

Getting the most out of statistics: A guide for users

Part 1 – Aims of this guide

The aims of this guide are to improve educational delivery by:

- 1 assisting staff in schools in thinking about the kinds of analyses of NQF data that might be appropriate for their purposes¹.
- 2 assisting staff in carrying out and presenting analyses.
- 3 providing staff with important points and pitfalls to take into account when interpreting analyses for the purpose of improving educational delivery. This forms the underlying principle for this guide.

This guide is designed to be as accessible as possible to school staff and the general public who may not have statistical knowledge and expertise. People who may not feel confident using statistics should not be put off – useful information can be gained quite easily by plotting graphs or making tables of percentages (for example, of students gaining a particular qualification). Any technical terms that are used are explained.

It will also help people with some statistical knowledge to extend their analyses of NQF / NCEA data. ‘Part 5 – Elementary concepts in data analysis’ contains some important concepts in data analysis including percentage data, plotting histograms, means and medians (measures of central tendency), the standard deviation (a measure of spread), methods and applications of aggregating data, measurement error/ confidence intervals and a statistical test (chi-square) that can be used to compare distributions of results.

We welcome your feedback about this guide. Please feel free to send any feedback to monitoring-research@nzqa.govt.nz.

¹ It is not the intention or the role of the NZQA to instruct schools and the wider public on what these purposes should be. There are many possible analyses that will vary in usefulness and relevance between schools and between staff within a school.

Getting the most out of statistics: A guide for users

Part 2 – Planning an analysis

The amount of data available on the statistics pages of the NZQA website means that some users unfamiliar with statistical analysis can feel overwhelmed. Asking appropriate and well-framed questions, however, makes data analysis more straightforward with a well-framed question indicating which data to focus on and compare. The starting point for any meaningful data analysis and the important question to ask is '*what do I want to know?*' Starting with this question will help determine and organise the analyses that need to be carried out.

Table 2.1 below outlines the main steps involved in planning, conducting and interpreting data analysis with additional explanatory text for those who require it.

Table 2.1 Stages of planning, conducting and interpreting an analysis of NCEA/NQF data

Stages of data analysis	Further explanation
High-level question/s: This is the starting point of a good analysis. It should be a succinct summary of the reason for conducting the analysis (i.e. what is the main knowledge that will be gained?).	It is often useful to begin with a broad question, perhaps one that does not directly refer to data or statistics. For example, a head of department might ask, ‘how can I improve the educational processes in my department’? While a question like this cannot be fully answered from assessment data (because there is much more to good educational outcomes than success in assessment), assessment data can provide the best available quantitative measure of educational success.
Specific questions: The high-level question/s need to be translated into specific questions that can be directly answered by the data analysis. The specific questions then determine which analyses will be used.	Having posed a broad question, the next step is to break it down into parts that can be directly answered by analysing specific data.
Analyses: At this stage the data are examined, organised and analysed in order to answer the specific question/s framed in the previous stage.	The next step is to plan actual data comparisons based on the more refined questions. Examples of this are shown in Table 2 below.
Interpretation: How do the results of the analyses reflect upon the specific question/s? How do these in turn reflect upon the high-level questions? Are there any aspects of the analyses that make interpretation problematic?	When interpreting an analysis, it is important to remember that while statistics can be useful in determining patterns and trends in assessment performance within schools and departments, a statistical analysis in isolation is not very useful. An analysis needs to be interpreted in conjunction with the professional judgment of teachers, Heads of Department, Principals and Principals’ Nominees. While assessment data provide a quantitative measure of educational outcomes, there are many more subjective measures that need to be taken into account, such as the extent to which students are engaged in a subject by a particular teacher, or the fact that the first job of a teacher is to deliver a curriculum rather than to teach to assessment.

Table 2.2 below outlines specific questions that may be used for data analysis. Notes are provided for each analysis, including potential pitfalls and ways of using the results to improve educational processes.

Please note that Table 2.2 provides an illustration of an approach to data analysis that can be used for a topic of interest. It does not provide an exhaustive list of specific

questions or analyses that could be conducted, or address all interpretative issues that might arise.

Table 2.2 Some specific questions, analyses, and interpretive notes addressing the broad question ‘how can educational processes in a department be improved?’

