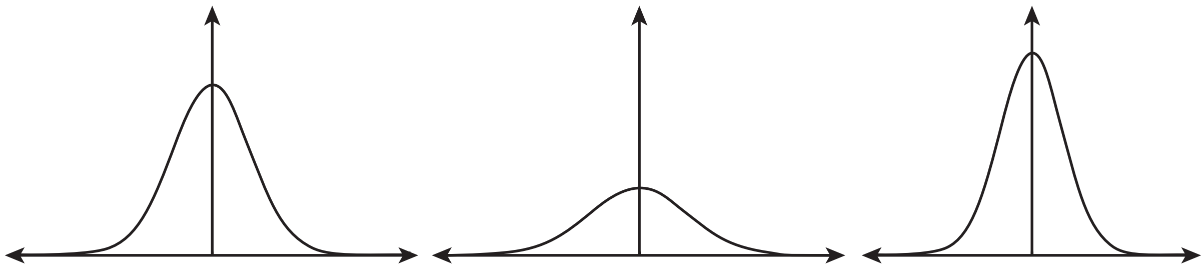
TOPIC 1: WORKING WITH STATISTICS

SUBTOPIC 1.1: NORMAL DISTRIBUTIONS

Normal Distributions

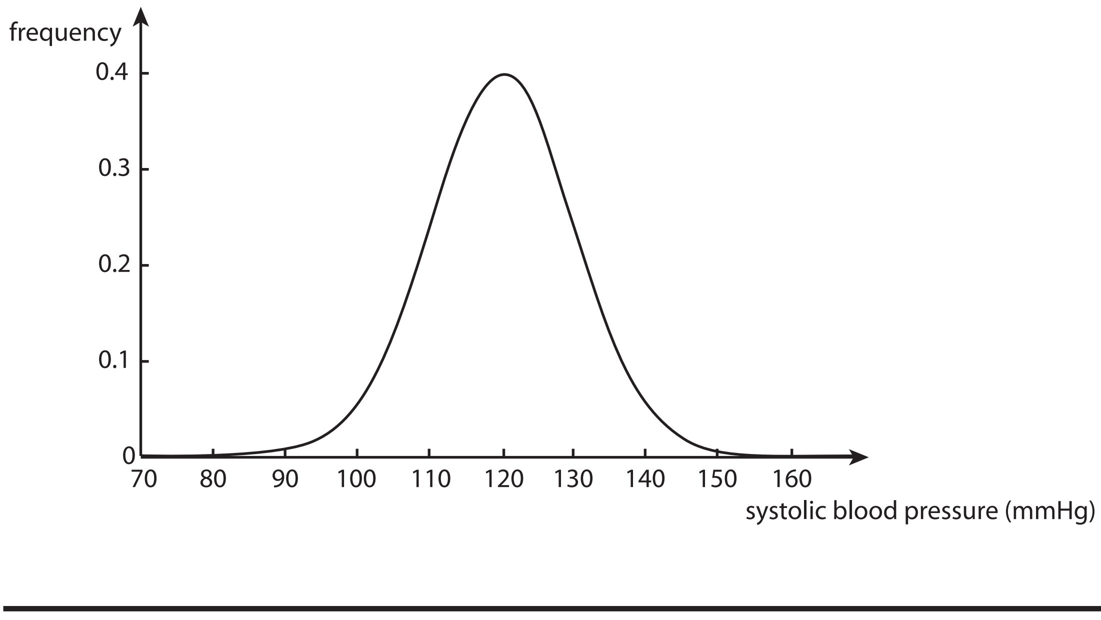
A normal distribution is a commonly occurring probability distribution with several notable characteristics. Firstly, it is symmetrical about its mean. That is, the probability of an event occurring depends only on how far away it is from the mean. The other main feature of the normal distribution is that the probability of an event occurring decreases the further away it is from the mean. The shape of the probability density function resembles a bell, and because of this it is commonly known as a "bell curve". As mentioned previously, because the normal distribution is a probability distribution, the area under the bell curve is always equal to 1.



