement error.

Table 5.1 Percentage of students of a particular teacher gaining credit in three level 1 English standards over three years

	2002 (25 students)	2003 (28 students)	2004 (23 students)	<b>Three year aggregate</b>
<b>90054</b> <i>Read, study and show understanding of extended written text(s)</i>	55%	63%	59%	<b>61%</b>
<b>90055</b> <i>Read, study and show understanding of a number of short written texts</i>	61%	59%	64%	<b>62%</b>
<b>90056</b> <i>View/listen to, study and show understanding of a visual or oral text</i>	62%	53%	54%	<b>57%</b>

Aggregation in the example above was done over time (three years), but data can also be aggregated over students (e.g., to compare performance of different demographic groups), over standards (e.g., to compare performance in one subject with performance in another), over schools (e.g., to compare performance in schools with different decile ratings) and in a number of other ways.

To aggregate data, do not simply work out the mean (average) of the components of the aggregate, but rather *weight* each component by its size. This is one instance in which raw number data rather than percentage data can be useful. To see how weighting works, consider the following example.

#### Further example

The head of a Science department wants to determine the overall success rate of Level 3 Physics students in his department. However, because the school has developed a flexible programme, not all of the Level 3 Physics standards offered by the school are undertaken by all Level 3 Physics students. For this reason, to accurately determine the overall success rate in Level 3 Physics, it is necessary to weight the calculation according to the number of students undertaking each standard.

Table 5.2 shows all Level 3 Physics standards, the number of students undertaking each, and the proportion gaining credit in each, at the hypothetical school. To work out the average in the usual way, add together the rate of success for each standard, and then divide by four (the number of standards). This gives an average success rate of 57.5%. However, in this instance, because there are different numbers of students undertaking each standard, to get an accurate idea of the average performance it is necessary to produce a *weighted average*. To do this, first multiply the candidature for each standard by its associated rate of success. For 90520, this is  $17 \times 58\% = 9.86$ ; for 90521,  $25 \times 63\% = 15.75$ ; for 90522,  $12 \times 51\% = 6.96$ ; and for 90523,  $19 \times 58\% = 11.02$ . Next, add these four values together:  $9.86 + 15.75 + 10.71 + 13.34 = 43.59$ . Finally, divide the sum by the number of *results* (not the number of standards):  $43.59 \div 73 = 59.7\%$ .

Table 5.2 Level 3 Physics standards with numbers of candidates and rates of success at a hypothetical school

Level 3 Physics standard	Number of candidates	Percentage gaining credit
90520	17	58%
90521	25	63%
90522	12	51%
90523	19	58%

### Measurement error and confidence intervals

Whenever a sample of data is taken, it is susceptible to measurement error i.e., chance variation amongst the cases that happen to be selected from the population. For example, each time a coin is flipped, there is a 50% chance that the result will be *heads* and a 50% chance that it will be *tails*. This does not mean, however, that if the coin is flipped, for example 20 times, the outcome will be exactly 10 heads and exactly 10 tails. While this outcome is more likely than any other specific outcome, most often the actual result will be at least slightly different. This variability in the actual number of heads and tails is the result of measurement error. In this case, the population is the (infinite) set of all possible coin flips.

Assessment data are similarly susceptible to measurement error. When students complete an assessment, they are providing a sample from the set of all possible work that they *might* have completed for that assessment. While the quality of a student's work might be consistently good or poor, there is always some variation due to measurement error when data are sampled from a population.

It is often useful when reporting a statistic (especially a mean or a percentage which is actually a kind of mean), to give a *confidence interval* for that statistic. A confidence interval is a range of values within which we can be confident to a certain level of probability that the actual population parameter falls within as estimated by our statistic. Traditionally, the probability level is set to 95%.

Formula 2 shows how to calculate a 95% confidence interval. Using this formula, if for example, 55% of 70 students in a school are successful in a particular assessment, then the 95% confidence interval is 55% +/- 12%, meaning that we can be 95% confident that the true percentage for that class is between 43% and 67%. Another way to think about this is that if the assessment were repeated (assuming this could be done without learning on the part of the students), we would be 95% confident that between 43% and 67% of them would be successful. This is quite a broad interval, illustrating that when a sample is small, estimates of the characteristics of the population from which that sample is drawn are not very stable.

Formula 2. The 95% confidence interval

$$95\% \text{ confidence interval} = 1.96 \times \sqrt{\frac{p(1-p)}{n}}$$

$p$  is a percentage value divided by 100 (meaning that it will always be between zero and one)

$n$  is the total number of observations (data)

In plain language, the formula above means the following: To calculate a 95% confidence interval for a percentage figure, follow these steps:

- Divide the percentage by 100. This gives  $p$ .
- Subtract  $p$  from one
- Multiply the result by  $p$  itself.
- Divide the result by  $n$  (the number of observations)
- Take the square root of the result.
- Multiply by 1.96 (This give the confidence interval for  $p$ ).
- Multiplying this by 100 will convert the confidence interval into percentage terms.

### **The chi-square test: comparing distributions of results**

One common kind of data analysis involves comparing two distributions of values to determine whether there is a difference between them. For example, a teacher might want to compare the relative performance of male and female students in a particular assessment, or a Head of Department might want to compare the overall performance in externally assessed standards between that department and a similar department at another school.

When distributions are compared in this way, the question arises of whether a measured difference is *significant* or not; that is, whether it reflects a meaningful difference, or whether it might be attributed to measurement error, which is an inevitable source of variability whenever something is measured (see the *measurement error and confidence intervals* section above). Generally speaking, measurement error diminishes as sample size increases, so it's more likely that a distribution of data based on a large sample is an accurate reflection of the actual situation, than that of a distribution based on a small sample.

There is a large number of statistical techniques that can be used to compare distributions for significant differences. The correct one to use depends both on the nature of the comparison and on the nature of the data. For comparing distributions of data for achievement standards, the correct test to use is called the *chi-square test*. The computational details of the chi-square test will not be discussed here. It can be run easily using Microsoft Excel.