Specific questions	Analyses and data comparisons	Notes
Does my department have specific areas of strength and weakness in assessment outcomes?	Compare success rates in individual achievement standards with aggregated figures from similar schools (e.g. decile band, single-sex boys/girls or co-educational).	If standards taught in schools have either a positive or negative difference from comparison data that are larger or in a different direction to most of the standards for the subject, then the material taught for those standards may indicate departmental areas of strength and/or weakness. NOTE: It is necessary to take into account any differences in the average ability levels of students undertaking particular standards that may not be evident at other schools.
Are assessment outcomes in my school / department improving over time, declining over time or remaining stable?	Monitor success rates in a set of standards over successive years.	NOTE: Several years’ data are required to identify trends. Even quite substantial changes from one year to the next may not indicate a trend with any particular underlying cause. It is also necessary to take into account any revisions of the standards of interest, and any change in the way in which a standard is applied in assessment.
Do some teachers (in the department) have successful teaching strategies which enable their classes to achieve consistently better assessment outcomes?	Over a five-year period, compare results from each teacher to determine whether any teachers’ classes show consistently better assessment performance.	If particularly effective teachers can be identified, it might be possible to share their successful teaching strategies with other teachers. NOTE: It is necessary to monitor trends over time as a single year’s data is insufficient to form an accurate picture of an individual teacher.
Are some teachers better with some groups of students, and others with other groups of students (e.g. male and female students)?	Similar analysis can be conducted, breaking down the results by any student demographics of interest (e.g. gender).	NOTE: Any systematic differences between the abilities of students typically taught by particular teachers would clearly need to be taken into account (e.g. one teacher might be responsible for students who are struggling with a subject).

Getting the most out of statistics: A guide for users

Part 3 – Different ways of comparing data

This section describes different ways of comparing data. Table 3.1 below summarises some of the main kinds of analyses to consider. The focus is on the uses and advantages of each kind of comparison, as well as factors that potentially complicate interpretation.

Please note that:

- the analyses presented here do not constitute an exhaustive list
- the final analysis chosen is dependent on the question/s to be answered
- some analyses may contain elements of more than one of the types described here e.g. an intra-subject analysis might also include a longitudinal component.

Table 3.1 Examples of questions and analyses that can be used with NCEA/NQF data and statistics

Type of question	Type of analysis
How are patterns of results changing over time?	Longitudinal analysis
What are the particular areas of strength and/or weaknesses in a teaching programme?	Intra-subject analysis
How can the particular strengths of different teachers in a department or school be used to assist one another to improve teaching practice?	Inter-staff analysis
What are the particular areas of academic strength and/or weakness within a school?	Inter-department analysis
How does a school compare with other similar schools in the performance of its students on the NQF?	Inter-school analysis

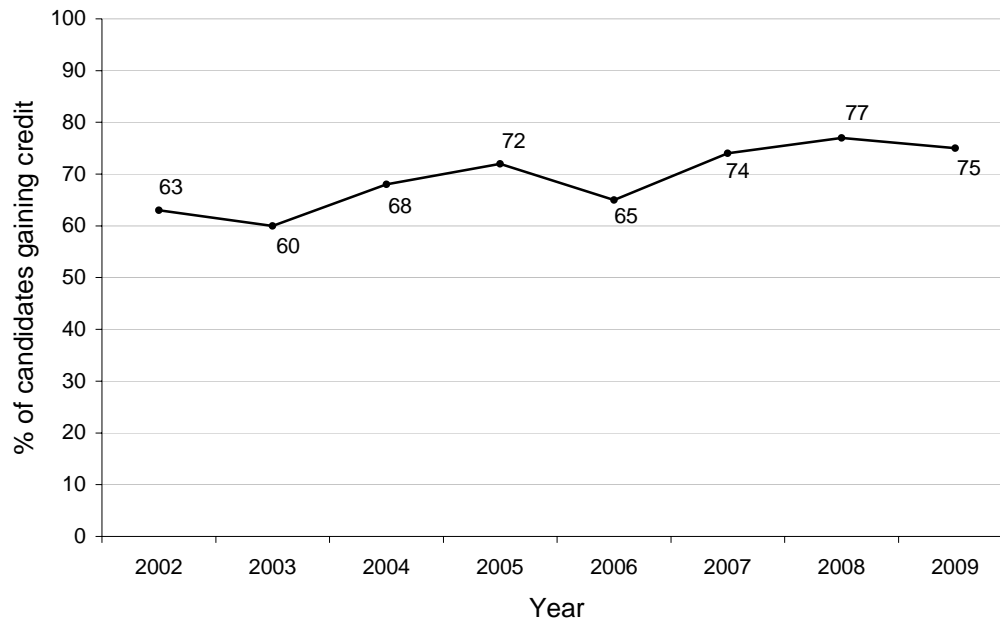
Longitudinal analyses

One of the most useful analyses for tracking results is *longitudinal*; that is, an analysis of how assessment results change over time. Currently, there are not enough years of NCEA data available to make such an analysis particularly meaningful. Longitudinal analyses will become more useful, however, as time goes on. The reason that a number of years of data are needed to make sense of a longitudinal analysis is that, particularly in analyses with relatively small numbers of results (e.g. at the level of a single department), numbers can be expected to fluctuate because of statistical measurement error (see part 5), which has nothing to do with the quality of the teaching or educational programmes. Even for analyses involving large numbers of students, some variability due to measurement error is to be expected.

