Part (D): Sampling distributions and confidence intervals

(Reading: Wild & Seber, Chapters 6,7, Freund, Chapters 8,11)

13 Sampling distributions

13.1 Introduction

Suppose we want to estimate the expected (mean) yield (μ) of a certain plant variety. We grow 100 plants and measure their yields:

Data: $x_1, x_2, \ldots, x_{100}$

We can estimate μ by

$$\hat{\mu} = \bar{x} = \frac{1}{100}\sum_{1}^{100} x_i$$

Suppose that in our experiment

$$\bar{x} = 1.56, \qquad s^2 = \frac{1}{99}\left\{\sum x_i^2 - \frac{(\sum x_i)^2}{100}\right\} = 0.26.$$

Then clearly $\hat{\mu} = 1.56$kg. But, how accurate is this as an estimate of μ?

To quantify the 'accuracy'of $\hat{\mu} = \bar{x}$ as an estimate of μ we require a probability model.

Model assumptions

We assume that each measured yield is the outcome of a random experiment such that:

(A) The outcome of each experiment is independent of all other experiments.

(B) The probability distribution for the result of each experiment is the same for all experiments.

Recall that (A) and (B) imply that the data represent a realisation of independent, identically distributed (i.i.d) random variables

$$X_1, X_2, \ldots, X_n$$

where X_k is the outcome of the k^{th} experiment, $n =$ sample size ($= 100$ here).

The probability model consists of:

- The variables $\{X_1, X_2, \ldots, X_n\}$ which are called a random sample, and
- The common distribution of $\{X_1, X_2, \ldots, X_n\}$ which is called the population distribution:

  $\mu = E(X) =$ population mean

  $\sigma = \sqrt{\text{var(X)}} =$ population standard deviation.

An important distinction:

Before we carry out the experiment we don't know what the measured yields will be. Therefore they are random variables $X_1, X_2, \ldots, X_n$ (UPPER CASE).

After the experiment we obtain actual observed values $x_1, x_2, \ldots, x_n$ (lower case).

Now, since $X_1, X_2, \ldots, X_n$ are random variables, then

$$\bar{X} = \frac{1}{n}\sum_{1}^{n} X_i$$

is also a random variable.

The value $\bar{x} = 1.56$kg observed in our plant-yield experiment is a realisation of the r.v. $\bar{X}$.

Each time we conduct the experiment (i.e. grow 100 plants and measure yield) we will obtain a different observed value $\bar{x}$.

Simulation study:

Suppose the population distribution is N$(5,4)$ and the sample size is $n = 25$. Then the sample mean (as a r.v.) is

$$\bar{X} = \frac{1}{n}(X_1 + X_2 + \ldots + X_{25})$$

where $X_i \sim N(5,4)$ (i.i.d).

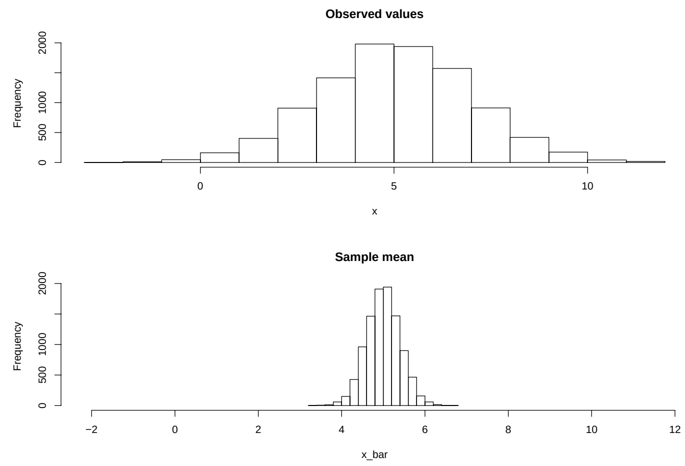
Now, we can generate (using a computer) a realisation (sample)

$$x_1, x_2, \ldots, x_{25}$$

and calculate $\bar{x}$.

Repeat this many times to generate many different realisations and build up a picture of the distribution of $\bar{X}$.

