Sampling distributions

1. Statistical inference based on sampling distributions

The process of statistical inference

The frequentist approach to **statistical inference** involves:

1. formulating a **null hypothesis** about a **characteristic**[1] of members of a **population**,
2. measuring that characteristic in a **sample** of the population,
3. selecting a **test statistic** to test that null hypothesis, and
4. assessing whether the calculated value of the test statistic is so large that it provides sufficient evidence to reject the null hypothesis.

1. The null hypothesis

The practice of science involves asserting and testing hypotheses that are capable of being proven false. The **null hypothesis** typically corresponds to a position of no association between measured characteristics or no effect of an intervention. The null hypothesis is paired with an **alternative hypothesis** that asserts a particular relationship exists between measured characteristics[2] or that the intervention has a particular effect.[3]

2. The sample

Sampling is used because usually it is infeasible or too costly to measure the characteristic(s) The of every member of the population. For statistical inference to be possible, the sample must be drawn at random or randomly allocated between the arms of a trial.

3. The test statistic

The test statistic for a null hypothesis is a summary value derived from the sample data. It could be the sample total or mean or proportion, or the difference between the sample totals or means or proportions of two or more groups. It could be a count or ratio of counts, a correlation coefficient, or a regression coefficient.

4. Assessment of the calculated value of the test statistic

In **parametric statistics**, the calculated value of the test statistic (the **signal**) is compared to all possible values of the test statistic (the **interference**).[4] It is assumed that **noise** (errors from the other sources of sample selection, measurement and confounding) has been eliminated during the earlier stages of study design, conduct and analysis.

This requires a quantitative measure of interference, which in turn requires an evaluation of all other possible values of the test statistic.

---

[1] Examples are weight, smoking status, colour, diagnosis with melanoma, response to an intervention.
[2] That blood pressure increases with body mass index, for example.
[3] Of reducing cholesterol, for example
[4] In some applications, engineers evaluate whether the power of the **signal** (the useful information from the source device) is large relative to the power of **interference** (the multitude of other signals from the same source). In other applications, they evaluate whether the power of the signal is large relative to the power of **noise** (the useless information from other sources that exists in the absence of the signal).

The sampling distribution of a test statistic

Example of a test statistic

Suppose in a study of systolic blood pressure (SBP) that the mean SBP of the one group of $n_1 = 100$ subjects is $\bar{x}_1 = 134$ mm and the mean SBP of a second group of $n_2 = 100$ subjects is $\bar{x}_2 = 126$. The purpose is to test whether the populations from which each sample group is drawn have identical SBP on average. For the null hypothesis, the difference of sample means $\bar{x}_1 - \bar{x}_2 = 134 - 126 = 8$ is an appropriate test statistic.

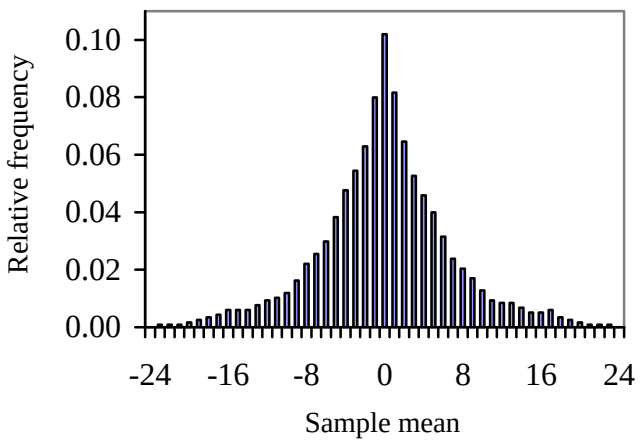
Is a difference of $\bar{x}_1 - \bar{x}_2 = 8$ mm **large enough to warrant rejection of the null hypothesis**?

Assessment of the interference

The sampling distribution shows the spread of the values of the test statistic **when the study is repeated under identical conditions** (using the same method of sampling with the same sample size, the same measurement procedures, and the same analytical techniques).

The only aspect of the study that can vary when it is repeated under identical conditions is the sample used. Sampling variation arises because each sample is different in membership (but not in size). Hence the test statistic will have a different value each time the study is repeated.