Longitudinal analysis example

Figure 3a shows the proportions of candidates at a hypothetical school achieving credit in an externally assessed standard over eight successive years.

Figure 3a Hypothetical data showing variability in the percentages of candidates achieving credit in a standard over eight successive years.



Note that while the results shown in Figure 3a fluctuate from one year to the next, over the full time period, an upward trend of an increase in success is evident. For this reason, changes in a results profile in a single year, in either direction, should not be a cause for either concern or celebration. A particularly *large* change in a single year, however, may be a signal that something affecting assessment results in an important way has changed e.g. a high turnover of teaching staff in a particular department, or the revision of one or more standards.

The appropriate basis for longitudinal analysis is the performance of the same item (i.e. a department or school) in the past. This is one of the advantages of longitudinal analyses; that it is self-referenced and does not require a comparison with data from other departments or schools. Interpreting inter-school or inter-subject comparisons can be problematic as the items being compared are often not comparable. While changes in teaching staff or the assessments themselves can still create some interpretative complication, a comparison of a department or school with itself over time is generally more able to be interpreted and more useful than a comparison of one department or school with another.

A candidates' performance in a norm-referenced assessment is always measured relative to the performance of other candidates. So, if there were a change in the strength of a cohort from one year to the next under normative assessment, it would be unlikely to show up in the final results data as results would be scaled to ensure the distributions matched for the two years i.e. any inter-year variability would be masked by scaling procedures. Under standards based assessment, however, candidates are assessed relative to a standard and not relative to one another. For this reason, real

changes in overall performance can be reflected in results data and a longitudinal analysis can highlight these changes.

Intra-subject (between standards) analyses

It can be useful, particularly for developing teaching programmes, to track differences in results between standards within a particular subject area. Such analyses can shed light on specific strengths and weaknesses within a particular department. When performing analyses of this kind, however, be aware of the following:

- 1 Analyses of differences between standards within a school typically involve comparatively few students' results. Statistics based on low numbers of data are prone to high variability, and so are not reliable estimates of an actual situation.

The most common, everyday example that illustrates this is a political poll. Any political poll has a margin of error associated with it, which means that it is highly likely (usually 95% likely) that the actual profile of the population will be within the margin of error relative to the poll result. So, if 42% of respondents to a poll say that they will vote for a particular party, and the margin of error for the poll is 3%, then the actual proportion of the population intending to vote for that party is 95% likely to be between 39% and 45%.

Note that a margin of error is actually the same thing as a confidence interval. A confidence interval is relative to the number of results analysed, with low numbers producing a wider confidence interval. An individual class can therefore be expected to have a high margin of error associated with its data, so that a single year of results data from a particular class may not accurately reflect the quality of teaching in that class. A handful of especially strong or weak students in a particular year will make a substantial difference to the results profile of the class – in a class of 25 students, for example, a single student constitutes 4% of the class, so if three students were to move from not achieving to achieving, the results profile would shift by 12 percentage points.

- 2 Results for unit and achievement standards (particularly external achievement standards) cannot be compared for two reasons. First, unit standards do not have differential grades; there are no Merit or Excellence grades available for them. Second and more seriously, results for internally assessed standards are only recorded in NZQA databases for students who have *achieved* a standard. There is no record of those who have attempted a standard but not achieved it. Because of this limitation in the data, unit standard statistics from the NZQA website cannot be used meaningfully in any analysis except longitudinal comparisons within schools, unless schools keep their own records for unit standards that include students who have attempted but not achieved in these standards.
- 3 If there appears to be a gap in performance at a school or in a department between two standards, this may not reflect a weakness in the learning programme with respect to the standard in which performance is poorer (nor, necessarily, a strength with respect to the standard in which performance is

stronger). To address the question of whether the performance profiles for a given two standards reflect any noteworthy issues for a school with respect to its teaching programmes, it is necessary to reference those profiles to the similar profiles in the national statistics. This can be done by referring to the NQF statistics, *results distributions by learning area* on the NZQA statistics website, on which the overall (national) percentages of students receiving grades of *Not Achieved*, *Achieved*, *Merit*, and *Excellence*, for every externally assessed achievement standard are shown.

Intra-subject analyses example

Consider the Level 1 English standards 90053 (*Produce formal writing*) and 90054 (*Read, study and show understanding of extended written text*). The national results profiles for these standards in 2004 are shown in Table 2. Note that the percentages of students receiving *Merit* and *Excellence* grades in each standard are very similar, whereas there is a substantially greater percentage of *Achieved* results in 90054 than in 90053, and a commensurately lesser percentage of *Not Achieved* results. These differences between standards need to be taken into consideration in any comparison of results from these standards within a school.

