

Mathematics



CS

Questioning assumptions

“When I use a word,” Humpty-Dumpty said, “it means just what I choose it to mean — neither more nor less.”

Lewis Carroll
*Through the Looking-glass,
and What Alice Found There.*

Using appropriate samples in statistics is crucial. The assumption that a sample is representative or of an appropriate size must often be questioned. Jane Watson investigated a claim made by a seafood company that “Seven in ten men who frequently eat canned tuna, sardines, salmon, mackerel or kippers admit to being ambitious”. The implication of this claim was that a diet of fish increased ambitiousness. Although the study, commissioned by John West Foods, used a total sample of 250 people, only 2.4 per cent were very frequent male consumers. Jane found that the claim was based on a real sample size of only six men (who all ate fish frequently), of whom four (or 66 per cent — rounded to 70 per cent) considered themselves more ambitious than their colleagues. Although the claim was true of this section of the sample, it would be misleading to suggest that the sample represented the entire population of fish-eating men in the country. The use of a proportional statement where the components of the proportion, “seven in ten men”, is bigger than the actual sample size is also a concern.

The way we describe proportion often implies a sense of sample size. Percentages imply large numbers. Using 80 per cent rather than 4 out of 5, when the sample size is 5, produces the illusion of a much larger sample. Excessive precision in measurements has a similar effect. Stating that the atmospheric carbon dioxide concentration has increased by 25.8324 per cent since 1850 introduces spurious precision. Recognising that the carbon dioxide concentration fluctuates on a daily and seasonal basis, the increase would more reasonably be reported as between 20 and 30 per cent.

Sometimes, instead of the numbers that are reported, it is the numbers that aren't presented that cause confusion. Ross Gittins provided an example of this in the *Sydney Morning Herald* of 4 February 1998. At the time, the youth unemployment rate was reported to be running at 28 per cent. The implication was that

more than a quarter of Australia's young was on the dole.

Examining the assumptions, the first thing to determine is the definition of “youth”. In this case it appears to mean people aged over 15 and under 20. According to Gittins, there are about 1.05 million people in this age group in Australia. Of these, 740 000 are still in the education system, 223 000 are in the full-time work force and 86 000 are unemployed. The 86 000 unemployed represent 28 per cent of the teenage workforce but only about 8 per cent of the total population in that age group.

Other examples of the need to question the assumptions behind statistical models are provided at the Nova Internet site
<http://www.science.org.au/nova>.

Sampling and variability

Understanding sampling and variability provides a basis for using statistics. We are very rarely interested only in what is happening in a sample. We are interested in a sample only to the extent it can provide information about the population. Consequently, we generally aim to make our sample as representative of the population as we can.

The General Mathematics syllabus outlines three methods to create representative samples.

1. Simple **random** sampling selects members from the population in a way that ensures that each member of the population has an equal chance of being selected. The selection of one member should also be independent of the selection of any other member of the population. However, random sampling does not guarantee that a particular sample will be exactly representative of a population.

Imagine that you had a large population, of which 50 per cent was male and 50 per cent was female. If you select a sample of four people you could have the entire sample of the same sex. Sample size clearly has an effect on how well the sample can predict the characteristics of the population.

2. **Stratified** random sampling is a process designed to produce a sample that reflects the proportions of various subgroups within the population. For example, if we wanted to select a sample in a high school of 600 students that had 150 students in Year 7 and 60 students in Year 11, a stratified random sample would need to have the same proportions.



Mathematics

3. In **systematic** sampling you select your sample by choosing every *nth* member from a list that includes all members of the population. This method is often used in quality control. If you had a list of 600 people and you wanted a sample size of 30, you could start by randomly selecting one of the first 20 names and then taking every 20th name from that point on.

Probability plays a significant role in using information obtained from the sample to predict features of the population. Also, probability is useful in defining the limitations of inferring characteristics of the population from the sample. If we revisit the problem of determining the distribution of gender in a population, we will obtain a clearer picture of how this works.

Starting with a sample size of two, we have three possible outcomes: two female; one female and one male; and two male. If gender is equally distributed in the population, the probability of randomly selecting two females from the population is $\frac{1}{4}$. If we use a sample size of two to predict the distribution of gender in the population, we would expect to predict an all-female population 25% of the time.

