# The Central Limit Theorem

## Prerequisites

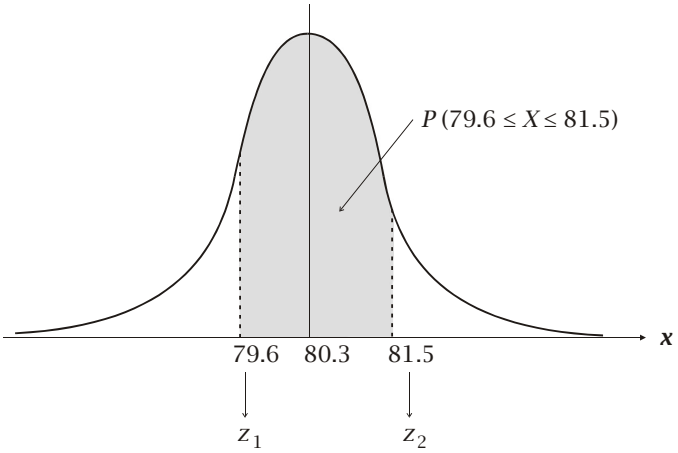
You should be familiar with the standardised normal variable.

**Example (6)**

Chocolate Company produces a *Tasty Bar* chocolate. The weight $X$ of *Tasty Bars* is believed to be normally distributed with mean 80.3 g and standard deviation 1.3 g. What is the probability that a *Tasty Bar* selected at random has a weight between 79.6 g and 81.5 g? Give your answer to 3 significant figures.

Solution

$X \sim N\left(80.3,(1.6)^2\right)$




We seek the $z$-values corresponding to the $x$-values $x_1 = 79.6$ and $x_2 = 81.5$. Substituting into $z = \dfrac{x-\mu}{\sigma}$

$z_1 = \dfrac{79.6-80.3}{1.3} = -0.538 \qquad \Phi(0.538) = 0.7046$

$z_2 = \dfrac{81.5-80.3}{1.3} = 0.923 \qquad \Phi(0.923) = 0.8220$

$$\begin{aligned} P(79.6 < X < 81.5) &= P(-0.538 < Z < 0.923) \\ &= 0.7046 + 0.8220 - 1 \\ &= 0.52.66 \\ &= 52.7\% \text{ (3 s.f.)} \end{aligned}$$

You should also know how to find a mean and variance of a set of sample data.

**Example (1) continued**

The manufacturers of *Tasty Bar* take a sample of 24 bars and record the following weights.

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 80.1 | 80.4 | 79.6 | 81.2 | 78.6 | 80.3 | 78.7 | 77.3 | 80.8 | 78.1 | 79.2 | 80.6 |
| 80.4 | 78.2 | 78.8 | 80.9 | 78.0 | 81.5 | 79.6 | 78.7 | 80.1 | 79.3 | 80.3 | 79.4 |

Find the sample mean and sample variance.

Solution

The sample mean is given by

$$\bar{x} = \frac{\text{Sum of values}}{n} \qquad \left[\bar{x} = \frac{\sum x}{n}\right]$$

Here $n = 24$ and the sum of the values is $\sum x = 1828.9$.

$$\bar{x} = \frac{1910.1}{24} = 79.5875 = 79.6 \text{ g (3 s.f.)}$$

The sample variance is given by

$$\sigma^2 = \frac{\text{Sum of squares}}{n} - (\bar{x})^2 \qquad \left[\sigma^2 = \frac{\sum x^2}{n} - (\bar{x})^2\right]$$

$$\sum x^2 = 152048.75$$

$$\sigma^2 = \frac{\sum x^2}{n} - (\bar{x})^2 = \frac{152048.75}{24} - (79.5875)^2 = 1.19 \text{ (3 s.f.)}$$

The manufacturers of *Tasty Bar* believe that they are producing bars with a mean weight of 80.3 g. However, the mean of the sample of 24 bars is just 79.6 g. Is it possible that something has gone wrong with their machines and that they are now consistently producing bars with a lower weight? Clearly, this is the kind of question that deeply concerns real-life manufacturers. However, before we can consider an answer to the question we must investigate further the properties of the sample mean.

# The sample mean

Let us suppose that Chocolate Company who produces the *Tasty Bar* regularly takes samples of size 24 of these bars in order to monitor the output and ensure that *Tasty Bars* remain, on average 80.3 g. Each of these samples produces a statistic that is the mean of the sample. Suppose that in five such samples, each comprising 24 observations, these means are

$\bar{x}_1 = 79.6$ $\bar{x}_2 = 79.8$ $\bar{x}_3 = 78.8$ $\bar{x}_4 = 80.5$ $\bar{x}_5 = 80.1$

These are 5 values that belong to a variable. We denote by $X$ the weight of one *Tasty Bar*. But the values $\bar{x}_1, \bar{x}_2, \bar{x}_3, \bar{x}_4, \bar{x}_5 \ldots$ above are **not** values of this variable $X$. This is indicated by the bar above each of $\bar{x}_1, \bar{x}_2, \bar{x}_3, \bar{x}_4, \bar{x}_5 \ldots$ which denotes that each one is **an average** of 24 values of weights and not a weight of **one** bar. The values $\bar{x}_1, \bar{x}_2, \bar{x}_3, \bar{x}_4, \bar{x}_5 \ldots$ are values of a new variable, which represents the mean of samples of 24 *Tasty Bars*. Thus we distinguish between two variables

$X$ denotes in this context the weight of one bar. Each value that $X$ takes is a single observation. The expression $X = x$ denotes that $X$ takes the value $x$.

$\bar{X}$ denotes in this context the mean of the weights of 24 bars in a sample. It is the average of 24 observations all part of one sample. The expression $\bar{X} = \bar{x}$ denotes that the sample mean takes a particular value $\bar{x}$. Essential to this notation is the difference between a capital letter $\bar{X}$ denoting a variable and the small letter $\bar{x}$ denoting a value of that variable.

This situation clearly generalises, so that if $X$ represents a variable that has a certain probability distribution drawn from a population and if samples of size $n$ are taken from this population, then $\bar{X}$ represents the mean of those samples.

Let us now suppose that the variable $X$ is normally distributed, so that it has a probability distribution $X \sim N(\mu, \sigma^2)$. Does it follow that the sample mean, $\bar{X}$, is also normally distributed in this way? In fact, it is a different variable and it turns out that we can show that whilst it is normally distributed it has different parameters. This important result is known as the central limit theorem.

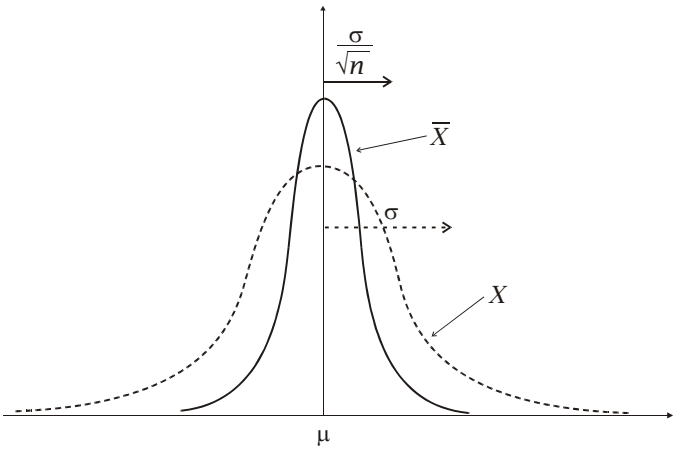
Let us first consider what type of distribution $\bar{X}$ is likely to have. Firstly, the expected value of $\bar{X}$ is surely the population mean $\mu$. That is $E(\bar{X}) = \mu$ where $\mu$ is the population mean. This is

expected because the sample is taken from the population. We assert that the result $E(\bar{X}) = \mu$ is correct and can be proven from the definition of expectation.

Now $\bar{X}$ is a variable determined by chance. It is an observation of a sample mean and is consequently likely to deviate from the population mean – so it too will have a variance. However, the sample mean will not deviate from the expected value to the same extent as a single observation, precisely because it is already an average and averages diminish the effect of random observations being widely separated from each other.

For example, suppose $X$ is a normally distributed variable describing a population and let $X \sim N(0,1)$. Suppose we start to sample this population. Suppose the first observation of $X$ is $X_1 = -1$ and the second is $X_2 = 1$. The value $X_1 = -1$ is one standard deviation away from the population mean $\mu = 0$. So too is the value $X_2 = 1$ though in the opposite direction. This is quite to be expected. Some of the observations will be less than the population mean and some will be more than it. On average we expect half the observations in any one sample to be less than the mean, and half to be more than it. After the first observation the sample mean is the same as the value and $\bar{X} = X_1 = -1$. However, after the second observation $\bar{X} = \frac{X_1 + X_2}{2} = \frac{-1+1}{2} = 0$. This illustrates how averages "iron out" variance. So, if the **population** has one variance, we expect the variance of the **sample mean** to be smaller. In fact, it can be shown that the variance of the sample mean is

$$\text{var}(\bar{X}) = \frac{\sigma^2}{n}$$




*The sample mean $\bar{X}$ has less variance than that of a single observation X.*

**The central limit theorem (for normally distributed parent populations)**

If $X \sim N\left(\mu,\ \sigma^2\right)$ then $\bar{X} \sim N\left(\mu,\ \frac{\sigma^2}{n}\right)$ for samples of size *n*.

This states that the sample mean of a normally distributed population is also normally distributed with the same mean. But, if the standard deviation of the population is $\sigma$, then the standard deviation of the sample mean is $\frac{\sigma}{\sqrt{n}}$, and thus the sample mean is more tightly distributed. It is not usual to prove the central limit theorem at this level.

**Example (2)**

A brewery produces kegs of beer whose alcohol content has mean 4% and standard deviation 0.5%. Calculate the probability that the mean alcohol content of a sample of 20 kegs will exceed 3.8%.

Solution

Let $X$ denote the alcohol content of a keg of bear, and let $\bar{X}$ denote the mean of a sample of 24 kegs of beer.

Then $X \sim N\left(4,0.5^2\right)$.

By the central limit theorem

$\bar{X} \sim N\left(4,\frac{0.5^2}{20}\right) = N\left(4,0.0125\right)$.

We require $P\left(\bar{X} > 3.8\right)$. Denote by $\sigma_{\bar{X}}$ the standard deviation of the mean $\bar{X}$. Then

$\sigma_{\bar{X}} = \frac{0.5}{\sqrt{20}} = \sqrt{0.0125} = 0.11180...$

The $z$-value corresponding to $\bar{X} = 3.8$ is

$z = \ z = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} = \frac{3.8 - 4}{\sqrt{0.0125}} = -1.789...$ $\qquad \Phi\left(1.789\right) = 0.9633$

$P\left(\bar{X} > 3.8\right) = P\left(Z > -1.789\right) = 0.9633 = 0.963\ (3\text{ s.f.})$

**Example (3)**

The random variable $X$ is normally distributed with mean 4.2 and variance 2.8. If $\bar{X}$ denotes the mean of a random sample of 12 observations of $X$ state the mean and variance of $\bar{X}$.

Solution

$X \sim N(4.2, 2.8)$

$E(\bar{X}) = 4.2$

By the Central Limit Theorem

$\text{var}(\bar{X}) = \frac{2.8}{12} = 0.233 \ \ (3 \text{ s.f.})$

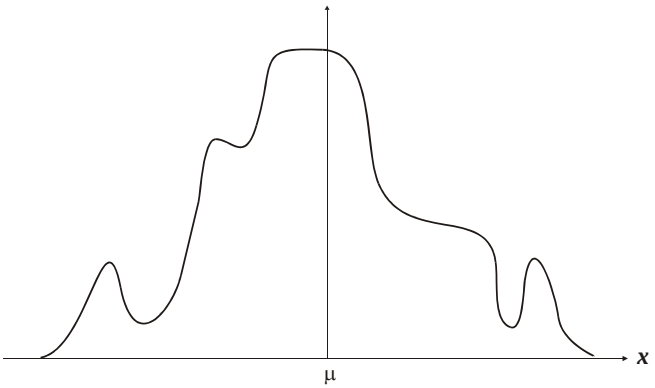
# Extending the central limit theorem

We can show that the central limit theorem applies to samples drawn from **any** population. That is, regardless of what the background distribution of the population is, the sample mean follows a normal distribution.

**The central limit theorem**

If $X_1, X_2, X_3, ...., X_n$ is a random sample of size $n$ from **any** distribution with mean $\mu$ and variance $\sigma^2$ then, for large $n$, the distribution of the sample mean $\bar{X}$ is *approximately normal* and $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ where $\bar{X} = \frac{1}{n}(X_1, X_2, X_3, ...., X_n)$. As $n \to \infty$ the approximation becomes better and better.

The significance of this theorem is that it provides an effective method of working with any population whatsoever, because by taking samples from that population of sufficiently large size, we can replace that population by a normally distributed sample. We can also estimate the mean and variance of this sample. For example, suppose the background population has a "weird" probability distribution, represented by this graph




Then, provided the samples of size $n$ are large enough, the distribution of the sample mean $\bar{X}$ is approximately normal and $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$. Here the sample size $n$ that is needed to be able to conclude that the sample mean is normally distributed is not specified. This is because it depends on the nature of the background distribution. If the parent population is already normally distributed, then automatically so is the sample mean. The less the background population is like a normal distribution to begin with, the larger the sample size must be in order to apply the central limit theorem. So the size of $n$ required in this version of the theorem is not fixed.

There can be some confusion in a student's mind as to when to use the central limit theorem. The simple answer is that if the question states the number of items in a sample and asks you to calculate a probability for the sample mean, $\bar{X}$, then you are using the central limit theorem.

**Example (4)**

The voltage supplied by an a/c to d/c converter is normally distributed with mean 9.0 V and standard deviation 3.0 V. Random samples of $n$ converters are taken. The mean voltage supplied for the sample is denoted by $\bar{X}$.

(*a*) State the distribution of $\bar{X}$, and give its mean and variance.

(*b*) Find the probability that $\bar{X}$ is greater than 9.2V when $n = 30$.

(*c*) Find the smallest sample size if the probability that $\bar{X}$ is greater than 9.2V is less than 0.005.

(*d*) Suppose that the voltage supplied has mean 9.0 V and standard deviation 3.0 V but the probability density function of the voltages is not known. If random samples of size $n$ are taken, what do we know about the distribution of the mean voltage when *n is large and when n* is small?

Solution

(*a*) $X \sim N\left(9.0, 3.0^2\right)$

$\bar{X} \sim N\left(9.0, \frac{3.0^2}{n}\right)$

$E\left(\bar{X}\right) = 9.0$

$Var\left(\bar{X}\right) = \frac{3.0^2}{n}$

(*b*) $n = 30$

$\bar{X} \sim N\left(9.0,\ \frac{3.0^2}{30}\right) = N\left(9.0, 0.3\right) = N\left(9.0, 0.5477^2\right)$

We require $P\left(\bar{X} > 9.2\right)$

$z = \frac{x-\mu}{\sigma} = \frac{9.2-9.0}{0.5477} = 0.365 \text{ (3 s.f.)}$ $\qquad \Phi\left(0.365\right) = 0.6425$

$P(\ \ ) = 0.6425$

$P\left(\bar{X} > 9.2\right) = 1 - P\left(Z < 0.365\right) = 1 - 0.6425 = 0.3575 = 0.368 \text{ (3 s.f.)}$

(*c*) The sample size is $n$ and $\bar{X} \sim N\left(9.0,\ \frac{3.0^2}{n}\right)$

Then $z = \frac{x-\mu}{\sigma} = \frac{9.2-9.0}{\sqrt{\frac{3.0^2}{n}}}$

Here we have $P\left(X > 9.2\right) = 0.005$. This corresponds to $z = 2.576$.

Then $2.576 = \frac{0.2}{\left(\frac{3.0}{\sqrt{n}}\right)}$

$\sqrt{n} = \frac{2.576 \times 3.0}{0.2} = 38.64$

$n = 1493.04...$

Then $n = 1494$ which is the next integer up.

(*d*) When $n$ is large then, by the central limit theorem, the distribution of the mean of the samples is approximately normal. But when $n$ is small, the distribution remains unknown.