Table 3.2 National percentages of students in each grade category for English standards 90053 and 90054 in 2004

	Not Achieved	Achieved	Merit	Excellence
Standard 90053 <i>Produce formal writing</i>	44.7%	38.8%	13.9%	2.5%
Standard 90054 <i>Read, study and show understanding of extended written text(s)</i>	37.6%	47.0%	12.5%	2.9%

In a hypothetical Year 11 English class, 79% of the students receive a grade of *Achieved* or better in standard 90053, compared with 68% for standard 90054. Nationally, around 55% of students received a grade of *Achieved* or better in 90053, compared with 63% in standard 90054. What can the teacher of this class conclude from this? First, the students seem to be achieving these standards at a somewhat higher rate than the national average. However, before celebrating this apparent success, two things need to be considered.

Remember that high variability is to be expected in statistics based on small numbers of cases. While the national data are based on a very large number of cases, class data are not. So, the apparently high rate of success for this class may actually be a chance effect. The class may just happen to be particularly academically strong in that year, or the exam may have happened to coincide especially well with work that they had completed close to the exam time. One way to test these possibilities is to compare data for a number of successive years (longitudinal data), and if this shows that the

teacher's classes perform *consistently* above the national average, it would be unlikely to be a statistical anomaly. Another way to test the possibilities is to calculate confidence intervals for the estimates under comparison. If the confidence intervals do not overlap, the observed difference is unlikely to be attributable to measurement error.

Teachers' professional judgment can also shed light on these issues, and consultation with colleagues over these kinds of analyses is encouraged. (In fact, professional judgment should always be used in interpreting any statistical analysis). Another consideration is the demographic circumstances of the school. Setting aside the issues of interpreting the *absolute* difference between the national results and the performance of the hypothetical English class, the analysis of the data shown in figure 1 was actually prompted by the question of what was learnt from the *difference in the results* profiles for the two standards about the strengths/weaknesses in a teaching programme. The national statistics suggest that 90053 is a slightly more difficult standard to meet than 90054, with grades of Achieved or better around eight percentage points more common for the latter. One issue to take into account here is whether the national cohorts undertaking these two standards are likely to be significantly different. Again, professional judgment is probably the best measure of this, but it is perhaps unlikely to be the case for two Level 1 English achievement standards.

Whereas in the national data, 90053 seems to be the more difficult of the two standards, in the hypothetical class results, there was a considerably higher success rate for 90054. As for the absolute comparison, the small size of the data set for the hypothetical class needs to be considered, and similar data from consecutive years would strengthen this conclusion. But on the face of it, the data for the hypothetical class suggest that the teacher has a particularly effective programme for reading and reading comprehension.

In the above example, just two standards were compared. In a real analysis, teachers would probably want to compare performance across all of the standards that they had taught in a particular course, or to a particular class. All of the same caveats would apply when making these comparisons, although an analysis of multiple standards would be likely to be more informative in many ways, because it would be possible to determine whether there is a common theme in standards that stand out as having particularly strong or weak rates of success.

Inter-staff analyses

Heads of department may wish to track the performance of individual teachers in order to improve the overall performance in their departments. It goes without saying that such analyses need to be approached with sensitivity and it is suggested that, if analyses such as this are carried out, results are communicated with a clear emphasis on teachers assisting one another to improve in a collegial environment. Furthermore, it is likely that some teachers will excel with some student groups or with particular curriculum material, and other teachers will excel with different groups of students or other material.

To have any validity, a comparison of teaching performance really needs to be carried out over time. As noted above, a single year's data from a single class does not give a reliable indication of the true nature of the teaching in that class.

Where classes are to some extent streamed, or grouped according to ability, a comparison of teachers' performance becomes next to impossible. Although it would be expected that a strong class would achieve a higher success rate than a weaker class, regardless of any differences in teaching effectiveness, to confirm this would require a reliable way of determining the expected results for each class, taking into account the difference in ability and there is no straightforward way of doing this. Furthermore, it might often be the case that classes of students with differing levels of ability undertake a somewhat different mix of standards in a given subject area. Standards, even within a subject and level, are not homogeneous in difficulty and, again, there is no straightforward way to correct for any differences in standard difficulty in any comparison of data for different classes.

Inter-department analyses

It may be of interest to a school to identify subject areas or departments of particular strength or weakness. As with comparisons of individual standards, a departmental comparison needs to take into account the different rates of achievement that are expected in the different subject areas. There are two reasons for this:

- 1 Any differences in standard difficulty are more likely to be evident across subjects than within subjects (although a range of difficulty is to be expected within subject as well). It is also likely that some subjects are intrinsically more difficult than others for the majority of students.
- 2 Some departments and subject areas tend to attract students of higher ability than others. Clearly these two factors will interact; a difficult standard undertaken by a strong cohort may be expected to have a similar results profile to a less difficult standard undertaken by an average cohort.