Increasing the sample size to six, the probability of predicting an all-female population is $\frac{1}{64}$ or less than 2%.

Another simple example using gender, drawn from the research on statistical reasoning (Nisbett et al., 1987) considers the ratio of births of boys to girls in a hospital. In a large hospital, imagine 120 babies are born every day. A smaller hospital nearby has only 12 births a day. On average, the ratio of boys to girls in each hospital is 1:1. One day, in one of the two hospitals, twice as many baby girls were born as baby boys. In which hospital was it more likely to happen?

Although the answer appears obvious, it highlights the importance of understanding variability in samples. Using the computer program ProbSim™ to simulate the problem, a result of 4 boys and 8 girls occurred on the second trial with the smaller hospital.

Event (U)	Count	Proportion
M	4	0.33
F	8	0.67

A result of 40 boys and 80 girls did not occur in 50 simulations with the larger hospital. The result is much more likely to happen in the small hospital. The probability of a random deviation from the population mean decreases with the increase in sample size.

In the General Mathematics Preliminary course unit on data collection and sampling (DA2), students are expected to recognise the effect of sample size on estimating the nature of the population. One of the suggested applications is to select a range of samples from a fixed population and to record the characteristics of each sample. It is possible to undertake this task using objects selected from a bag that contains equal numbers of counters of two colours. The prediction of what is in the bag (see **CURRICULUM SUPPORT**, Vol. 4, No. 1, p. 2) is a useful introduction to using sampling as a means of predicting the characteristics of a population. Computer simulation provides a useful method of generating the number of samples necessary to explore the effect of the size of the sample on the predicted characteristics of the population.

The speed with which computers and calculators can run simulations assists the investigation of quite complex statistical concepts. It has given rise to the popularity of Monte Carlo experiments to empirically demonstrate many results that do not yet have theoretical proofs. The term *empirical* in statistics means *observable* or *observed*.

Extending the hospital problem

If we consider the hospital problem further, how common is it to obtain a result of twice as many baby girls being born as baby boys for different sized samples? That is, when drawing samples from a population with equal numbers of boys and girls, what is the probability of selecting a sample that contains twice as many girls as boys? What happens to this probability if we consider sample sizes of 6, 12, 18, 24 and 30?

Using binomial probability, the chance of twice as many baby girls with a sample size of six is

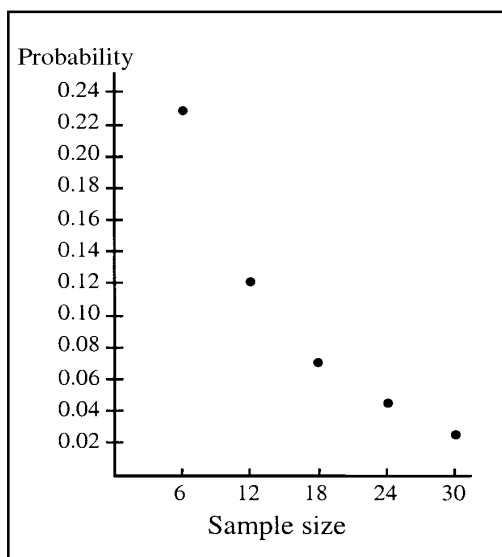
${}^6C_2\left(\frac{1}{2}\right)^6$ or $\frac{15}{64}$. Although students in the General

Mathematics course are not expected to use binomial probability, they could determine an approximation of this and similar results empirically from a simulation program, actually drawing samples of different sizes from a population.

Mathematics

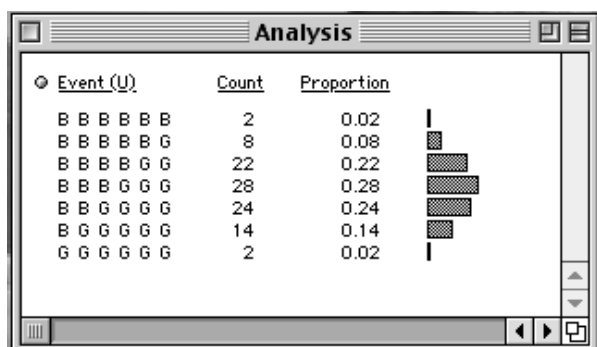


Continuing the examination of the effect of sample size on the probability of twice as many baby girls as baby boys being born, the theoretical results are as follows. Using samples of 12, 18, 24 and 30 produces probabilities of ${}^{12}C_4(\frac{1}{2})^{12}$, ${}^{18}C_6(\frac{1}{2})^{18}$, ${}^{24}C_8(\frac{1}{2})^{24}$ and ${}^{30}C_{10}(\frac{1}{2})^{30}$. If we plot these probabilities as a function of sample size we see that the probability of such an atypical sample decreases as the sample size increases.