Why Normal Distributions Occur

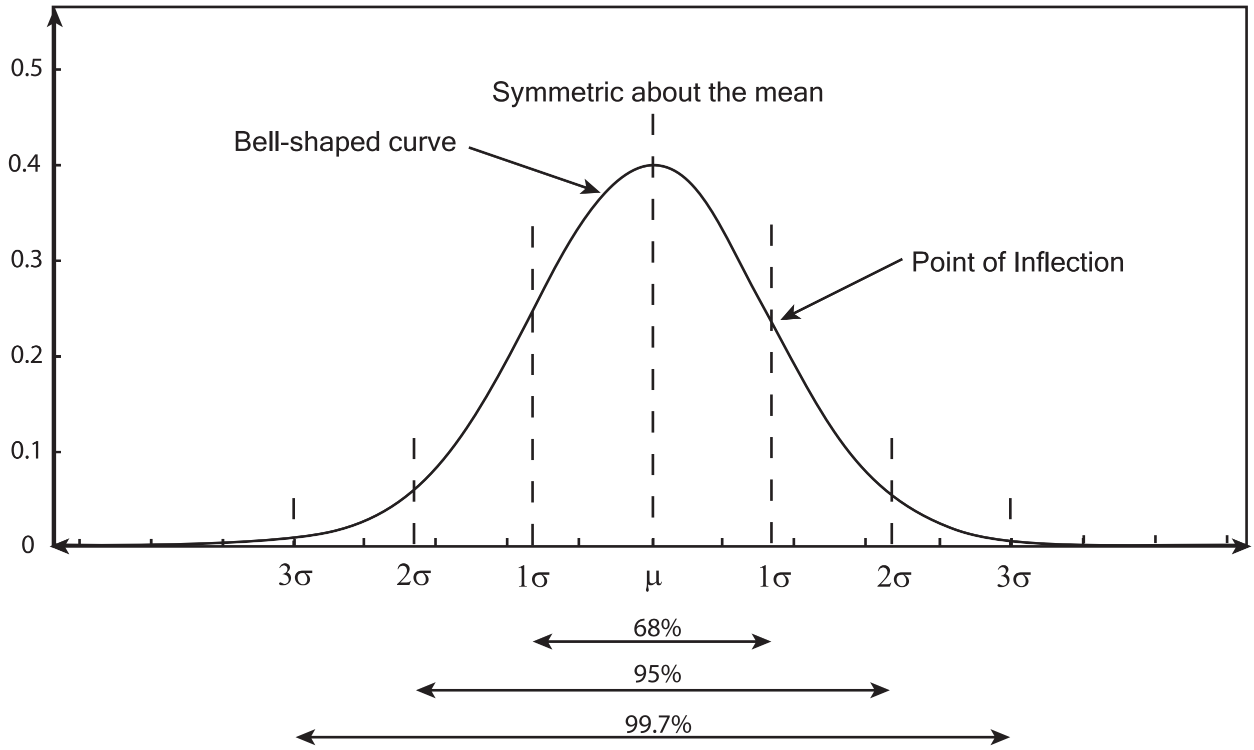
Normal distributions occur commonly in nature because the properties noted above are very common. Events which are closer to the mean are more likely to occur than more extreme events. This result is identical on both sides of the mean. This produces the typical bell curve; symmetrical about the mean and decreasing on either side, as we go further from the mean.

An example of a normal distribution is the distribution of systolic blood pressure in the population; variations in systolic blood pressure are usually normal as shown below, with a mean of 120 mmHg.