One way to test this hypothesis is to compare differences in standard-aggregated results for two departments, and to determine whether or not this difference is greater than, less than, or about equal to the difference expected on the basis of the national data, or preferably a subset of national data drawn from schools with similar demographic characteristics to your own (e.g. mid-decile co-ed). The following example compares just two subjects. In an actual analysis, it may be more appropriate to take into account a broad range of departments.

Inter-department example

In this example, the relative performance in Level 3 standards of two departments, Geography and Mathematics, are compared for a hypothetical school. Table 3 shows the numbers of results in each grade category for each Level 3 standard in Geography, and each Level 3 standard in Mathematics (excluding probability and statistics standards). The *Total* rows in the table show all of the results in each grade category

across all standards in both subjects. The *subject aggregate* rows show the percentage of the total results for each subject falling into each grade category. It seems clear from these data that the success rate in Level 3 Mathematics at this school is higher than that in Level 3 Geography, with 71% of results in mathematics gaining credit compared with 55% for Geography. What is not clear from these data, is why this is so. Is it because the teaching in Mathematics is superior at this school? Is it because the Level 3 Geography standards are more challenging than the Level 3 Mathematics standards? Is it because the Mathematics cohort is more able?

Table 3.3 Results for a hypothetical school in Level 3 Geography and Mathematics

Geography					
Standard	Total number of results	Number of results in each grade category			
		Not Achieved	Achieved	Merit	Excellence
90701	45	21	14	7	3
90702	30	18	8	3	1
90704	57	20	25	10	2
<i>Total</i>	<i>132</i>	<i>59</i>	<i>47</i>	<i>20</i>	<i>6</i>
<i>Subject aggregate (%)</i>		45%	36%	15%	4%
Mathematics					
Standard	Total number of results	Number of results in each grade category			
		Not Achieved	Achieved	Merit	Excellence
90638	84	23	41	14	6
90639	90	30	42	10	8
90644	57	15	22	14	6
90635	69	12	31	18	8
90636	75	27	31	14	3
Total	375	107	167	70	31
Subject aggregate (%)		29%	44%	19%	8%

To answer the question of why the success rate in Level 3 Mathematics is higher than that in Level 3 Geography at this school, it is necessary to compare the school's results in these subjects with those of similar schools nationally. Assume that this hypothetical school is in the mid-decile range. Table 4 shows the numbers of results for each Level 3 Geography and Mathematics standard gained by students at Decile 4–7 schools nationally, as well as the percentages of these results in each of the four grade categories. These are real data on the NZQA website from the 2004 external assessment round².

² <http://www.nzqa.govt.nz/qualifications/ssq/statistics/natl-nqf.do?statsYear=2004#t>

Before it is possible to directly compare these data with the data for the hypothetical school, it is necessary to aggregate the percentages in each grade category across the standards in each subject. The aggregate percentages are shown in the bottom row for each subject.

Table 3.4 Numbers of results and percentage of results in each grade category for each externally assessed standard at Level 3 in Geography and Mathematics, across all Decile 4–7 secondary schools

Geography					
Standard	Total number of results	Percentage of results in each grade category			
		Not Achieved	Achieved	Merit	Excellence
90701	2,256	55%	34%	8%	3%
90702	2,306	57%	30%	11%	2%
90704	2,335	30%	43%	19%	8%
<i>Subject aggregate (%)</i>		<i>48%</i>	<i>36%</i>	<i>12%</i>	<i>4%</i>
Mathematics					
Standard	Total number of results	Percentage of results in each grade category			
		Not Achieved	Achieved	Merit	Excellence
90638	3,011	40%	41%	16%	3%
90639	2,882	45%	43%	11%	1%
90644	4,369	30%	52%	14%	4%
90635	3,050	33%	43%	20%	4%
90636	3,039	47%	39%	14%	0%
Subject aggregate (%)		39%	43%	15%	3%

Compare the subject aggregate percentages of results for the hypothetical school (Table 3) with those for all Decile 4–7 schools (Table 4). For Geography, the percentages are very similar – identical for the *Achieved* and *Excellence* categories, and differing by just three percentage points for each of the *Not Achieved* and *Merit* categories. These small differences are well within the margin of error (95% confidence interval) for the estimates of the performance in Geography at the hypothetical school; the estimate of the *Not Achieved* rate, for example, has a margin of error of +/-8% (check this using the formula for confidence intervals, given in Part 5), so that the confidence interval for this estimate is 37% – 53%. Thus the results here give no reason to suspect that the geography department at the hypothetical school is any stronger or weaker than average.

For Mathematics, the situation is a bit different. The percentage of students gaining credit at the hypothetical school is 10 points higher than it is for the Decile 4–7 aggregate. The margin of error for this estimate is narrower, because the sample is larger; there are 375 mathematics students at the hypothetical school, compared with 132 geography students. The 95% confidence interval for the success rate in mathematics turns out to be 24% – 34%, the upper limit of which is still below the *Not Achieved* rate for the Decile 4–7 aggregate (39%). So we can be over 95% confident that the success rate in Mathematics standards at the hypothetical school is higher than would be expected based on the school’s demographics. The conclusion of this analysis is probably that the school has an excellent Mathematics department, with the only real alternative being that, for some reason, the students at this school are particularly mathematically gifted.

Inter-school analyses

Many schools use their NQF/NCEA statistics to compare themselves with other schools in their local area, with the national statistics, or with a subset of the national statistics such as other schools of the same decile rating. Conclusions from such comparisons should be drawn with a great deal of care. Wrong conclusions can be serious and damaging in terms of the prestige and public understanding of a school.

The league tables produced by the media each year when the NQF/NCEA results are published are essentially meaningless when it comes to comparing schools’ performance, because there are a great many factors beyond the control of a school that affect outcomes in assessment.

Inter-school example

Consider the data presented in Table 5, which shows that success rates in Level 3 NCEA are very much affected by the demographic characteristics of a school. (In fact, similar disparities are evident in almost all secondary school assessments.) The range of success in Level 3 NCEA makes clear just how profound the effects of demographics are, from low decile, single-sex boys’ schools with a Year 13 achievement rate of around 20%, to high decile, single-sex girls’ schools, with a Year 13 achievement rate above 75%.

Table 3.5 Percentages of Year 13 students at schools with various demographic classifications achieving Level 3 NCEA in 2004

	Co-ed schools	Single sex boys’ schools	Single sex girls’ schools
Decile 1 – 3	28.8	19.7	34.5
Decile 4 – 7	44.2	47.3	56.7
Decile 8 – 10	57.5	55.9	75.5

The point clearly made by the data presented in Table 5, is that any comparison between schools at least needs to take into account the decile rating of the schools, as well as their co-educational or single sex status. A comparison with average statistics

for schools with similar demographics will be far more meaningful than a comparison with national averages. Consider, for example, two schools, one a Decile 10 single-sex girls' school with a Year 13 success rate in Level 3 NCEA of 67%, and a Decile 1 single-sex boys' school with a success rate of 30%. In a league table, the former school would look like it is doing a much better job for its students than the latter, particularly if the reader was ignorant of the demographic differences. In the context of the data in Table 5, however, it is clear that the reverse is true; students at the low decile boy's school are achieving at much higher rates than expected for a school of that type, whereas those at the high decile girl's school have a much poorer rate of success than students at other similar schools.

Even a comparison of schools with similar decile and gender characteristics has its difficulties. Demographic variables do not tell the full story about a school's circumstances, and there are many other important predictors of success in assessment that are outside of a school's control such as its geographic location and its ethnic and cultural mix. The decile statistic itself is problematic because it accounts only for the proportion of students in a school that is likely to be below the poverty line. It does not describe the *distribution* of economic circumstances within a school, so two schools with the same decile ratings might actually have quite different socioeconomic profiles.

Another approach would be for a school to choose a group of other schools with similar demographics with which to compare its results. Often, teachers and principals have a better idea of which other schools have similar circumstances to their own, than can be provided through official decile ratings or other administrative data.

Getting the most out of statistics: A guide for users

Part 4 – Creating a Report

This section provides some general guidelines for structuring a report and ways of discussing analyses. Whenever data analysis is carried out for reporting purposes (e.g. to a school board or to the Education Review Office), it will be necessary to organise it into a report. Even when analyses are performed by an individual teacher or head of department for their own uses, writing the results into a brief report can be useful. A report provides a focus and a format in which various analyses and data comparisons can be discussed. Even if a set of analyses is carried out primarily for the benefit of the person performing them, it may also be of interest to others, and a well considered report is more valuable than a set of apparently disconnected graphs and tables. In addition to making the reason for, and results of, an analysis explicit to a reader, writing a report can also clarify the findings for the author. Keeping a report brief and to the point is important. The exact form that a report will take depends on its purpose and the types of analyses performed.

The sample report at the end of this section gives a few ideas for structure and content – it is not intended to be a definitive model for the creation of a report.

Introducing the report

A report should include a brief introduction that clearly states the high-level question or questions to be addressed, the reason for asking these questions and a brief description of how the questions are to be addressed analytically. This helps people to make sense of the statistics.

If the report is being prepared for an authority (such as a school board), a summary of the findings might also be useful. A summary would typically be placed at either the very beginning or very end of a report.