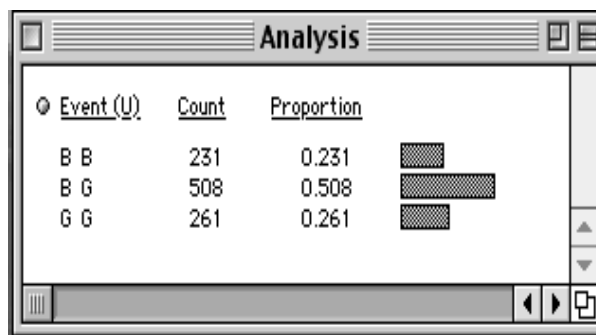
Students in the General Mathematics course can use simulation programs to empirically answer questions such as the following:

If the chances of having a baby girl or a baby boy are equally likely, what is the approximate probability that, if I select a simple random sample of 6 babies, there will be twice as many girls as boys in the sample?

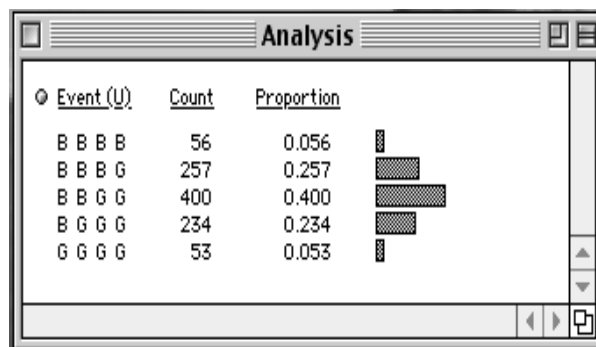


Simulating 100 samples of 6 births produces an experimental probability of 0.24. It is not difficult to understand why such Monte Carlo experimental procedures are commonly used.

We can use simulation to return to the general relationship between the sample size and the likelihood of the sample reflecting the population. Starting with a sample size of 2 and drawing 1000 samples using ProbSim™ produced the following results:



Predicting the nature of the population from the sample size of two suggests that we would be incorrect about half the time. Repeating the process with a sample size of four produced the following results:

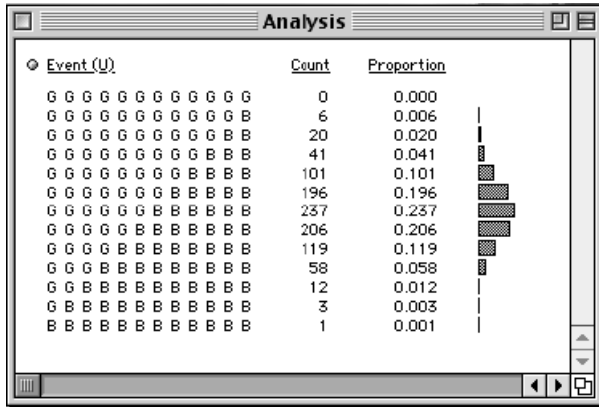


The number of outcomes will increase with increasing sample size. How then can we tell if larger samples are better predictors of the nature of the population? Indeed, the proportion of the 1000 samples that is representative of the population is decreasing as the sample size increases.

The “clustering of the outcomes” around the population mean, however, tends to increase. That is, if we look at the shape of the sampling distribution, its spread appears to decrease as the sampling size increases.



Mathematics



If we are interested in the mean proportion of girls to boys in the sample, these results lead towards the ideas of sampling error and the standard error of the mean. The estimated standard error of the mean is based on the sample standard deviation divided by the square root of the sample size.

CS Mathematics and coincidence

Coincidences intrigue us. We feel compelled to find a way to attach significance to these events. Quite often coincidences should be expected because of the nature of variability of distributions. Particular events only appear improbable after they occur. The chance that some general event will occur which will be characterised as an unlikely coincidence is very likely.