Features of a Normal Distribution

- Bell-shaped.
- Area under curve equals 1. This is because it is a probability distribution.
- Symmetry about the mean.
- Characteristic spread. Every normal distribution has a characteristic spread; approximately 68% of the data values are contained within 1 standard deviation on either side of the mean, 95% are within 2 standard deviations on either side of the mean and 99.7% are within 3 standard deviations on either side of the mean.
- First standard deviation at the point of inflection. The points of inflection of the bell curve coincide with the first standard deviation on either side of the mean (for an explanation of point of inflection, see *Subtopic 2.1: Using Functional Models*).



We can completely describe a normal distribution by its mean and standard deviation. Therefore, if a normal distribution has mean μ and standard deviation σ, we denote it by $N(\mu,\sigma^2)$ (note that we use σ^2 rather than σ).

The Standard Normal Distribution

Rather than using technology to store information about every possible normal distribution, we can translate any normal distribution into the "standard normal distribution". The standard normal distribution is a normal distribution with a mean of 0 and a standard deviation of 1. Thus we can denote the standard normal distribution by $N(0,1)$.

We can translate any normal distribution into the standard normal distribution because they have the same fundamental features, as outlined above. The only difference between the standard normal distribution and every other normal distribution is the mean and standard deviation values. However, the standard deviation affects the shape of the bell curve, so standardising a normal distribution requires translation to account for the different mean and scaling to account for the different standard deviation.

We standardise a normal distribution by converting the data values of random variable X into a universal random variable Z. Z is related to X in the following manner:
$Z = \dfrac{X-\mu}{\sigma}$, where μ is the population mean of X and σ is the standard deviation of X.

This universal random variable Z now has a mean of 0 and a standard deviation of 1. Therefore it has a standard normal distribution.

This allows us to use technology to find information about the standard normal distribution, which we can then translate back to the original normal distribution X.

Example 1:
The weight w of baby dugongs at birth are normally distributed with mean 14.9 kg and standard deviation 0.9 kg. Find by first converting to z-values:

a) $\Pr(w \geq 15.5)$

b) $\Pr(w \leq 13.2)$

c) $\Pr(14.5 \leq w \leq 15.6)$

Solutions:

a) $\Pr(w \geq 15.5) = \Pr(\dfrac{w-14.9}{0.9} \geq \dfrac{15.5-14.9}{0.9})$

$= \Pr(z \geq \dfrac{2}{3})$

$= 0.252$ (using technology)

b) $\Pr(w \leq 13.2) = \Pr(\dfrac{w-14.9}{0.9} \leq \dfrac{13.2-14.9}{0.9})$

$= \Pr(z \leq -1.89)$

$= 0.0294$ (using technology)

c) $\Pr(14.5 \leq w \leq 15.6) = \Pr(\dfrac{14.5-14.9}{0.9} \leq \dfrac{w-14.9}{0.9} \leq \dfrac{15.6-14.9}{0.9})$

$= \Pr(-\dfrac{4}{9} \leq z \leq \dfrac{7}{9})$

$= 0.453$ (using technology)

Example 2:

Find k such that $\Pr(x \le k) = 0.45$, where x is normally distributed with mean 200.7 and standard deviation 12.0.

Solution:

$\Pr(x \le k) = 0.45$

$\therefore \Pr(z \le \frac{k - 200.7}{12}) = 0.45$

using technology, $\frac{k - 200.7}{12} = -0.1257$

$\therefore k = 199.2$

Example 3:

José scored 175 (out of a possible 200 marks) in the Stage 2 Chemistry exam, and 135 out of 200 for Specialist Maths. If the state's scores for Chemistry are normally distributed with mean 130 and standard deviation 21, and for Specialist Maths 95 and 15 respectively, in which subject did José score better in comparison to the rest of the state?

Solution:

Start by standardising both distributions. We need to do this in order to obtain the proportion of students in the state scoring less than José in each subject.

Let s be a student's score in Chemistry. Thus the probability of a student scoring lower than José is given by:

$$\begin{aligned}\Pr(s \le 175) &= \Pr(z \le \frac{175 - 130}{21}) \\ &= 0.984 \text{ (using technology)}\end{aligned}$$

Considering José's performance in Specialist Maths in the same way, we find that:

$$\begin{aligned}\Pr(s \le 135) &= \Pr(z \le \frac{135 - 95}{15}) \\ &= 0.996 \text{ (using technology)}\end{aligned}$$

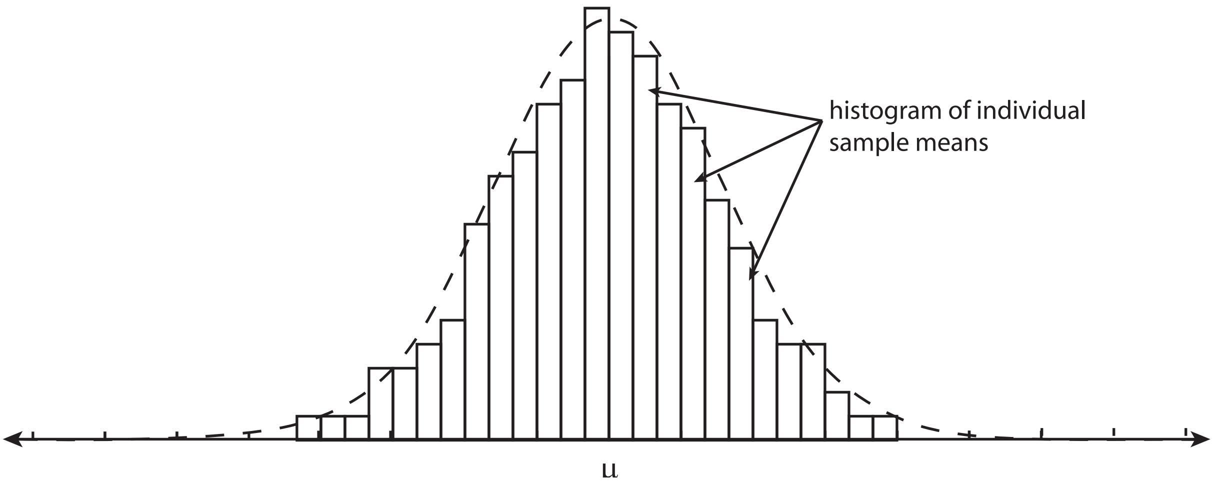
Thus José has scored better than 98.4% of the state in Chemistry, and 99.6% of the state in Specialist Maths. So, in comparison to the rest of the state, José scored better in Specialist Maths.

TOPIC 1: WORKING WITH STATISTICS

SUBTOPIC 1.2: CENTRAL LIMIT THEOREM

The Result of Plotting the Mean of Each Sample when Sampling Repeatedly from Any Population

Consider any population. Say we select a number of random samples, each with n members from the original population, and plot a histogram of the means of each sample. At first, we obtain an apparently random distribution. However, we can continue to build up a distribution of sample means by taking more samples of size n. As we keep taking more and more samples, eventually a distribution similar to a normal distribution arises:



It should be noted that the mean of the sample means will be equal to the mean of the original population, μ.

Hence, even though the original population may not have been normally distributed, the distribution of the means of random samples is approximately normal, with the approximation to a normal distribution improving as the sample size increases. This result forms the basis for the Central Limit Theorem.

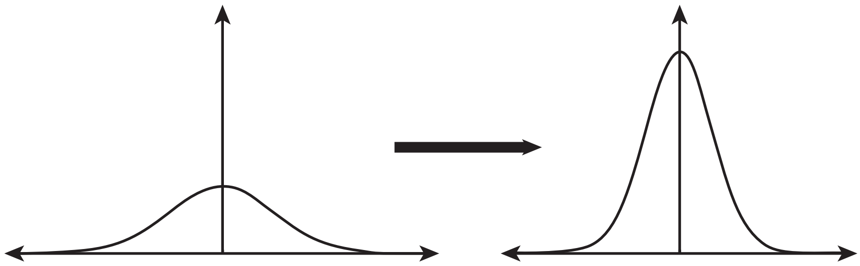
The Central Limit Theorem

Sample Mean and Population Mean

The means of random samples taken from any population will be approximately normally distributed with a mean equal to the population mean. The approximation will improve as the sample size increases.

The Central Limit Theorem states that even if the population does not have a normal distribution, if we pick a number of *random* samples and calculate the mean of each sample, then these means will be approximately normally distributed.

The sample means are distributed symmetrically about μ, the mean of the original population. However, the standard deviation of this normal distribution is dependent on the size of the samples. As the number of elements in each sample increases, the standard deviation of the sample mean distribution decreases. This results in the following change of shape of the resultant distribution, as n is increased:



Notice that the area under the curve remains equal to 1, because this is a probability distribution. So, as the peak becomes thinner, it will also become taller.

Mean and Standard Deviation of Distribution of Sample Means

Mean of sample means $= \mu$

Standard deviation of sample means $= \dfrac{\sigma}{\sqrt{n}}$

In this case, n refers to the size of each sample, *not* the number of samples. The number of samples merely determines whether or not we have enough information to use the Central Limit Theorem to predict probabilities.

As the sample size increases, the approximation to a normal distribution improves. Note that it is often stated that a sample size of $n \geq 30$ constitutes a large sample, such that we can accurately approximate the sample mean distribution to a normal distribution. However, this isn't a concrete rule, but merely a rule of thumb obtained experimentally. There are some cases in which a smaller sample size may still result in an accurate approximation to a normal distribution. For example, if the original population is already normally distributed, then sample sizes of less than 30 are often sufficient to obtain a very good approximation to a normal distribution of sample means.

Simple Applications of the Central Limit Theorem

Example 1:

The average number of olives on a supreme pizza at Pizza Zone is 63 with a standard deviation of 4.

a) Calculate the probability that an individual pizza will have less than 60 olives on it.
b) Calculate the probability that the mean $\bar{x}$ of a sample of 49 pizzas is less than 60 olives.

Solutions:

a) Let x = number of olives on the pizza.

$\Pr(x<60)=0.227$ (using technology)

b) In this case, the mean of the sample means is 63 olives and the standard deviation is

Std dev $=\dfrac{\sigma}{\sqrt{n}}=\dfrac{4}{\sqrt{49}}=\dfrac{4}{7}$.

$\Pr(\bar{x}<60)=7.60\times10^{-8}$ (using technology)

Example 2:

The average head circumference of an individual alien on the (as yet) undiscovered planet Lozoph is 207 centimetres with standard deviation 14 centimetres.

a) Calculate the probability that an individual alien will have a head circumference less than 200 centimetres.

b) Calculate the probability that the mean $\bar{x}$ of a sample of head circumferences of 25 aliens is less than 200 centimetres.

Solutions:

a) Let x = head circumference of alien.

$\Pr(x<200)=0.309$ (using technology)

b) In this case, the mean of the sample means is 207 centimetres and the standard deviation is

Std dev $=\dfrac{\sigma}{\sqrt{n}}=\dfrac{14}{\sqrt{25}}=\dfrac{14}{5}$.

$\Pr(\bar{x}<200)=0.00621$ (using technology)

The Importance of the Central Limit Theorem for Statistics

"In statistical inference of means or proportions, it can be assumed that sample means and sample proportions have approximately normal distribution, regardless of the original distribution from which observed values were sampled" (SACE Stage 2 Mathematical Studies Curriculum Statement, Subtopic 1.2). In general, normal distributions are easier to work with than other probability distributions. This is because information about normal distributions is easy to find by using the standard normal distribution (see *Subtopic 1.1: Normal Distributions*). For other distributions, such as the one below, it may or may not be possible to find such information.