Visual presentation of data

The main body of a report should contain tables or graphs/histograms containing the data being reported. Each graph or table should be as simple as possible. Too many numbers in one illustration can be off-putting, especially for people who are not familiar with reading data analyses.

It is a good idea to write a brief interpretation after each table or graph to highlight points of interest and, in particular, how the data reflect upon the questions asked. Such interpretations should be focused on the data at hand, whilst the end of a report is more appropriate for wide-ranging discussion.

Any formal statistical analyses (eg a chi-square test) should also be reported in the text and related to the data.

Discussion of data

A good report contains a clear explanation of the statistical analyses. Many people are not statistical experts, so it is important to include any caveats or limitations on interpretation that apply. Examples of potential limitations have already been discussed throughout the guide, but a summary of points to consider, and the circumstances in which each caveat might apply is given in Table 1.

Table 4.1 Some limitations and caveats to bear in mind when interpreting analyses

Type of analysis	Limitations / caveats
Comparing success rates of students at different schools in external assessments to determine which school is the most successful.	Schools vary greatly in their demographic characteristics. Demographic characteristics influence performance in assessments. A comparison of schools is therefore meaningless without taking careful account of demographic differences between them.
Comparing success rates of students studying different courses (at the same school) to determine which department is the most successful.	In addition to demographic differences between students undertaking various courses, some subject areas are on average, more challenging than others. The relative difficulty of assessments therefore needs to be taken into account in any comparison of this kind.
Comparing rates of qualification acquisition for, say, Year 12 students, in successive years, to determine whether their success rate is increasing, decreasing, or stable.	Changes from year to year, especially small ones, do not necessarily indicate a trend. Small changes due to natural variability are to be expected. Also, even quite large changes from one year to the next might be one-off effects, rather than indicating a long-term trend.

This is not an exhaustive list, but is intended to provide some examples of critical thinking about analyses and comparisons of data. Finally, it is important to note any conclusions from the data. The main conclusions should relate to the question or questions that prompted the analysis, but any other interesting observations are also worth reporting.

Getting the most out of 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 cases³. 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 standards, we divide 680,838 by 1,072,837, which gives 0.635. Multiplying this

³ 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.

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.

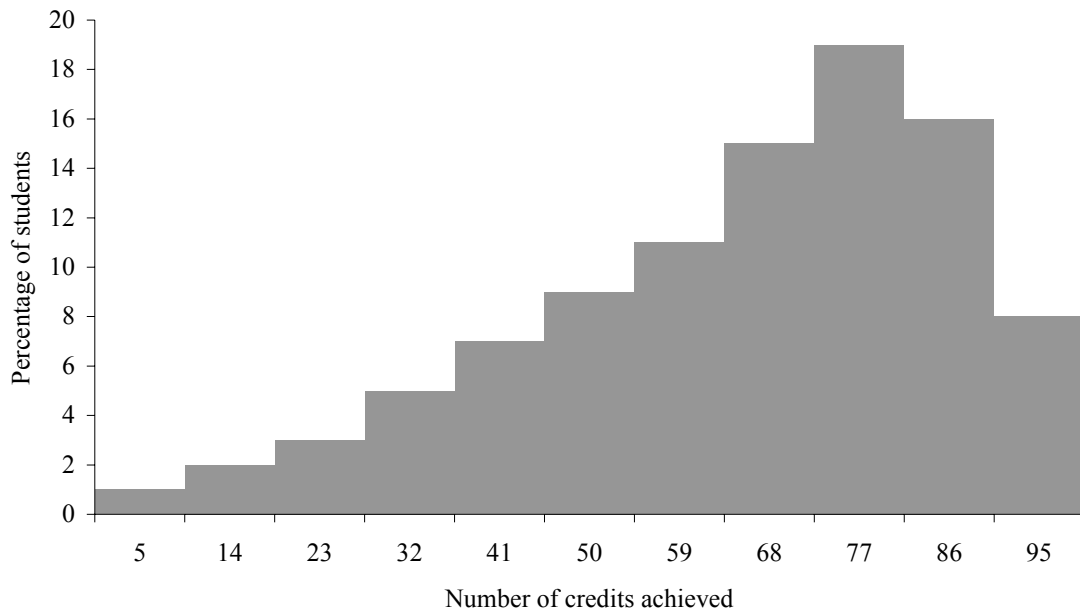
When reporting percentage data, it is important to make it clear what a percentage is a percentage *of*. It is not useful, for example, to say that a particular school has a success rate of 70% in Level 2 NCEA. Does this mean that 70% of *all students* at this school *leave* with Level 2 NCEA? Does it 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, a percentage is more difficult to interpret, leading to potential 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 totalling 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).

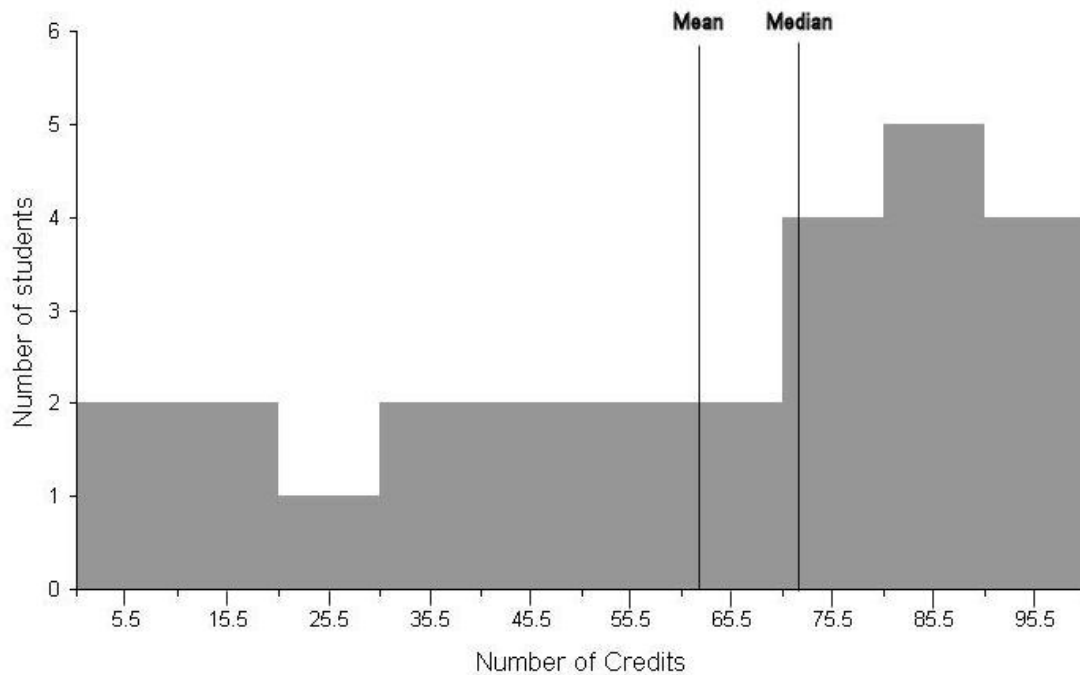
The reason that the mean and median of an asymmetrical distribution are different is that in determining the median, only the rank order of the values is taken into account, whereas in determining the mean, the actual values themselves are used. When extreme values are exactly balanced, as they are in 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 (ie 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.

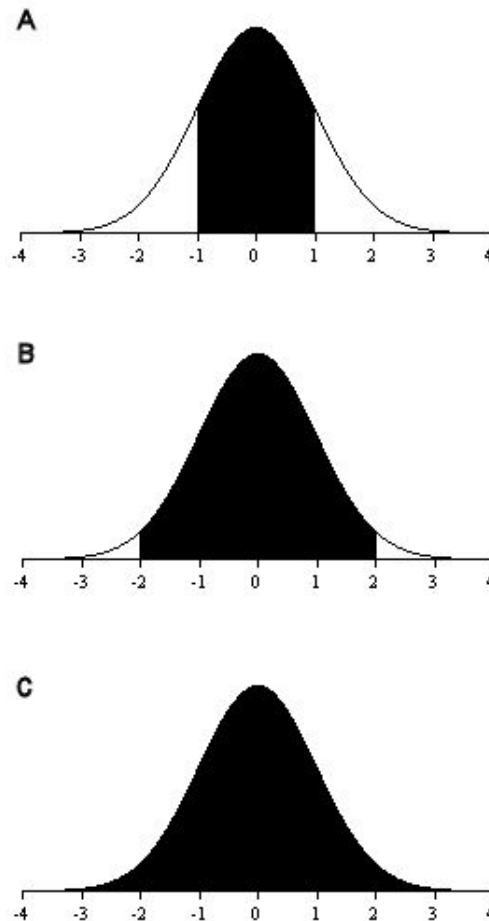
Formula 1 describes how to calculate the standard deviation, however, it is more important to understand what it is, rather than the details for calculating it.

Formula 1. The standard deviation

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

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 influences on student performance that are systematically stronger in some years than others e.g. a high teacher turnover or a 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 cancelling 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)	Three year aggregate
90054 <i>Read, study and show understanding of extended written text(s)</i>	55%	63%	59%	61%
90055 <i>Read, study and show understanding of a number of short written texts</i>	61%	59%	64%	62%
90056 <i>View/listen to, study and show understanding of a visual or oral text</i>	62%	53%	54%	57%

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

$$1.96 \times \sqrt{p(1-p) \div n}$$

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.