The famous birthday problem from probability theory is an example of how likely generic coincidences can be. In a random gathering of people, you would need to have 367 people present to ensure that 2 of them share a birthday. Yet to have a 50-50 chance of this happening, you need a sample size of only 23 people. Imagine empirically testing the birthday problem in a large high school. Sending a student to each mathematics class containing at least 23 students should result in about half of these classrooms recording the “coincidence” of 2 students sharing a birthday. Of course, with the variability typical of sampling, sometimes this won’t happen!

Whenever coincidences are being discussed, it is important to distinguish between generic events and particular events. For example, the chance of the

particular event that you win the lottery is quite small, whereas the generic outcome, that someone wins the lottery, is not. The same is true of the birthday problem. When all we require is that 2 people have some birthday in common, then 23 people suffice to produce a probability of one-half. In comparison, 253 people are needed in order for the probability to be 0.5 that just one of them has a specific birthday, say, 25 December. The probability that someone does not have a 25 December birthday is $364/365$; the probability that neither of 2 people has a 25 December birthday is $\left(\frac{364}{365}\right)^2$ and none of 253 is $\left(\frac{364}{365}\right)^{253}$, which is approximately equal to 0.4995. The complement of this event, that at least one of the 253 people has a 25 December birthday, is $1 - \left(\frac{364}{365}\right)^{253} \approx 0.5$.

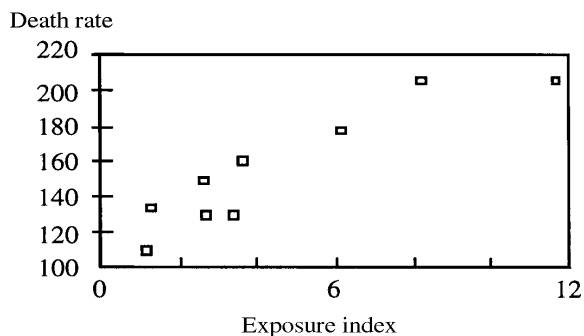
When it comes to considering the mathematics of coincidence it is sensible to heed the words of John Paulos, who wrote in *Beyond Numeracy*: “In reality, the most astonishingly incredible coincidence imaginable would be the complete absence of all coincidences” (1991, p. 41).

CS Fitting a curve to data

When you fit a curve to data, how can you tell if it is a good fit? Most of the points on a scatter plot will not fall exactly on the fitted curve. The question we are interested in is: “How close to the curve are the data points?” Since we often fit curves in order to predict y -values, we usually study the vertical distance between each observed point and the fitted curve. This difference in y -values is called a *residual*.

The *Journal of Environmental Health* in 1965 reported on radioactive waste that had been seeping into the Columbia River. The report provided information on the rate of cancer deaths compared to exposure indices for nine areas around the Columbia River. Presented as a scatter plot, the information was as follows.

Mathematics



To fit a median-median line to the data, follow these steps:

1. Separate the data into three groups of equal size (or as close to equal as possible), according to the values of the horizontal coordinate.
2. Find the summary point for each data group based on the median x -value and the median y -value.
3. Find the equation of the line through the summary points of the outer groups. We will call this line L .
4. Keeping the gradient the same, slide L one-third of the way vertically to the middle summary point.
5. Find the equation of this line.

For the above data the fitted median line is approximately $y = 10x + 112$. This linear equation forms a model for the data. The y -intercept corresponds to the cancer death rate we predict when there is no radioactive contamination, that is, 112 deaths per 100 000. The value of the slope is the amount that y -values change for unit changes in x -values.

How well does the line fit the data? The best way to answer this question is to calculate the residuals.

x	y (data)	y (fit)	Residual = data - fit
1.3	114	126	-12
1.6	138	129	9
2.5	147	138	9
2.6	130	139	-9
3.4	130	148	-18
3.8	162	152	10
6.4	178	179	-1
8.3	210	199	11
11.6	208	233	-25

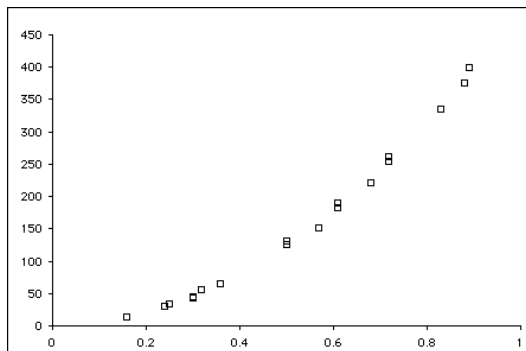
Residuals that are small relative to observed y -values provide evidence of a good fit. If there are many large residuals, we may need to choose a different mathematical function to model the relationship.