Suppose that the sampling distribution of the difference of sample means of SBP in samples of size $n = 100$ drawn from each population is as shown at right. The figure shows the relative frequency of each possible difference of sample means grouped in ranges of 1 mm Hg. Around 10% of the differences of sample means lie between –0.05 and 0.5, around 8% lie between –1.5 and –0.5 mm, around another 8% lie between 0.5 and 1.5, and so on. The distribution is centred at 0, the hypothesised value, but it need not be.



How is the information of the sampling distribution used?

The sampling distribution can be used for statistical inference and to calculate p-values and confidence intervals.

For example, by carefully inspecting the sampling distribution, it is possible to determine that:

1. 11.0% of sample means are ≥ 7.5mm, so $p = 0.110$ (one-tailed test);
2. Another 11.1% of sample means are ≤ –7.5mm, so $p = 0.221$ (two-tailed test).

Similarly we could calculate 95% confidence intervals for the difference of population means. Careful inspection reveals them to be approximately ±15 mm,

The standard error

An alternative would be to summarise the spread of the sampling distribution in terms of its standard deviation (SD). The SD of the sampling distribution is referred to as the **standard error** (SE) of sampling. It measures the error in the sample mean that arises because not all members of the population are included in the sample. If all members had been included in the sample, the sample mean would be identical to the population mean – there is no error.

Obtaining information about the sampling distribution

Information about the sampling distribution can be obtained in at least four ways:

1. by complete enumeration of all possible values of the test statistic (rarely feasible);
2. by repeated sampling from the population (rarely feasible);
3. by re-sampling from the sample data, and
4. by assuming that it is well-approximated by a known distribution (eg normal, binomial, Poisson).

2. Complete enumeration of the sampling distribution of the sample mean

The set of all possible values of the sample mean $\bar{x}$ in samples of size n is its **sampling distribution** in samples of size n. Usually it is too vast to be enumerated fully. For example, in selecting samples of size $n = 100$ from a population of $N = 10000$, there are $N = 6.5208 \times 10^{241}$ different samples that can be drawn.

There is a different sampling distribution for each possible sample size $n = 1, 2, \ldots N - 1$. Each sampling distribution has its own mean[5] and SD.

Example of a sampling distribution determined by complete enumeration

Consider estimating the mean height of persons in a population of $N = 5$ people in samples of size $n = 2$ members. The heights (in cms) of the members of the population are:

162, 168, 174, 192, 204

The set of all possible samples of size $n = 2$ are shown below:

Possible samples of size $n = 2$				
{162,168}	{162,192}	{168,174}	{168,204}	{174,204}
{162,174}	{162,204}	{168,192}	{174,192}	{192,204}

By calculating the mean of each possible sample, we find that the sampling distribution of the sample mean in samples of size $n = 2$ is:

Possible values of the sample mean in samples of size $n = 2$						
Very low	Low	Little low	About right	Little high	High	Very high
				183		
	168			183		
165	171	177	180	186	189	198

In this case, the sampling distribution has been fully enumerated, and the mean of the sampling distribution of sample means is an exact estimate of the population mean μ_χ. It is:

$$\hat{\mu}_\chi = \frac{165+168+171+177+180+183+183+186+189+198}{10} = 180 \text{ cm}$$

The SD of the sampling distribution of samples of size $n = 2$ is the standard error (SE) of sampling with samples of size $n = 2$:

$$\text{SE}(\bar{x}) = \sqrt{\frac{(165-180)^2 + (168-180)^2 + (171-180)^2 + \ldots + (198-180)^2}{10}} = 9.6 \text{ cm}$$

[5] The mean of the sampling distribution of sample means is the mean of sample means!

