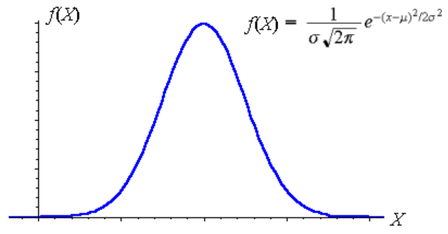
Assessing a population distribution from a sample – notes for teachers



Probability in the New Zealand Curriculum at levels 7 and 8 includes comparing theoretical distributions with experimental distributions. Students at level 7 are expected to be familiar with the normal distribution, uniform distributions and triangular distributions. At level 8 they also learn the binomial and Poisson distributions. Whenever a situation is modelled by a continuous distribution the area under the curve always adds to one. We only look at a sample to find out about the population it came from. We are never interested in the sample itself. We are asking whether the distribution of this sample provides evidence that it does not come from a population which could be modelled by a given distribution.

In the real world, distributions are *never* exactly normally distributed. The normal distribution is a theoretical model which is useful because it enables us to calculate probabilities for a distribution based only on the mean and standard deviation. If the population we are considering has a distribution which is approximately normal and we have good estimates for μ and σ, then the probability estimates we make for the population using a normal distribution model are going to be close to the actual population probabilities. We call such populations *normally distributed* to indicate that they have the characteristics of a normal distribution and can usefully be modelled by a normal distribution. Similar considerations apply to modelling using uniform, triangular, binomial or Poisson distributions.

The normal distribution can give reasonably accurate estimates for probabilities of distributions of populations if they have the following characteristics: unimodal; reasonably symmetric; frequency of the observations falls off rapidly as measurements get further from the central value; few or no extreme values. A uniform distribution is the best choice of model when there are lower and upper limits to possible values and little information about the shape of the distribution, or when the context and shape of the sample distribution suggest a uniform distribution. A triangular distribution is the best choice when there is information about lower and upper limits and the mode, but little information about the shape apart from that, or when the context and shape of the sample distribution suggest a triangular distribution.

Whether a sample distribution is consistent with being from a population which could be modelled by a given distribution is a matter of judgement. Contextual knowledge should be used to decide whether a model is useful, along with the characteristics of the distribution. For example, a small sample (eg $n<30$) may not look normal but could be consistent with being from an underlying normal distribution, while a large sample (eg $n>200$) from a normally distributed population would be expected to look approximately normal. Uniform and triangular distributions can be discrete or continuous. Unlike real world distributions, the underlying true distribution of a probability situation may in some cases be exactly modelled by a theoretical distribution. Students should have the opportunity to see how changing the bin (class) width changes the appearance of a histogram.

At level 7 NZC we are expecting students to make calls about whether or not a sample is consistent with what we would expect of a sample of that size from a population with a particular theoretical distribution. Sample distributions of various shapes should be discussed in teaching and learning. For assessment, sample distributions should either be reasonably obvious or the schedule should accept a justified decision either way. Use samples of over 200 where possible. We should not be expecting students to make borderline calls about the relative importance of a few values at the fringes vs the overall shape for a given sample size. At level 8 NZC, we are expecting students to make calls about the appropriate distribution to model a situation, based on a number of factors, including the nature of the situation and the characteristics of the sample distribution.