The second feature of the residuals that we need to examine is whether they follow any trend or pattern as the x -values vary. Are the residuals in the middle all

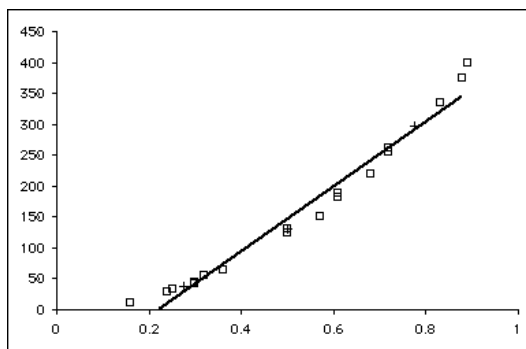
positive or all negative and at the ends of opposite signs? What does this mean?

Often it is easier to analyse the residuals when they are paired with corresponding x -values and studied as a new data set. A scatter plot of this new data set, called a *residual plot*, lets us check the size and pattern of the residuals. If our model provides a good fit for the original data, the residual plot will show points scattered randomly within a horizontal band about the horizontal axis.

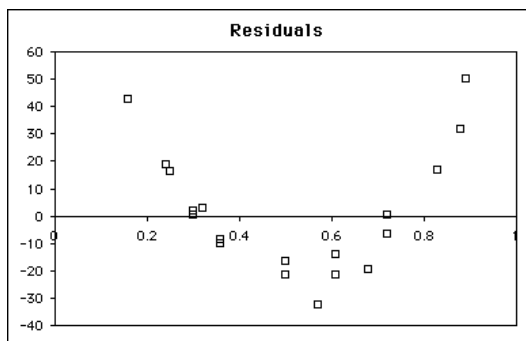
If we fit a straight line to data when the true relationship is not linear, the residual plot will usually show a pattern. We could, for example, fit a median line to the following data.



With the median line superimposed on these data there is an approximately equal distribution of points on either side of the line.



If we now calculate and plot the residuals as a function of the x -values, we notice the following pattern.

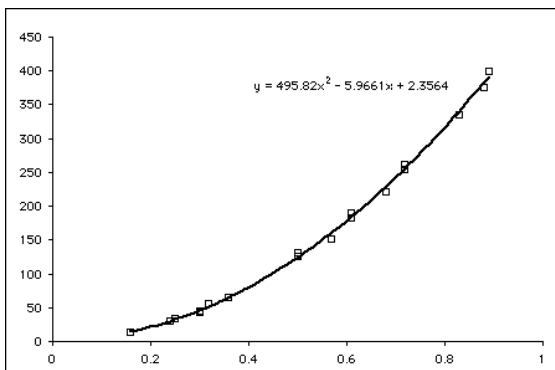




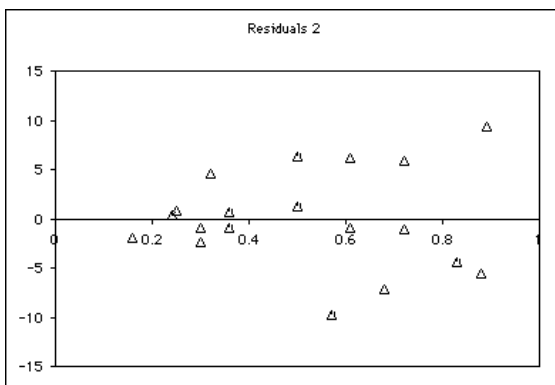
Mathematics

If the residuals were randomly scattered throughout the plot we would consider the line to be a good fit. The residuals, however, show a clear pattern, with each end positive and the middle negative. Indeed, the residuals appear to follow a parabolic distribution.

This suggests that fitting a quadratic curve to the data might be appropriate. That is, the data might be better modelled using a quadratic function. Modelling linear and non-linear relationships (AM4) in the General course clearly benefits from the use of technology.