Another example of a sampling distribution determined by complete enumeration

There is a different sampling distribution for each sample size. Suppose the study involves selecting samples of size $n = 3$ from the example population. The possible samples are:

Possible samples of size $n = 3$				
{162,168,174}	{162,168,204}	{162,174,204}	{168,174,192}	{168,192,204}
{162,168,192}	{162,174,192}	{162,192,204}	{168,174,204}	{174,192,204}

The sampling distribution of the sample mean of height in samples of size $n = 3$ has the elements:

Possible values of the sample mean in samples of size $n = 3$						
Very low	Too low	Little low	About right	Little high	Too high	Very high
			178			
			178			
		174	180	186		
	168	176	182	188	190	

Note that this sampling distribution is more symmetric and "bell-shaped" than that for samples of size $n = 2$. Its mean is:

$$\hat{\mu}_{\chi} = \frac{168+174+176+178+178+180+182+186+188+190}{10} = 180 \text{ cm}$$

The SD of the sampling distribution of samples of size $n = 3$ is the standard error (SE) of sampling with samples of size $n = 3$:

$$\text{SE}\left(\overline{x}\right) = \sqrt{\frac{\left(168-180\right)^2 + \left(174-180\right)^2 + \left(176-180\right)^2 + \ldots + \left(190-180\right)^2}{10}} = 6.4 \text{ cm}$$

Note that the standard error is lower in samples of size $n = 3$ than in samples of size $n = 2$.

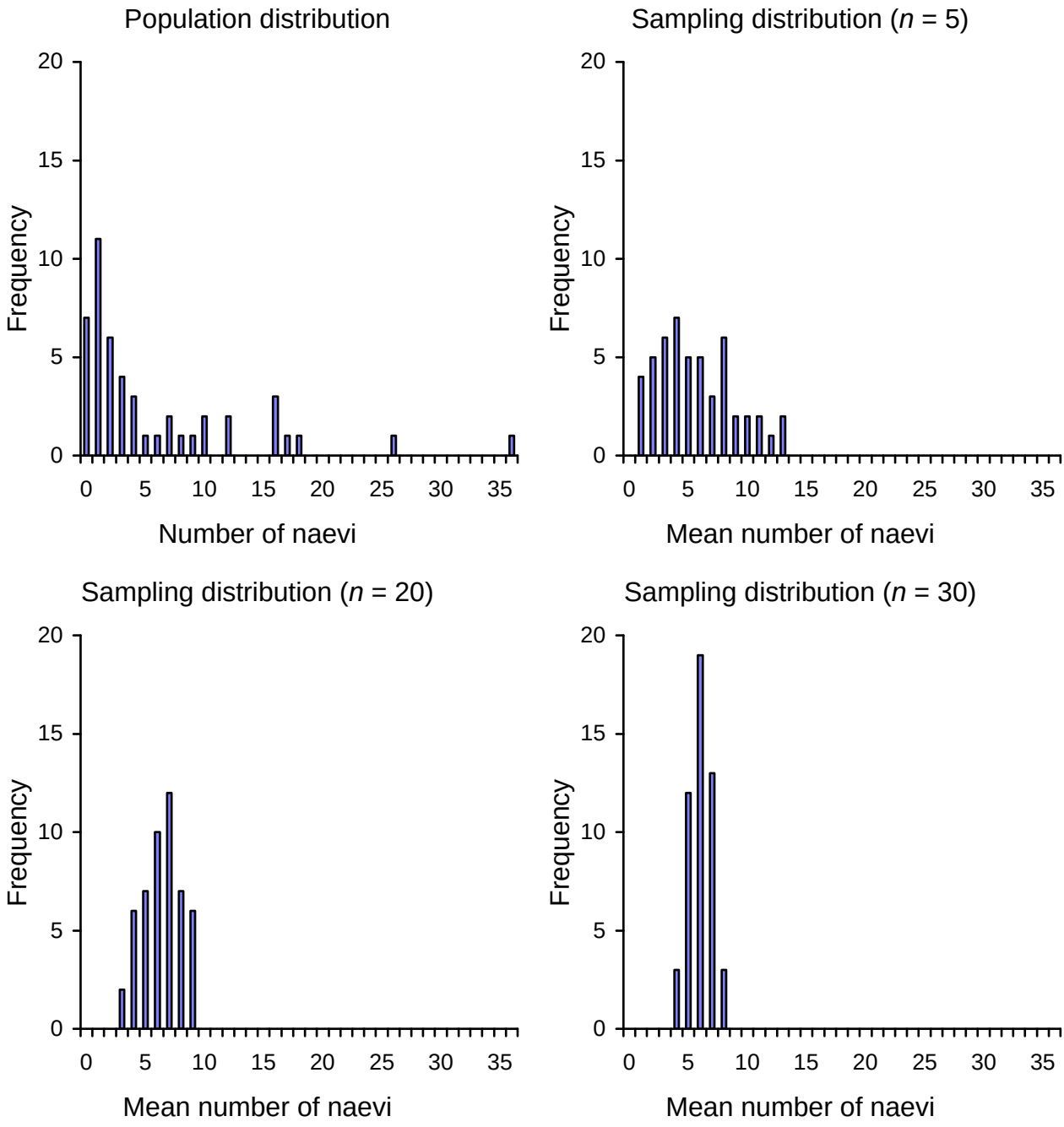
These results highlight three properties of the sampling distribution of the sample mean:

1. the sample mean is an ***unbiased*** estimator of the population (on average, the sample mean is correct – the mean of the sample means will closely approximate the population mean);
2. the standard error of sampling becomes smaller as the sample size increases (as more of the population is included in the sample).
3. the sampling distribution of the sample mean becomes more like a "***normal***" distribution as the sample size increases.

3. Approximation of the sampling distribution by repeated sampling

Consider the population of second-year medical students at the University of Tasmania in year 2000. Each member had the number of common melanocytic naevi (pigmented moles) of at least 2mm in diameter on their left arms counted by an experienced study nurse.

The figure at top left depicts the population distribution of the naevi counts. Each column corresponds to a naevi count, and the height of the each column represents the number of members of the population who had that naevi count in 2000. Repeatedly drawing samples of size 5, 20 and 30 persons and calculating the sample mean in each (with each sample mean rounded to the nearest integer) gave rise to the three sampling distributions depicted.



The population distribution was grossly asymmetric, but the sampling distributions are reasonably symmetric and become more symmetric and bell-shaped as the sample size increases.

4. Approximation of the sampling distribution by re-sampling

Re-sampling

In statistics, re-sampling from a sample of data is used for a variety of purposes:

1. estimating the standard error of a test statistic using subsets of available data (**jackknifing**) or drawing randomly with replacement from the data (**bootstrapping**) to obtain the entire sampling distribution;
2. exchanging labels on data points when performing permutation tests (also called exact tests);
3. validating models by using random subsets (cross-validation).

Jackknifing

This involved repeatedly computing the test statistic in a subset of the sample data formed by deleting one or more observations at a time. From this set of replicates of the statistic, an estimate of the bias (if any) and sampling variance of the test statistic – and thereby its standard error, which is the square root of the sampling variance – can be calculated.

For many test statistics (including the sample mean, correlation coefficient and regression coefficient), the jackknife estimate of variance is consistent (tends almost surely to the true value as the sample size become larger).

The jackknife, like the original bootstrap, is dependent on the independence of the data. Extensions of the jackknife to allow for dependence in the data have been proposed.

Bootstrapping

Bootstrapping is a statistical method for estimating the sampling distribution of a test statistic by random sampling with replacement from the sample data, most often with the purpose of deriving robust estimates of the standard error of a test statistic and confidence intervals for a population parameter such as a mean, proportion, correlation coefficient or regression coefficient. It may also be used for testing hypotheses in statistical inference. It is often used as a robust alternative to inference based on parametric assumptions when those assumptions are in doubt, or where the standard error is difficult or impossible to determine.

The naevi example in the previous section was described as repeated sampling from the population of second year medical students at UTAS in 2000. If that population was instead regarded as a sample of medical students, the repeated sampling procedure would amount to re-sampling (bootstrapping) provided that the re-sampling was at random with replacement..

Comparison of jackknifing and bootstrapping.

The jackknife, originally used for bias estimation, is a specialized method that provides estimates of the sampling variance of the test statistic. This can be enough for basic statistical inference (hypothesis testing, p-values, confidence intervals).

The bootstrap, on the other hand, first estimates the whole distribution of the test statistic and then, if necessary, the sampling variance can be calculated from that. While powerful and easy, this can become highly computer intensive.

5. Approximation of the sampling distribution by a theoretical distribution

In the previous methods, the sampling distribution was completely enumerated or approximated. This enabled tests of inference about a population parameter based on the particular value of a test statistic. The information contained in the sampling distribution made it possible to calculate the percentage of all possible values of the test statistic that are at least as large as the value observed.

An alternative approach is to assume that the sampling distribution is well-approximated by a known theoretical distribution. That makes it possible to use mathematical formula appropriate to that distribution to estimate the standard error of many test statistics.

Two very useful standard error formulae

Two important standard errors are those for the sample mean and proportion. They are based on the assumption that their sampling distributions are normal in large samples, as will increasingly be the case in larger and larger samples if sampling is random.[6].

The standard error of the sample mean

The standard error of the sample mean $\bar{x}$ in a sample of size n drawn from a population with mean μ_χ and variance $\sigma_\chi{}^2$ is:

$$SE\left(\bar{x}\right) = \sqrt{\frac{\sigma_\chi{}^2}{n}} = \frac{\sigma_\chi}{\sqrt{n}}$$

The standard error of the sample proportion

Consider a characteristic that has nominal (unordered) attributes, such as eye colour, and suppose we are interesting in estimating the proportion π_X of persons in the population who possess one particular attribute (eg blue eyes) of the characteristic (eye colour).

The sample proportion $\hat{p}$ is:

$$\hat{p} = \frac{\text{number of subjects possessing the attribute}}{\text{total number of subjects in the sample}}$$

The sample proportion is closely related to the sample mean. It is the mean ($\bar{x}$) of a binary (0/1) response variable X coded as 1 = blue eyes and 0 = eyes of any other colour. The SD formula for this binary variable reduces to $s_x = \hat{p}\left(1-\hat{p}\right)$.

The standard error of the sample proportion $\hat{p}$ in a sample of size n drawn from a population with population proportion π is:

$$SE\left(\hat{p}\right) = \sqrt{\frac{\pi\left(1-\pi\right)}{n}}$$

That it has this form can be verified from the formula $SE\left(\bar{x}\right) = \frac{\sigma_X}{\sqrt{n}}$ for a binary variable.[7]

[6] By the **Central Limit Theorem**, the sampling distribution of any test statistic that can be interpreted as a sum of random variables is approximately normal when the sample is large. The sample mean and sample proportion can be interpreted a sum of random variables, and hence their sampling distributions are approximately normal in large samples.

Two other standard error formulae

A general formula for the variance of the difference between two random variables X and Y is:

$$Var(X-Y)=Var(X)+Var(Y)-2\times Cov(X,Y)$$

where $Cov(X,Y)$ denotes the covariance[8] between X and Y, and $\bar{X}$ and $\bar{Y}$ denote the means of X and Y respectively. If X and Y are independent variables, $Cov(X,Y)=0$ and:

$$Var(X-Y)=Var(X)+Var(Y)$$

The standard error of a difference between the means of independent groups

Label the means of a characteristic in two independent population groups as $\bar{\chi}_1$ and $\bar{\chi}_2$, and the group variances as ${\sigma_1}^2$ and ${\sigma_2}^2$. Samples of size n_1 and n_2 are drawn, and sample means $\bar{x}_1$ and $\bar{x}_2$ are calculated. The test statistic for testing the difference $(\bar{\chi}_1-\bar{\chi}_2)$ between the population means is the difference $(\bar{x}_1-\bar{x}_2)$ between the sample means. From the formula for the difference between independent random variables:

Population group	Population mean	Population variance	Sample size	Sample mean
1	$\bar{\chi}_1$	${\sigma_1}^2$	n_1	$\bar{x}_1$
2	$\bar{\chi}_2$	${\sigma_2}^2$	n_2	$\bar{x}_2$

$$SE(\bar{x}_1-\bar{x}_2)=\sqrt{\left[SE(\bar{x}_1)\right]^2+\left[SE(\bar{x}_2)\right]^2}$$

and using the formula for the standard error of the sample mean:

$$SE(\bar{x}_1-\bar{x}_2)=\sqrt{\left(\frac{\sigma_1}{\sqrt{n_1}}\right)^2+\left(\frac{\sigma_2}{\sqrt{n_2}}\right)^2}$$

The standard error of a difference between proportions of independent groups

Label the proportions of persons possessing an attribute in two independent population groups as π_1 and π_2. Random samples of size n_1 and n_2 are drawn, and the sample proportions $\hat{p}_1$ and $\hat{p}_2$ are calculated. The test statistic for testing the difference $(\pi_1-\pi_2)$ between the population proportions is the difference $(\hat{p}_1-\hat{p}_2)$ between the sample proportions. From the formula for the difference between independent random variables:

Population group	Population proportion	Sample size	Sample proportion
1	π_1	n_1	$\hat{p}_1$
2	π_2	n_2	$\hat{p}_2$

$$SE(\hat{p}_1-\hat{p}_2)=\sqrt{\left[SE(\hat{p}_1)\right]^2+\left[SE(\hat{p}_2)\right]^2}$$

and using the formula for the standard error of a sample proportion:

$$SE(\hat{p}_1-\hat{p}_2)=\sqrt{\frac{\pi_1\times(1-\pi_1)}{n_1}+\frac{\pi_2\times(1-\pi_2)}{n_2}}$$

---

[7] It is a simplified form of the formula for $SE(\bar{x})$ that is possible if each x_i is restricted to the values 0 or 1.

[8] $Cov(X,Y)=\frac{\sum_{i=1}^{N}(X_i-\bar{X})\times(Y_i-\bar{Y})}{N}=\frac{(X_1-\bar{X})\times(Y_1-\bar{Y})+(X_2-\bar{X})\times(Y_2-\bar{Y})+\ldots+(X_N-\bar{X})\times(Y_N-\bar{Y})}{N}$

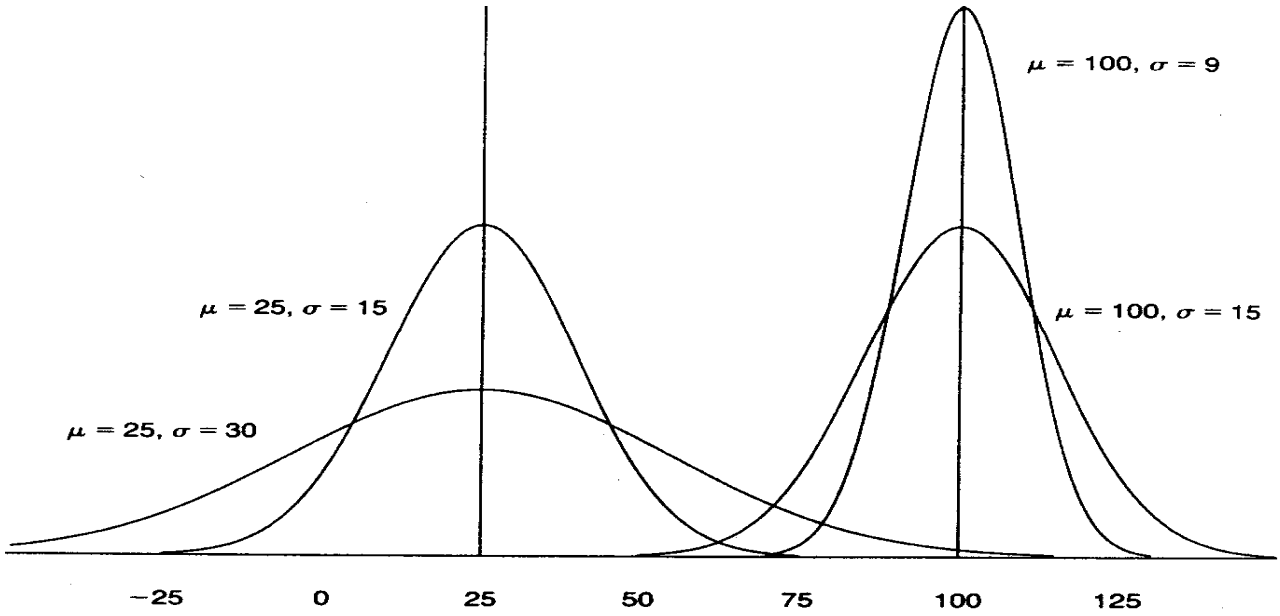
6. Large-sample properties of the sampling distribution of the sample mean

The Central Limit Theorem

The Central Limit Theorem can be used to justify assumptions about the sampling distribution of the sample mean.[9] One result is as follows. Suppose S_n is a statistic that can be expressed as the sum of n identically and independently distributed random variables $X_1, X_2, ... X_n$ each with mean μ_X and SD σ_X. Then as n tends towards infinity, the probability distribution of $\frac{S_n - \mu_X}{\sigma_X / n}$ tends towards the normal distribution with zero mean and unit SD.

The Normal distribution

The normal distribution is continuous, bell-shaped and symmetric about its mean. It is important because the distribution of many measurements (cholesterol, blood pressure, height, weight, the logarithm of incubation periods of communicable diseases) have this shape.



Examples of normal distributions.

The normal distribution is characterised by its mean μ and SD σ. Published tables are available for the **standard normal distribution** with $\mu = 0$ and $\sigma = 1$. To use the tables to calculate probabilities for a random variable X that is normally distributed with mean μ_X and SD σ_X, we calculate the "Z-score" of the particular value (say x) of X:

$$Z = \frac{x - \mu_X}{\sigma_X}$$

The "Z-score" gives the number of SDs by which x exceeds μ_X, the mean of X. In a normal distribution:

10% of values (5.0% in each tail) are more than 1.64 SDs from the mean
5% of values (2.5% in each tail) are more than 1.96 SDs from the mean
1% of values (0.5% in each tail) are more than 2.58 SDs from the mean.

[9] The Central Limit Theorem (CLT) is not a single theorem. It is a collection of related theorems that establish conditions under which a sum of random variables will be distributed in accordance with a specific probability distribution (such as the normal).

7. The 95% confidence interval

Recalling that 5% of values (2.5% in each tail) in a normal distribution fall more than 1.96 SDs from the mean, and assuming that the sampling distribution of the sample mean is normally distributed[10], a 95% confidence interval for the population mean μ_X is constructed as:

$$95\% \text{ CI for } \mu_X = \bar{x} \pm Z_{(1-\alpha/2)}\left[SE\left(\bar{x}\right)\right]$$

where $Z_{(1-\alpha/2)}$ denotes the $100\left(1-\alpha/2\right)$ percentile of the unit normal distribution.

For 95% confidence limits, $Z_{(1-0.05/2)} = 1.96$ so that the lower limit is $\bar{x} - 1.96\left[SE\left(\bar{x}\right)\right]$ and the upper limit is $\bar{x} + 1.96\left[SE\left(\bar{x}\right)\right]$. We are 95% confident that μ_X lies within these limits.

Similarly, a 95% confidence interval for the population proportion π is constructed as:

$$95\% \text{ CI for } \pi = \hat{p} \pm 1.96\left[SE\left(\hat{p}\right)\right]$$

The choice of 95% for confidence intervals, rather than 90% or 99%, is arbitrary, but common.[11]

Why are we 95% confident

Suppose that we conduct a study in order to make inferences about the mean height of Australians. It involves choosing a random sample of Australians from the electoral roll, and measuring their height.

We conduct the study and obtain a value of the sample mean, say $\bar{x} = 171$ cm. Proceeding as outlined above, we would calculate 95% confidence limits in make an inference about the likely value of the population mean. The 95% confidence interval would be centred at $\bar{x} = 171$ cm.

Suppose now that we repeated the study for some reason. We would obtain another value of the sample mean, say $\bar{x} = 168$ cm, and another 95% confidence interval that is now centred at $\bar{x} = 168$ cm. It would be a little different from the previous one.

There are two things to notice here. Firstly, the sample mean was different. Secondly, so too was the 95% confidence interval.

In fact, by repeating this study thousands of times, we would obtain thousands of different 95% confidence intervals.

The basis of our 95% confidence in the results from any one study is that **95% of those confidence intervals would include the population mean**. We interpret that to mean that there is a 95% probability that the 95% confidence limit we obtain on any single performance of the study will contain the population mean.

---

[10] It will be if the sample is large enough (Central Limit Theorem), or if the variables that are summed to calculate the mean are normally distributed.

[11] The eminent British statistician, Sir Ronald Fisher, once remarked over coffee that a 1-in-20 occurrence was a rare event. That remark prompted the "p = 0.05 rule" of statistical significance and the choice of 95% confidence intervals. There is no scientific basis for it.