If we can understand the distribution of $\bar{X}$ then we can begin to quantify how accurate $\bar{x}$ is likely to be as an estimate of the population mean.



Notice that:

a) values of $\bar{x}$ cluster around the population mean, 5

b) values of $\bar{x}$ are less variable than individual observations.

14 The distribution of $\bar{X}$ and the Central Limit Theorem

Again, let $X_1, X_2, \ldots, X_n$ be a random sample of size n. The population distribution is not completely specified, but we assume that X has

$$E(X) = \mu; \qquad \mathrm{var}(X) = \sigma^2 \qquad \text{(both finite)}.$$

What can we say about the distribution of $\bar{X} = \frac{1}{n}\sum_1^n X_i$?

14.1 Mean and variance of $\bar{X}$

We have the following results:

$$\begin{aligned} \mathrm{E}(\bar{X}) &= \mathrm{E}\left\{\frac{1}{n}(X_1 + X_2 + \ldots + X_n)\right\} \\ &= \frac{1}{n}\{\mathrm{E}(X_1) + \mathrm{E}(X_2) + \ldots + \mathrm{E}(X_n)\} \\ &= \frac{1}{n}(\mu + \mu + \ldots + \mu) \\ &= \frac{1}{n} \times n\mu = \mu \end{aligned}$$

$\Rightarrow$ The average of $\bar{X}$ over many experiments is the 'true' population mean.

$$
\begin{aligned}
\mathrm{var}(\bar{\mathrm{X}}) &= \mathrm{var}\left\{\frac{1}{\mathrm{n}}(\mathrm{X}_1 + \mathrm{X}_2 + \ldots + \mathrm{X}_\mathrm{n})\right\} \\
&= \frac{1}{n^2}\{\mathrm{var}(\mathrm{X}_1) + \mathrm{var}(\mathrm{X}_2) + \ldots + \mathrm{var}(\mathrm{X}_\mathrm{n})\} \qquad (X_{i's}\text{independent}) \\
&= \frac{1}{n^2} \times n\sigma^2 = \frac{\sigma^2}{n}
\end{aligned}
$$

$\Rightarrow$ The variance of $\bar{X}$ is inversely proportional to the sample size n.

[s.d. of $\bar{X} = \frac{\sigma}{\sqrt{n}}$]

14.2 The Central Limit Theorem (CLT)

(Reading: Wild & Seber, 7.2, Freund, 8.2)

Although we have not specified exactly the population distribution (X), we can say a lot more about the distribution of $\bar{X}$.

Theorem (CLT):

The distribution of $\bar{X}$ (the sample mean as a r.v.) is approximately Normal with mean μ and variance $\frac{\sigma^2}{n}$ $\left[\Rightarrow \text{s.d.} = \frac{\sigma}{\sqrt{n}}\right]$.

Remarks:

i) If the population is $N(\mu, \sigma^2)$ then $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$ exactly.

ii) Otherwise the quality of the approximation increases with the sample size n.

We illustrate this result with a simulation study:

Case 1

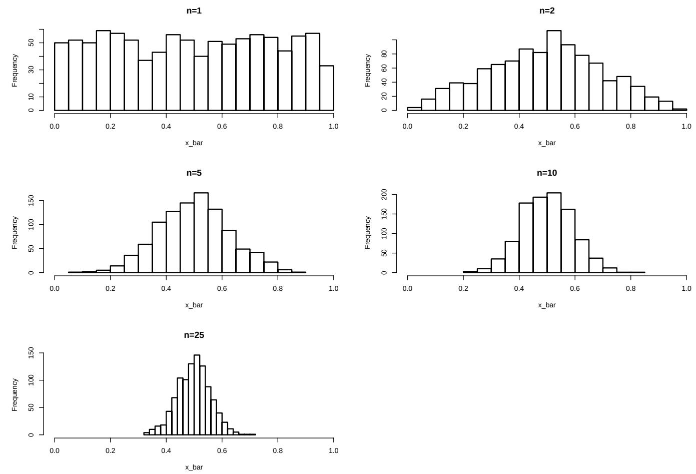
Population distribution is $U(0,1)$. [Recall $\mathrm{E}(\mathrm{X}) = 0.5, \mathrm{var}(\mathrm{X}) = \frac{1}{12}$.]

Simulate random sample $X_1, X_2, \ldots, X_n$.

Look at distribution of $\bar{X}$ by forming a histogram from 1000 such samples of size n.

Do this for $n = 1, 2, 5, 10, 25$.

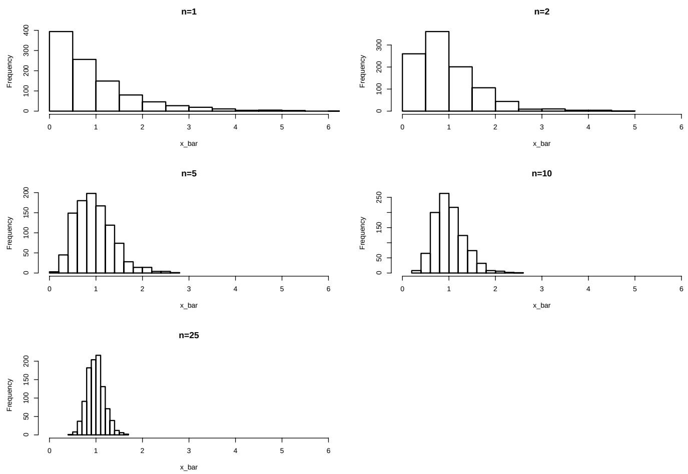
Sample mean of $U(0,1)$ distribution



Case 2

Same as before with population distribution Exp(1). $[\mathrm{E(X)} = 1, \mathrm{var(X)} = 1]$

Sample mean of Exp(1) distribution



Note how the histograms of the sample mean appear to look 'Normal' as n increases.

iii) The CLT also tells us what the distribution of $\sum_i^n X_i$ looks like:

$$\sum_i^n X_i = n\bar{X} \overset{approx}{\sim} N(n\mu, n\sigma^2)$$

[To see this, recall that $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$].

Example

The weight of a certain variety of apple (in grams) has a distribution with mean $\mu = 150$ and variance $\sigma^2 = 100$. A box is packed with 40 randomly selected apples. What is the probability that the total weight of apples exceeds 6.1 kg?

Solution...

Application (Normal approximation to binomial)

Show that if the r.v. $X \sim$ bin(n, p), then

$$X \overset{approx}{\sim} N(np, np(1-p))$$

as $n \to \infty$.

Proof...

Example

A gambler plays 99 times at roulette and always bets on red. What is the probability that he wins at least 50 times?

[*Roulette: P(red)* $= \frac{18}{37} = 0.4865$]

Solution...

15 Constructing confidence intervals

Let $X_1, X_2, \ldots, X_n$ denote a random sample (i.i.d.) from a population with unknown mean μ. We assume for the moment that the population variance σ^2 is known.

From the CLT we know that

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$$

which implies that

$$\frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim N(0, 1)$$

The left hand side is a random variable, given as a function of the data and involving the unknown parameter μ.

We can use it as a 'pivotal' quantity to derive the probability statement

$$P\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

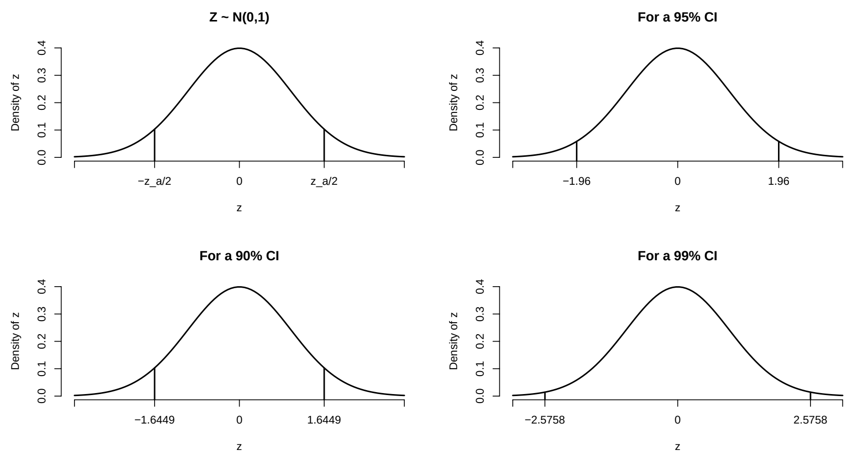
where $z_{\frac{\alpha}{2}}$ is the 'percentage' point of the $N(0,1)$ distribution such that

$$P\left(Z > z_{\alpha/2}\right) = 1 - \Phi\left(z_{\alpha/2}\right) = \frac{\alpha}{2}.$$

For example, $z_{0.025} = 1.96$, since

$$P(Z > 1.96) = 1 - \Phi(1.96) = 0.025.$$

(Table 5, Lindley & Scott, p.35, also see picture).



In the case of $z_{0.025}$ we say that there is a $(1-\alpha)100\% = 95\%$ chance that the random interval

$$\left(\bar{X} - 1.96\,\frac{\sigma}{\sqrt{n}},\ \ \bar{X} + 1.96\,\frac{\sigma}{\sqrt{n}}\right)$$

'covers' or contains the true population mean μ.

We call the above interval a 95% confidence interval (C.I) for the mean μ.

[Note that in general, and for different $z_{\alpha/2}$, we can determine appropriate $(1-\alpha)100\%$ confidence intervals.]

For a particular realisation (i.e. a set of observations)

$$x_1, x_2, \ldots x_n$$

we say that the interval

$$\left(\bar{x} - 1.96\,\frac{\sigma}{\sqrt{n}},\ \ \bar{x} + 1.96\,\frac{\sigma}{\sqrt{n}}\right)$$

is the observed 95% CI.

A subtle but important point:

Given a particular realisation it would be wrong to state that

$$P\left(\bar{x} - 1.96\,\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + 1.96\,\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

No random variables involved!

We really mean:

The interval $\left(\bar{x} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}\right)$ is a realisation from a population of intervals, 95% of which contain the true value μ.

Or:

If we obtain a large number of such intervals (with a different independent sample each time), we expect 95% of them to contain the true value of μ.

Therefore we are fairly confident that a single such interval contains μ.

(But if we have been 'unlucky', then it doesn't contain μ.)

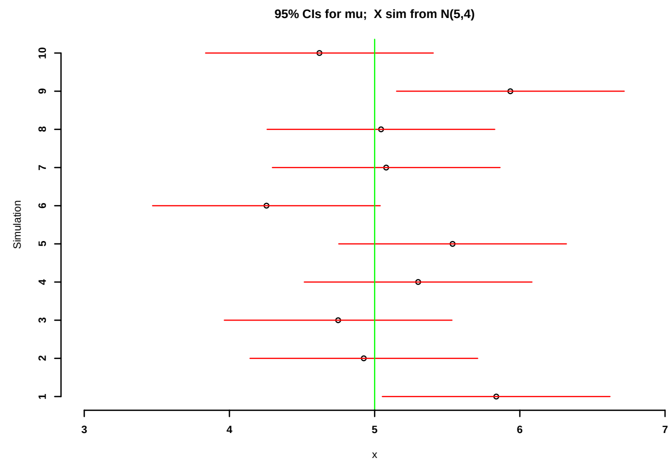
Simulation study

Recall our simulation study for the case of $n = 25$, $X \sim N(5, 2^2)$.

Let's look at the 95% CIs over 10 realisations. (We suppose that we don't know $\mu = 5$, but we do know $\sigma^2 = 4$.) Then given the sample mean $\bar{X}$ we obtain

$$\left(\bar{X} - 1.96\,\frac{2}{5},\ \ \bar{X} + 1.96\,\frac{2}{5}\right)$$

as our 95% CI. [Roughly $(\bar{X} - 0.8, \bar{X} + 0.8)$]



Here we have been 'unlucky' twice, as 2 out of the 10 CIs do not contain the true value of $\mu = 5$.

In the long run we would find that $\frac{1}{20}$ of the realisations (CIs) did not contain μ.

Example

The times (in seconds) of a certain chemical reaction is known to be distributed as $N(\mu, 0.5^2)$, where μ is unknown. The times of a random sample of 10 such reactions are measured to be:

$$3.9,\ 4.7,\ 6.1,\ 5.2,\ 5.4,\ 4.8,\ 4.5,\ 5.0,\ 4.7,\ 4.9$$

Calculate $50\%, 90\%$ and 99% CIs for μ.

Solution...

16 Constructing CIs when σ is unknown

In many practical situations we don't know the population variance σ^2. However, given the observations $x_1, x_2, \ldots, x_n$, we can calculate the sample variance

$$s^2 = \frac{1}{(n-1)}\left\{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right\}.$$

and the standard deviation $s = \sqrt{s^2}$.

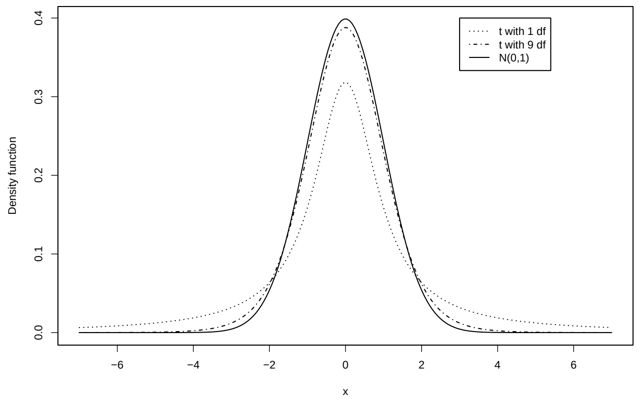
We can use s instead of σ to construct a CI, using the following distribution.

The t_ν distribution

The t distribution (or *Student's* t – named after the pseudonym 'Student' that W.S. Gosset used) is a continuous distribution, with shape similar to that of the $N(0,1)$ distribution.

The distribution is characterised by a parameter ν, called the *degrees of freedom* of the distribution, and we denote it by t_ν.

Its probability density function is plotted in the graph below.



Remarks

1. The t distribution is symmetric around zero, with longer tails than the $N(0,1)$.

2. As $\nu \to \infty$, the t distribution approaches the $N(0,1)$ distribution.

3. The values of its cdf are tabulated (e.g. see *NCST* p42–45).

Theorem:

If the population distribution is $N(\mu, \sigma^2)$ and the sample size is n, then

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}.$$

This is the t-distribution with $n-1$ degrees of freedom. As before, $\bar{X}$ and S denote the sample mean and sample s.d. and n is the sample size.

Remark:

This result holds approximately for non-normal distributions. This is important because in practice we can never know that the data come from (exactly) a Normal population.

To calculate a CI for μ (e.g. a 95% CI) from $\bar{X}$ and S we need to find t such that

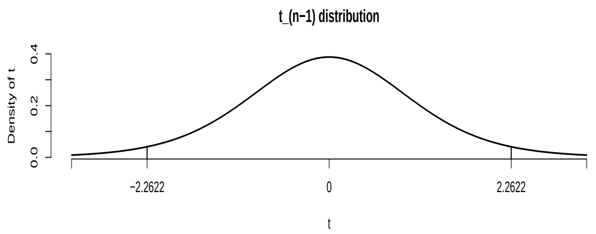
$$P\left(-t < \frac{\bar{X} - \mu}{S/\sqrt{n}} < t\right) = 0.95$$

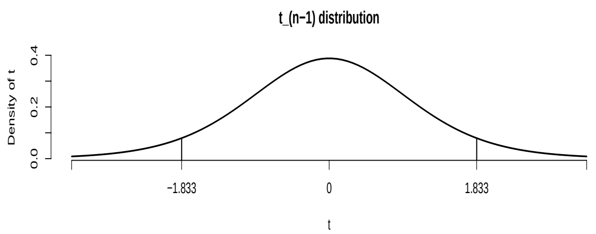
$$\Leftrightarrow \quad P\left(\bar{X} - t\frac{S}{\sqrt{n}} < \mu < \bar{X} + t\frac{S}{\sqrt{n}}\right) = 0.95$$

where t is 'percentage' the point of the t_{n-1} distribution such that

$$P(T > t) = 0.025, \qquad \text{with } T \sim t_{n-1}.$$

We usually denote this point by $t_{n-1,0.025}$.





These values can be obtained from Table 10 (Lindley & Scott, p.45)

Note: The percentage points decrease as ν (no. of degrees of freedom) increases (t-distribution becomes narrower).

Also, as $\nu \to \infty$, t-distribution approaches the Normal distribution.

Example

Recall the chemical reaction times example. Data:

$$3.9,\ 4.7,\ 6.1,\ 5.2,\ 5.4,\ 4.8,\ 4.5,\ 5.0,\ 4.7,\ 4.9$$

Suppose we don't know σ^2. Find a 90% CI for the population mean μ.

Solution...

Example

Return to the plant-yield example. Here we had 100 measured yields with sample mean $\bar{x} = 1.56$ and sample variance $s^2 = 0.26$.

We can quantify the accuracy of our estimate $\hat{\mu} = \bar{x} = 1.56$ by calculating e.g. a 95% CI for μ as

$$\left(1.56 - t_{99,0.025}\frac{s}{\sqrt{100}},\ \ 1.56 + t_{99,0.025}\frac{s}{\sqrt{100}}\right)$$

Now $t_{99,0.025} \approx 2.0$, $s = \sqrt{0.26} = 0.51$.

Therefore our 95% CI is

$$(1.458, 1.662)$$

'We are 95% confident that the value of μ lies in the interval calculated.' On 95% of times that we carry out the experiment we will be correct.

17 Comparison of two populations

In many practical situations we are interested in comparing the means of 2 populations to investigate whether or not they are different.

To do this we can calculate a CI for $\mu_1 - \mu_2$ where μ_1, μ_2 are the unknown means of the 2 populations. We then consider whether the value 0 lies in the CI.

[$\mu_1 - \mu_2 = 0 \Leftrightarrow$ means are equal]

We consider 2 cases.

17.1 σ_1^2 and σ_2^2 known

Let $X_{11}, X_{12}, \ldots, X_{1n_1}$ and $X_{21}, X_{22}, \ldots, X_{2n_2}$ be the random samples from the two populations with unknown means μ_1, μ_2 and known variances σ_1^2, σ_2^2. The sample means are

$$\bar{X}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} X_{1i}, \qquad \bar{X}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} X_{2i}$$

Now, assuming the populations are Normal we have

$$\bar{X}_1 \sim N\left(\mu_1, \frac{\sigma_1^2}{n_1}\right), \qquad \bar{X}_2 \sim N\left(\mu_2, \frac{\sigma_2^2}{n_2}\right)$$

and further assuming that the two samples are independent of each other we have

$$\bar{X}_1 - \bar{X}_2 \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right).$$

We can construct e.g. a 95% CI for $\mu_1 - \mu_2$ using

$$P\left(\bar{X}_1 - \bar{X}_2 - 1.96\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \mu_1 - \mu_2 < \bar{X}_1 - \bar{X}_2 + 1.96\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right) = 0.95$$

i.e. the 95% CI is

$$\left(\bar{X}_1 - \bar{X}_2 - 1.96\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}},\ \bar{X}_1 - \bar{X}_2 + 1.96\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$$

Example

The heights (in ft) of two species of plant are known to be normally distributed with unknown means μ_1 and μ_2 and variances $\sigma_1^2 = 0.25, \sigma_2^2 = 0.36$. Independent samples of size $n_1 = 20, n_2 = 25$ are drawn from the two populations. The observed sample means are $\bar{x}_1 = 4.4$ and $\bar{x}_2 = 5.2$. Calculate a 95% CI for the difference in the means $\mu_1 - \mu_2$.

Solution...

17.2 σ_1^2 and σ_2^2 unknown

If we don't know the population variances σ_1^2 and σ_2^2 but both samples are large (say $n_1, n_2 \geq 30$) then we can use the sample standard deviations s_1 and s_2 to substitute for σ_1 and σ_2 and construct a CI as for the case where the variances are known.

If either or both samples are small, then things are more complicated:

If we can assume that $\sigma_1^2 = \sigma_2^2$ (i.e. the unknown population variances are equal) then we can proceed as follows.

Let S_1^2 and S_2^2 denote the 2 sample variances. These can be combined to give a pooled estimator of σ^2:

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

Our $(1-\alpha)100\%$ CI for $\mu_1 - \mu_2$ can be constructed as

$$\left(\bar{X}_1 - \bar{X}_2 - t\, S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}},\quad \bar{X}_1 - \bar{X}_2 + t\, S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right)$$

where $S_p = \sqrt{S_p^2}$ is the pooled sample standard deviation (estimator of σ) and

$$t = t_{n_1+n_2-2,\ \alpha/2}$$

is the point of the $t_{n_1+n_2-2}$ distribution such that

$$P(T > t_{n_1+n_2-2,\ \alpha/2}) = \frac{\alpha}{2}, \qquad \text{with } T \sim t_{n_1+n_2-2}.$$

Example

A random sample of 10 cigarettes of type 1 had an average nicotine content of 3.1 milligrams with a standard deviation of 0.5 mg. A sample of 8 cigarettes of type 2 had mean and s.d. 2.7 mg and 0.7 mg respectively.

Assuming that the two sets of data are independent random samples from normal populations with the same variance, construct a 95% CI for the difference between the mean nicotine contents of the brands.

Solution...

18 Confidence intervals for unknown proportions

Assume that we want to estimate a proportion of a population having a specific characteristic.

This can be expressed as a probability (e.g. probability of a car failing a safety check), percentage (e.g. percentage of votes in a YES/NO referendum), or rate (e.g. mortality rate of a disease).

In many cases, we can express the above quantities as the probability p in a binomial distribution.

Recall that if $X \sim \text{Bin}(n,p)$, then CLT gives

$$X \overset{approx}{\sim} N(np, np(1-p))$$

This also implies that

$$\frac{X}{n} \overset{approx}{\sim} N\left(p, \frac{p(1-p)}{n}\right)$$

But notice that $\frac{X}{n}$ is an estimate of the unknown probability (proportion) p, and we write

$$\frac{X}{n} = \hat{P}.$$

Then we also have that

$$\frac{\hat{P}-p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0,1)$$

It follows that we can construct a $(1-\alpha)100\%$ CI for p, based on:

$$P\left(-z_{\alpha/2} < \frac{\hat{P}-p}{\sqrt{\frac{p(1-p)}{n}}} < z_{\alpha/2}\right) = 1-\alpha$$

$$\Leftrightarrow \quad P\left(\hat{P} - z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} < p < \hat{P} + z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}\right) = 1-\alpha$$

Now, as this expression involves the unknown true proportion p, we can further approximate it by using the estimate $\hat{P}$ to obtain the $(1-\alpha)100\%$ CI:

$$\left(\hat{P} - z_{\alpha/2}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}}, \quad \hat{P} + z_{\alpha/2}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}}\right)$$

19 CIs for differences between proportions

Suppose now that we want to estimate the difference between two proportions p_1 and p_2 based on two samples of size n_1 and n_2 from two binomial populations.

For large samples, we can use the approximation

$$Z = \frac{(\hat{P}_1 - \hat{P}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \overset{approx}{\sim} N(0,1)$$

to obtain a a $(1-\alpha)100\%$ CI for the difference $p_1 - p_2$ of the form

$$(\hat{P}_1 - \hat{P}_2) \pm z_{\alpha/2}\sqrt{\frac{\hat{P}_1(1-\hat{P}_1)}{n_1} + \frac{\hat{P}_2(1-\hat{P}_2)}{n_2}}$$

Notice that, as before, the unknown proportions p_1 and p_2 have been substituted in the variance with their estimates $\hat{P}_1$ and $\hat{P}_2$, to give the above approximate CI.

Example

(a) A poll was taken of University students before a student election. Of the 78 *male students contacted,* 33 *said they would vote for candidate A. Obtain a* 95% *CI for the proportion of male voters in the University population in favour of this candidate.*

(b) Consider now a second sample of 86 *female students, of which* 26 *said they would vote for candidate A. By obtaining an appropriate* 95% *CI, can you support the view that the percentage of voters for candidate A is the same among male and female students?*

Solution ...