The above curve was fitted to the data using the trendline function in the Excel spreadsheet program. If we now recalculate the residuals and plot the results we see that, not only are the residuals much smaller, but they are more randomly distributed. There is no longer an obvious pattern in the residuals.



The power of technology and programs such as modern spreadsheets make the process of simple exploratory data analysis accessible to a wide audience. The tedium of the calculations associated with optimising fitting a curve to data means that such processes are products of the technology age.

Curve fitting in mathematical modelling has many applications in the study of science.

The Monte Carlo method

The term “Monte Carlo method” was coined by Metropolis during the development of the atomic bomb in the Manhattan Project of World War II. The similarity of statistical simulation to games of chance led to the method being named after the capital of Monaco, a well-known centre for gambling.

The Monte Carlo method provides approximate solutions to a variety of mathematical problems by performing statistical sampling experiments.

Isolated applications of Monte Carlo methods have been used for centuries. However, only in the past several decades has the technique gained the status of a full-fledged numerical method capable of addressing the most complex applications.

Buffon’s needle problem is an early example of the Monte Carlo method. In the second half of the nineteenth century, a number of people performed experiments, in which they threw a needle in a haphazard manner onto a board ruled with parallel straight lines and inferred the value of π from observations of the number of intersections between needle and lines. Dropping a needle of length l onto a plane ruled with parallel lines distance d apart ($l < d$) results in a theoretical probability of success of $\frac{2l}{\pi d}$.

If the needle is of length 1 unit and a distance of 2 units separates the parallel lines, the proportion of successes will approximate $\frac{1}{\pi}$.

The real use of Monte Carlo methods as a research tool stems from work on the atomic bomb during the Second World War. A German scientist, Lise Meitner, discovered in 1939 that if neutrons bombarded uranium, the uranium atom appeared to be actually splitting—breaking in half. The behaviour was named “fission”.

When a massive uranium-235 nucleus breaks into halves, it no longer requires all its neutrons. Not only does the break-up produce 7000 times the energy of the neutron that caused it, but it also liberates two or three neutrons to go on and do exactly the same thing to other nuclei.

Fission takes place spontaneously in the rare uranium-235, but most of the neutrons liberated in a small piece of it find their way through its sides without striking nuclei.

Monte Carlo methods were employed to create a direct simulation of the probabilistic problems concerned with this random neutron diffusion in fissile material.

Mathematics



CS Programming the General Mathematics Course

Creating a program in mathematics or, indeed, any subject, is a process of forward planning. Consequently, programming commonly starts with using a program planner to produce a sequence of topics. The support document for the General Mathematics course provides a range of examples of program overviews.

“Costing” a program

Traditional approaches to programming will begin by attempting to “cost” the various units in terms of expected teaching time. This produces a simple initial estimate of the required time with a new course.

Preliminary	Weeks	HSC	Weeks
Financial Mathematics		Financial Mathematics	
FM1: Earning money	1	FM4: Credit and borrowing	2
FM2: Investing money	2	FM5: Annuities and loan repayments	3
FM3: Taxation	2	FM6: Depreciation	2
Data Analysis		Data analysis	
DA1: Statistics and society	1	DA5: Interpreting sets of data	3
DA2: Data collection and sampling	2	DA6: The normal distribution	2
DA3: Displaying single data sets	2	DA7: Correlation	2
DA4: Summary statistics	2		
Measurement		Measurement	
M1: Units of measurement	1	M5: Further applications of area and volume	2
M2: Applications of area and volume	2	M6: Applications of trigonometry	2
M3: Similarity of two-dimensional figures	2	M7: Spherical geometry	2
M4: Right-angled triangles	3		
Probability		Probability	
PB1: The language of chance	2	PB3: Multi-stage events	2
PB2: Relative frequency and probability	2	PB4: Applications of probability	2
Algebraic Modelling		Algebraic modelling	
AM1: Basic algebraic skills	2	AM3: Algebraic skills and techniques	3
AM2: Modelling linear relationships	2	AM4: Modelling linear and non-linear relationships	3
Total	28	Total	30

This initial costing allows assessment and revision to be built into the program overview.

Remembering mathematics

Do your students have trouble remembering mathematics? Take heart, it can happen even to distinguished mathematicians. The mathematician J.J. Sylvester (1814–1897) objected on one occasion to a mathematical statement made by a companion. Sylvester insisted that the statement had never been heard of and, moreover, simply could not be true. The companion responded by showing the amazed Sylvester a paper written by Sylvester himself, in which Sylvester had announced his discovery of the statement concerned and had written out its proof!



Mathematics

CS

Stage 6 Special Program of Study

For students with special education needs

From 2001, HSC students with special education needs following a Special Program of Study (SPS) will be eligible for the award of the Higher School Certificate.

Special Program of Study courses

Students who meet the SPS eligibility requirements will be able to undertake Board-developed Life Skills courses, regular Board-developed courses and/or Board-endorsed courses.

Board-developed Life Skills will be 2 unit, 240-hour courses. The following courses have been endorsed and are currently being developed:

- English Life Skills
- Mathematics Life Skills
- Personal Development, Health and Physical Education Life Skills
- Citizenship and Society Life Skills
- Science Life Skills
- Creative Arts Life Skills
- Technological and Applied Studies Life Skills
- Workplace and Community-based Learning Life Skills.

Industry Curriculum Framework courses include

- Tourism and Hospitality
- Business Services (Administration)
- Retail Operations
- Primary Industries
- Information Technology
- Metal and Engineering
- Construction.

Students entered for an SPS may undertake the Industry Curriculum Framework courses either:

- under regular course arrangements, or
- by units of competency selected through the individual transition planning process from a 240-hour course (for example, 7 units of competency rather than 12 units over 240 hours, including 70 hours of work placement).

Eligibility requirements

Students who meet the SPS eligibility requirements are students with disabilities in special schools, support classes or regular classes.

The eligibility requirements for the SPS are that:

- students generally will have completed at least 4 Life Skills courses for the School Certificate
- students' planning must be undertaken through an individual transition planning process
- under special circumstances students will be allowed access to Stage 6 Special Program of Study courses, e.g. if the student has:
 - a deteriorating condition;
 - undertaken regular syllabuses in Stage 6 but has experienced **significant** difficulty.

Decisions about whether to enrol students in Special Program of Study courses for Stage 6 will be made by the school. The principal will be required to certify on the Preliminary and HSC entry forms that individual transition planning for each student entering for Life Skills courses in Stage 6 has occurred.

Note: The majority of eligible students will have an intellectual disability.

Pattern of study

Students undertaking an SPS follow the same pattern of study requirements for the HSC as other students. These are a minimum of:

- at least 6 units of Board-developed courses
- at least 2 units of Board-developed English
- at least 3 courses of 2 unit value
- at least 4 subjects.

Please refer to the HSC Calendar of Events for the Special Program of Study Events in November and December. (<http://www.newhsc.schools.nsw.edu.au>)

Curriculum Support in 2000

Subscriptions

CURRICULUM SUPPORT is available free of charge to teachers in NSW government schools.

It is available on subscription to teachers in non-government schools, to libraries and to others.

See your principal for a copy of the flier with details of how to subscribe, subscription rates and an application form.

As subscriptions determine the number of copies printed, we would be grateful to receive your order and cheque no later than Friday 25 February, 2000.

Evaluation fax sheet

Fax back to: 9886 7571

Your views on this year's CURRICULUM SUPPORT (Mathematics)

We would appreciate your views on this year's four editions of **CURRICULUM SUPPORT** and, in particular, the HSC supplement.

Please take some time to complete this page and fax it back to us so we can plan for next year's **CURRICULUM SUPPORT**.

LOOKING BACK OVER 1999	Strongly agree	Agree	Disagree	Strongly disagree
CURRICULUM SUPPORT keeps me well informed about current developments in my area of teaching.				
CURRICULUM SUPPORT provides me with many useful and practical ideas for teaching in my area.				
The HSC supplement has been a useful source of information on resources and ideas to assist me to plan for new HSC courses next year.				
It is important that all teachers have a personal copy of CURRICULUM SUPPORT for their area of teaching.				

LOOKING FORWARD TO 2000	Strongly agree	Agree	Disagree	Strongly disagree
I would like to see CURRICULUM SUPPORT changed in terms of				
• layout				
• size				
• design				
• content				

I would like next year's **CURRICULUM SUPPORT** to address the following issues in my KLA/area of teaching (please specify):

I would like next year's HSC supplement to provide me with information and ideas on the following areas (please specify):

Other comments or suggestions: