Chapter 4

Estimation and Hypothesis Testing

4.1 Introduction

A probability model for the distribution of variable of interest will usually depend on one or more unknown *parameters*. For example, if we propose a normal distribution for a particular variable, then we need to know the mean μ and standard deviation σ of that normal distribution, in order to use our model to make predictions about future observations.

Just as we used sample data in §3, to assess whether a particular distributional model is appropriate for a variable of interest, we can also use sample data to estimate the parameters of our model.

4.2 Estimating a mean

The most straightforward situation is where the parameter of interest is the mean of the population distribution, for example the normal parameter μ or the exponential parameter β. There are also cases where it may be sufficient to estimate the mean of a population without necessarily specifying a complete distributional model for the population.

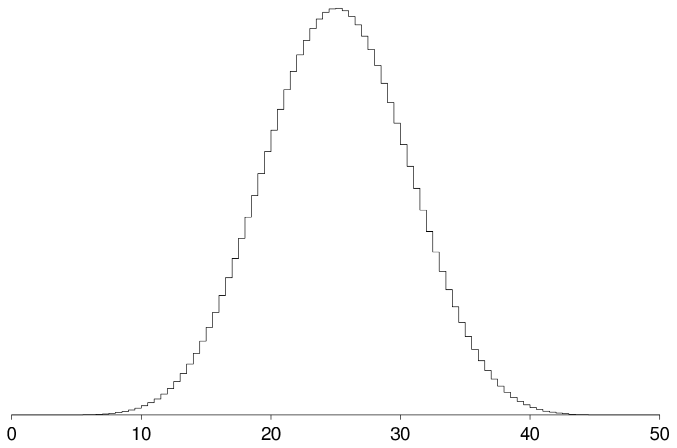
Suppose that we have a sample of size n, $x_1, \ldots x_n$ from a population of interest. It seems obvious that we should use the sample mean $(\overline{x})$ to estimate the population mean μ. This procedure, estimating a population quantity using a sample quantity, is called **point estimation**.

When we calculate a point estimate, it is important that we have some idea how accurate that estimate is likely to be. So how accurate are we, when we use the mean of a sample of size n to estimate the mean of a population distribution?

Samples from a population are variable, and therefore estimates calculated using sample data are also variable, and we can consider their distribution. When a sample of size n is observed from a distribution, the sample mean $\overline{x}$ is a single observation from the distribution of $\overline{x}$ for all such samples. An important question is 'What does the distribution of $\overline{x}$ look like?' and in particular 'How does it compare with the distribution of the original observations $x_1, x_2, \ldots$?'

The following example is artificial, but serves to illustrate the point

♡ **Example** 4.1. Suppose that the distribution of interest consists of the integers from 1 to 49, each with probability 1/49. Twice a week a 'sample' of six observations is taken from this distribution in the National Lottery. From §2, we know that there are 13 983 816 possible samples of size 6. We can illustrate the distribution of $\overline{x}$ across these possible samples by a histogram.



We can also calculate the mean and standard deviation of this distribution, 25 and 5.7735 respectively.

Note that the mean and standard deviation of the original distribution (the numbers 1 to 49, each with probability 1/49) are 25 and 14.1421 respectively.

We immediately notice three facts about the distribution of $\overline{x}$

1. It has the same mean as the original distribution.
2. It has a smaller standard deviation than the original distribution.
3. The histogram seems 'bell-shaped' suggesting that the distribution may be close to a normal distribution, even though the original distribution is far from normal.

In general Suppose also that $x_1, \ldots, x_n$ is a sample of size n from a distribution with mean μ and standard deviation σ.

Then the distribution of $\overline{x}$, the sample mean has the following three properties.

1. It also has mean μ.
2. It has standard deviation $\frac{\sigma}{\sqrt{n}}$.

   For larger sample sizes n, the distribution of sample means has smaller standard deviation, so the sample means for larger samples are less variable and generally closer to μ.

3. It is approximately normally distributed if n is large, **regardless of the shape of the original distribution**.

- This is surprising and remarkable. It is the **Central Limit Theorem**, and one of the reasons why the normal distribution is so important for data analysis.

How large must a sample be before we can assume that the sample mean $\overline{x}$ is from a normal distribution?

There is no ready answer to this question. If the original distribution is 'close to normal', then quite small samples may be adequate. Indeed if the original distribution is exactly normal, then this assumption is appropriate for any size of sample. However, for highly non-normal distributions (very skew or multimodal) larger samples will be required.

What remains true for all distributions is that the larger the sample size, the closer the distribution of sample means is to a normal distribution.

Now, when we use a sample mean $\overline{x}$ to estimate the mean μ of the underlying distribution, we know that $\overline{x}$ can be considered as a single observation from the distribution of sample means for samples of size n.

We know that the mean of this distribution is also μ, but that its standard deviation is $\sigma/\sqrt{n}$. Therefore, 'on average', $\overline{x}$ is equal to μ, the quantity which we want to estimate, so $\overline{x}$ is a sensible estimate. (This property, being 'correct on average', is called **unbiased**).

Furthermore, $\sigma/\sqrt{n}$, the standard deviation, is a measure of the spread of possible sample means around μ, and gives an indication of the error involved when we use a single sample mean $\overline{x}$ to estimate μ.

Unfortunately, $\sigma/\sqrt{n}$, the standard deviation of the distribution of $\overline{x}$, is not known, as it depends on the σ, the standard deviation of the original distribution, which is an unknown quantity. However, if we use the standard deviation, s, of the **sample** to estimate σ, we can use $s/\sqrt{n}$ as a measure of the accuracy of $\overline{x}$ as an estimate of μ.

The quantity $s/\sqrt{n}$ is called the **standard error of the mean** and should be quoted whenever a sample mean $\overline{x}$ is used to estimate a population mean μ, as an indication of the accuracy of the estimate.

In MINITAB

Stat→Basic Statistics→ Display Descriptive Statistics

4.3 Confidence interval for a mean

A better approach to estimating μ than using a single point estimate $\overline{x}$, together with the standard error $s/\sqrt{n}$ as a measure of precision, is to combine the two quantities to give a **range** of plausible values for μ. A **confidence interval** provides this.

Suppose that $\overline{x}$ is the mean of a sample of observations $x_1, \ldots, x_n$ from a distribution with mean μ and standard deviation σ.

Furthermore, suppose that either the sample size n is 'large', or the distribution of interest is close to normal.

Then we know that $\overline{x}$ is a single observation from (approximately) a normal distribution with mean μ and standard deviation $\sigma/\sqrt{n}$. We can standardise the variable $\overline{x}$ by subtracting μ and dividing by $\sigma/\sqrt{n}$. The standardised sample mean will have a standard normal distribution. We can write

$$z = \frac{\overline{x} - \mu}{\sigma/\sqrt{n}}$$

However, as σ is not known, we need to estimate it by the sample standard deviation s. Then,

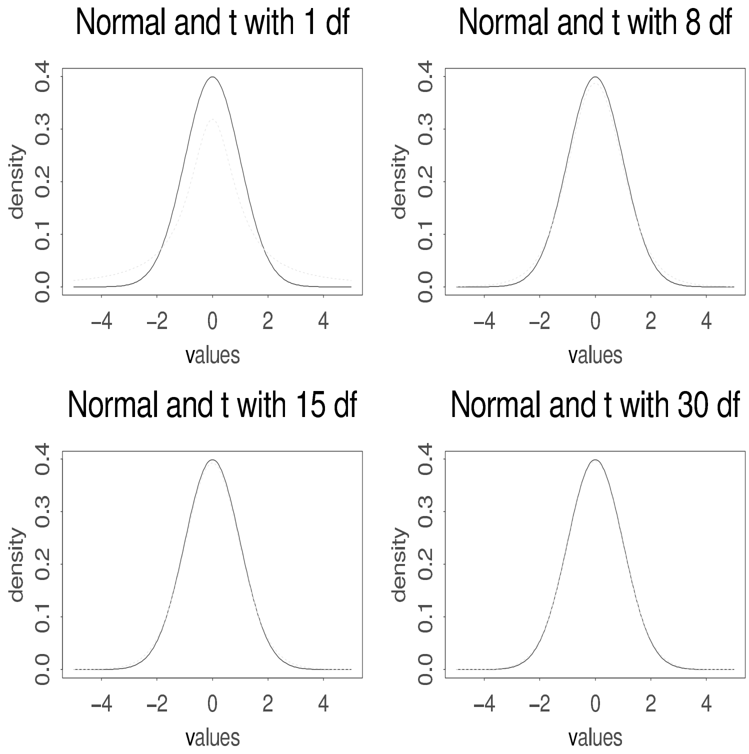
$$t = \frac{\overline{x} - \mu}{s/\sqrt{n}}$$

is an observation from a **t distribution** 'with $n-1$ degrees of freedom'.

The t distribution is a known distribution, with a density curve which looks similar to the standard normal distribution, but has a standard deviation larger than 1. The mean of a t distribution is always zero, but the standard deviation depends on the **degrees of freedom**, and is larger if the degrees of freedom is small.

When the degrees of freedom is large, the t distribution is very similar to the standard normal distribution, and its standard deviation is very close to one.

To distinguish between different t distributions (with different degrees of freedom), we denote the t distribution with k degrees of freedom by t_k.

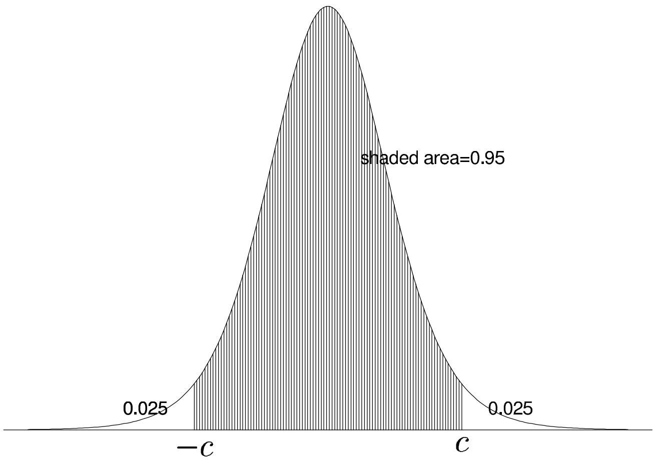


Because the t distribution is well understood, we can make statements about observations from a t distribution. As

$$t = \frac{\overline{x} - \mu}{s/\sqrt{n}}$$

is an observation from a t_{n-1} distribution. then we can (in principle), by calculating areas under the density curve of the t_{n-1} distribution, find the value c such that

$$P\left(-c \leq \frac{\overline{x} - \mu}{s/\sqrt{n}} \leq c\right) = 0.95.$$



Therefore

$$P\left(\overline{x} - c\frac{s}{\sqrt{n}} \;\le\; \mu \;\le\; \overline{x} + c\frac{s}{\sqrt{n}}\right) \;=\; 0.95.$$

or, in other words, **for 95% of samples** of size n, drawn from a distribution with mean μ, the interval between

$$\overline{x} - c\frac{s}{\sqrt{n}} \qquad \text{and} \qquad \overline{x} + c\frac{s}{\sqrt{n}}$$

will include μ.

Hence, we can calculate the endpoints of an interval, which will, for 95% of samples, include the population mean μ.
(The endpoints of the interval are often written as $\overline{x} \pm cs/\sqrt{n}$).

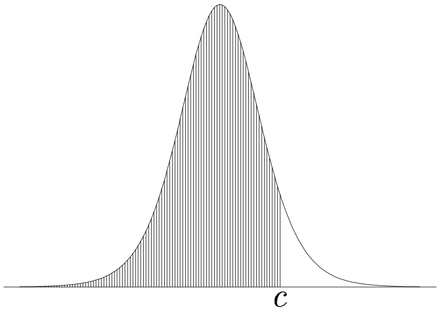
We call this interval a **95% confidence interval** for μ.

It is an interval within which we can be '95% certain' that μ lies. It provides a suumary of the 'most plausible' values for the population mean μ in light of the observed data.

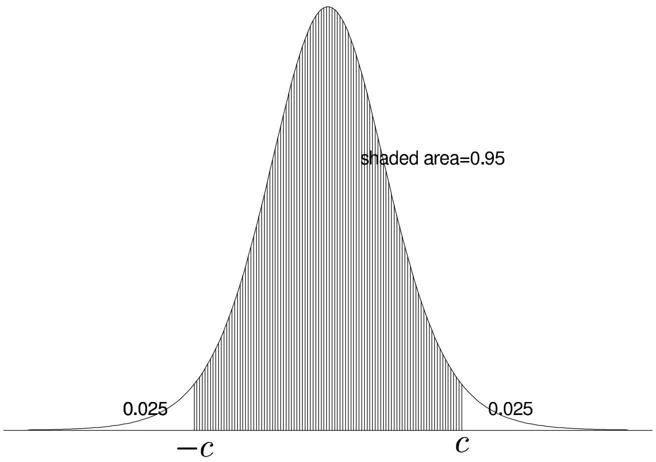
Statistical tables can be used to find c for any sample size n.

Because the t_k distribution is similar to the standard normal distribution for large values of k, then if the sample size n is large, in which case $n - 1$, the degrees of freedom will also be large, then c can be calculated using a standard normal distribution.

Table gives values of c for some $P(t_k \le c)$



	$P(t_k \le c)$				
k	0.9	0.95	0.975	0.99	0.995
1	3.08	6.31	12.71	31.82	63.66
2	1.89	2.92	4.30	6.96	9.92
3	1.64	2.35	3.18	4.54	5.84
4	1.53	2.13	2.78	3.75	4.60
5	1.48	2.02	2.57	3.36	4.03
6	1.44	1.94	2.45	3.14	3.71
7	1.41	1.89	2.36	3.00	3.50
8	1.40	1.86	2.31	2.90	3.36
9	1.38	1.83	2.26	2.82	3.25
10	1.37	1.81	2.23	2.76	3.17
11	1.36	1.80	2.20	2.72	3.11
12	1.36	1.78	2.18	2.68	3.05
15	1.34	1.75	2.13	2.60	2.95
20	1.33	1.72	2.09	2.53	2.85
25	1.32	1.71	2.06	2.49	2.79
30	1.31	1.70	2.04	2.46	2.75
40	1.30	1.68	2.02	2.42	2.70
50	1.30	1.68	2.01	2.40	2.68
60	1.30	1.67	2.00	2.39	2.66
100	1.29	1.66	1.98	2.36	2.63
∞	1.28	1.64	1.96	2.33	2.58



The area upto the point c in this graph is 0.975 and we use c to find the 95% CI. Hence

1. For 90% confidence, use the 0.95 values in the table above.

2. For 95% confidence, use the 0.975 values in the table above.

3. For 99% confidence, use the 0.995 values in the table above.

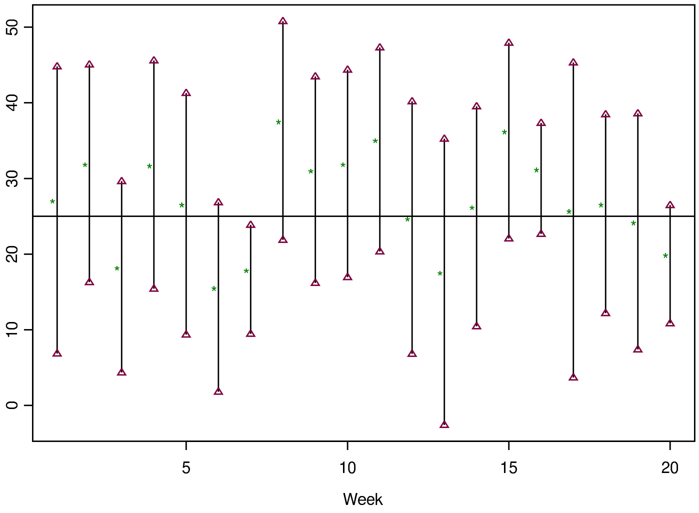
Notes

1. A confidence interval for μ is only a statement about the population mean μ. It does not say anything about other properties of the distribution of interest. In particular, we should not expect 95% of observations from a distribution to lie in the 95% confidence interval. They are likely to be much more variable.

2. If the exact value of k required is not in the table, then use the nearest value that is, or interpolate.

♡ **Example** 4.2. **Lottery example** [This is used purely as an illustration of how confidence intervals behave. The sample size of 6 is not really large enough to be happy with, but we do know in this case that the distribution of sample means is approximately normal. The confidence intervals and the sample means are plotted in the figure below.]

Date			Sample				$\overline{x}$
7/3/98	4	11	14	39	43	44	25.83
4/3/98	6	28	30	34	41	45	30.67
28/2/98	1	7	15	18	30	31	17.00
25/2/98	9	21	27	36	42	48	30.50
21/2/98	2	16	25	27	37	45	25.33
18/2/98	1	5	10	13	25	32	14.33
14/2/98	8	13	14	17	20	28	16.67
11/2/98	11	32	38	42	46	49	36.33
7/2/98	9	25	27	31	42	45	29.83
4/2/98	13	17	32	35	42	45	30.67
31/1/98	17	22	30	40	46	48	33.83
28/1/98	4	12	15	31	32	47	23.50
24/1/98	1	4	6	14	24	49	16.33
21/1/98	5	12	24	35	36	38	25.00
17/1/98	14	31	33	38	46	48	35.00
14/1/98	20	27	28	31	33	41	30.00
10/1/98	3	10	11	27	47	49	24.50
7/1/98	7	14	25	32	36	38	25.33
3/1/98	1	13	26	28	35	45	24.67
31/12/98	8	13	18	21	23	29	18.67



In MINITAB

Stat→Basic Statistics→ 1-sample t

Real examples

♡ **Example** 4.3. The file `concrete.dat` contains the compression strength (Nmm^{-2}) of 180 concrete cubes. Suppose that we are interested in the mean of the distribution of compression strength of all such cubes. The sample size of 180 is large, so there is no problem here.

The sample mean is $\bar{x} = 61.098$, the sample standard deviation is $s = 3.963$, and therefore, as the sample size $n = 180$, the standard error $s/\sqrt{n} = 0.295$. The value of the constant c for a t_{179} distribution is 1.9733 (approximately the same as for a standard normal).

Therefore, a 95% confidence interval for μ, the mean compression strength of all such cubes is $(60.515, 61.681)$.

This interval may also be presented as

$$60.515 \leq \mu \leq 61.681 \qquad \text{or} \qquad 61.098 \pm 0.583$$

A 99% confidence interval for μ is $(60.329, 61.867)$.

A 90% confidence interval for μ is $(60.609, 61.586)$.

♡ **Example** 4.4. Consider the data in the file `latent.dat` (also presented on the introductory handout), which are measurements of the latent heat of water using two methods.

Measurements subject to error are often assumed to be normally distributed with mean μ equal to the ‘true’ value. Normal probability plots of the sample data produce straight lines for both samples, we can assume that the distribution of measurements is approximately normal, and calculate a confidence interval for the true value μ without any concerns.

Using sample data for method A, the sample mean is $\bar{x} = 80.021$, the sample standard deviation is $s = 0.024$, and therefore, as the sample size $n = 13$, the standard error $s/\sqrt{n} = 0.007$, and $c = 2.18$ using the tables. Therefore, a 95% confidence interval for μ, the true latent heat of water is $(80.006, 80.035)$.

Using sample data for method B, the sample mean is $\bar{x} = 79.979$, the sample standard deviation is $s = 0.031$, and therefore, as the sample size $n = 8$, the standard error $s/\sqrt{n} = 0.011$. Therefore, a 95% confidence interval for μ, the true latent heat of water is $(79.953, 80.005)$.

4.4 Estimating other parameters

When estimating the mean μ of a distribution, a sample mean is an obvious estimator. However, for other parameters, a convenient estimator may not be quite so obvious. For example, how should we estimate the parameters α and β of a Weibull or EVG1 distribution?

The answer is that there exists a flexible estimation procedure, applicaable in simple or complex models, which can be proved theoretically to produce estimators with good properties. The method is called **maximum likelihood estimation**.

To discuss this method in general is betyond the scope of this course. However, we shall attempt to give a flavour of the method, by considering estimating the parameter β of the exponential distribution (See §3 for details).

Suppose we have observations $x_1, x_2, \ldots, x_n$ from an exponential distribution with parameter β. The probability density evaluated at each of these observations is

$$\frac{1}{\beta}e^{-\frac{x_1}{\beta}},\ \frac{1}{\beta}e^{-\frac{x_2}{\beta}},\ \ldots,\ \frac{1}{\beta}e^{-\frac{x_n}{\beta}}.$$

As the observations are assumed to be independent we write their **joint** probability density as

$$\frac{1}{\beta}e^{-\frac{x_1}{\beta}} \times \frac{1}{\beta}e^{-\frac{x_2}{\beta}} \times \ldots \times \frac{1}{\beta}e^{-\frac{x_n}{\beta}} = \frac{1}{\beta^n}e^{-\frac{x_1+\ldots+x_n}{\beta}} = \frac{1}{\beta^n}e^{-\frac{1}{\beta}\sum_{i=1}^n x_i}.$$

This function reflects how likely the observed data are in terms of how great the probability density is at each of the observed data values. The method of maximum likelihood estimates β by the value which makes the observed data more likely than any other value of β would. In other words, we maximise

$$\frac{1}{\beta^n}e^{-\frac{1}{\beta}\sum_{i=1}^n x_i}$$

as a function of β.

Differentiating this expression with respect to β, we get

$$\begin{aligned}\tfrac{d}{d\beta}\left[\beta^{-n}e^{-\beta^{-1}\sum_{i=1}^n x_i}\right] &= -n\beta^{-n-1}e^{-\beta^{-1}\sum_{i=1}^n x_i} + \beta^{-n}\beta^{-2}\sum_{i=1}^n x_i e^{-\beta^{-1}\sum_{i=1}^n x_i} \\ &= \left[-n\beta + \sum_{i=1}^n x_i\right]\beta^{-n-2}e^{-\beta^{-1}\sum_{i=1}^n x_i} \\ \\ &= 0 \qquad \text{if} \quad \beta = \tfrac{1}{n}\sum_{i=1}^n x_i = \overline{x}.\end{aligned}$$

Therefore, setting β equal to the sample mean $\overline{x}$ makes the observed data $x_1, \ldots, x_n$ more likely than any other value of β, so the sample mean is the maximum likelihood estimate for β. [Recall that β is the mean of an exponential distribution, so estimating it by a sample mean seems intuitively sensible. Similarly, the maximum likelihood estimate for the mean μ of a normal distribution os also the sample mean, although the maximum likelihood estimate for the standard deviation σ is not the sample standard deviation s. It is $\sqrt{\frac{n-1}{n}s}$]. In practice, maximum likelihood estimates can be calculated for any parameter.

MINITAB provides maximum likelihood estimates for model parameters using **Graph→Probability plot**, along with the probability plot to check whether the model is appropriate.

Ideally, we would also like to be able to get standard errors or (even better) confidence intervals for these parameters but **MINITAB** does not provide these in general. However, what **MINITAB** does provide are estimates of the distribution function $F(x) = P(X \leq x)$ based on the parameter estimates.

In particular, for a number of values of p (and more can be specified using `Options`), estimates of the value of x for which $P(X \leq x) = p$ are given. Furthermore, the uncertainty in these estimates is represented by confidence intervals.

♡ **Example** 4.5. Consider the data in file `stress1.dat` relating to the stresses resulting from wave action on the joints of an off-shore oil-drilling platform. In §3 we used a Weibull distribution as a model for this variable, and estimated the parameters as $\alpha = 0.98$ and $\beta = 21.8$. Suppose, for design purposes we want to estimate the value of stress which is exceeded with probability 0.01. Then the relevant MINITAB output is as follows.

```
Percentile Estimates

                           95% CI       95% CI
                          Approximate  Approximate
Percent  Percentile  Lower Limit  Upper Limit

 1          0.203      0.0374        1.104
 2          0.413      0.0950        1.796
 ...
98         87.508     52.2537      146.548
99        103.300     59.5150      179.299
```

Hence, although the stress value exceeded with probability 0.01 is estimated to be 103.3, there is considerable uncertainty, as the 95% confidence interval $(59.5, 179.3)$ is very wide.

4.5 Hypothesis Tests for the Mean of a Population

Sometimes, sample data are collected with the purpose of examining a conjecture or **hypothesis** concerning a distribution. For example, data may be collected to ensure that certain standards are being satisfied, or a change may have been made to a process and data is collected on an output of that process to see if its distribution has changed from the (known) previous distribution.

We will focus on hypotheses which concern the (unknown) distribution mean μ, although in principle a hypothesis may concern any property of the distribution which might be of interest (for example, any parameter of the distribution).

Again, we suppose that $\overline{x}$ is the mean of a sample of n observations $x_1, \ldots, x_n$ from a distribution with mean μ and standard deviation σ. Let the hypothesised value of μ be denoted by μ_0.

A confidence interval gives a range of plausible values of μ, based on the observed data, so a sensible procedure would seem to be to reject the hypothesis that $\mu = \mu_0$ if the value of μ_0 does not lie inside our confidence interval.

For example, if we have a 95% confidence interval, and the mean of the distribution is indeed equal to μ_0, then for 95% of samples, the confidence interval will include μ_0. If it does not then, either we have been unlucky, and observed one of the 5% of samples with erroneous confidence intervals, or the mean of the distribution is not, in fact, equal to μ_0.

Therefore, if the hypothesised value, μ_0 does not lie in the 95% confidence interval, we use this fact as evidence against the hypothesis that $\mu = \mu_0$ and reject the hypothesis **at the 5% level of significance**. Another way of saying this is that the evidence against the hypothesis $\mu = \mu_0$ is **statistically significant** at the 5% level.

More generally, the significance level for the test is 1− the confidence level of the associated interval. Smaller significance levels correspond to wider confidence intervals, so require greater evidence in order to reject.

Recall that, for large samples, or distributions which are close to normal, we calculate confidence intervals based on the fact that

$$t = \frac{\overline{x} - \mu}{s/\sqrt{n}}$$

is an observation from a t_{n-1} distribution. Hence, if our hypothesis that $\mu = \mu_0$ is true, then

$$T = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}$$

is an observation from a t_{n-1} distribution.

The hypothesised mean μ_0 will fall inside the confidence interval if

$$\begin{aligned} & \overline{x} - c\frac{s}{\sqrt{n}} \leq \mu_0 \leq \overline{x} + c\frac{s}{\sqrt{n}} \\ \Rightarrow \quad & -c \leq \frac{\overline{x} - \mu_0}{s/\sqrt{n}} \leq c \\ \Rightarrow \quad & -c \leq T \leq c \end{aligned}$$

where c is the relevant value calculated using the t_{n-1} distribution

If $|T| > c$, then μ_0 is outside the confidence interval and the hypothesis is rejected.

Therefore, a hypothesis test involves calculating the **test statistic** T and seeing if it falls in the rejection region $|T| > c$ evaluated using the t_{n-1} distribution together with the significance (confidence) level for the test. This is intuitively sensible. The test will reject when the sample mean $\overline{x}$ and the hypothesised distribution mean μ_0 are far apart.

Sometimes we calculate a p-value for a hypothesis test. A p-value is the highest significance level at which the hypothesis would **not** be rejected. Therefore the p-value is the significance value of the test for which T lies right on the edge of the rejection region *i.e.* $|T| = c$. Hence the p-value is the probability that an observation from a t_{n-1} distribution is greater than $|T|$ or less than $-|T|$.

Recall that smaller significance levels require greater evidence, so if the p-value is small, the data are providing strong evidence against the hypothesis, because the hypothesis is rejected, even at small significance levels.

Usually, we reject the hypothesis if the p-value $p < 0.05$. In other words, we tend to use a 5% significance level. Other values of significance which are commonly used are 1% and 0.1%. These imply even stronger evidence against H_0.

If the hypothesis is not rejected, then that is exactly what has happened – we have not rejected it. This does not mean that we have accepted it. Try to avoid using the word 'accept' when talking about statistical hypotheses. The reason that we have not rejected the hypothesis may be simply that we have not observed a sufficiently large sample for the evidence against it to be statistically significant.

Another issue to be aware of is the difference between statistical and practical significance. We reject a hypothesised mean μ_0 because the data provide strong evidence that the true distribution mean μ is not equal to μ_0. However, this does not necessarily mean that there is a large discrepancy between μ and μ_0. Indeed, in practical terms it is possible for the discrepancy between μ and μ_0 to be of a magnitude which is relatively unimportant in the application concerned. What constitutes a practically significant discrepancy depends on the application and is not a statistical issue. A confidence interval is a particularly useful summary, as it enables you to assess both statistical significance (is the hypothesised value in the interval) and practical significance (how far is the interval away from the hypothesised value).

♡ **Example** 4.6. The file `concrete.dat` contains the compression strength (Nmm^{-2}) of 180 concrete cubes. Suppose that the cubes are required to be manufactured with a mean compression strength of 62 Nmm^{-2}. Test the hypothesis that the process is manufacturing cubes to the required standard.

Here we are required to test the hypothesis that $\mu = 62$ (cubes are of the required standard).

Recall that a 95% confidence interval for μ is $(60.515, 61.681)$. As this interval does not include 62, we reject the hypothesis that $\mu = 62$ at the 5% level of significance. Furthermore, as a 99% confidence interval for μ is $(60.329, 61.867)$, and this we also reject the hypothesis that $\mu = 62$ at the 1% level of significance.

The test statistic for this test is

$$T = \frac{\overline{x} - 62}{s/\sqrt{n}} = \frac{61.098 - 62}{0.295} = -3.05$$

which gives a p-value of 0.003. This is a very small p-value, indicating that these data provide extremely strong evidence against the hypothesis that $\mu = 62$.

On the other hand, if the required standard for the mean of the distribution is 61.5 Nmm^{-2}, this value falls inside the 90% confidence interval so, even at the 10% level of significance, there is no evidence that μ is not equal to 61.5. The test statistic for this test is

$$T = \frac{\overline{x} - 61.5}{s/\sqrt{n}} = \frac{61.098 - 61.5}{0.295} = -1.36$$

which gives a p-value of 0.175. This is quite a moderate p-value, indicating that these data provide no significant evidence against the hypothesis that $\mu = 61.5$.

In MINITAB

Stat→Basic Statistics→1-sample t

4.6 Comparing Two Distributions

Often, the most interesting hypotheses arise when we are comparing two (or more) distributions. Usually, we are interested in whether the observations of one distribution are larger than those of the other.

To investigate this, we again focus on the distribution means, and use samples from each of the distributions concerned to test a hypothesis concerning the distribution means.

Suppose that we observe a sample of n observations $x_1, \ldots, x_n$, from the distribution of variable X and a sample of m observations $y_1, \ldots, y_m$, from the distribution of variable Y.

We assume that the distribution of X has mean μ_x and standard deviation σ_x, and that the sample $x_1, \ldots, x_n$ has sample mean $\overline{x}$ and sample standard deviation s_x. Similarly, the distribution of Y has mean μ_y and standard deviation σ_y, and the sample $y_1, \ldots, y_m$ has sample mean $\overline{y}$ and sample standard deviation s_y.

Two cases arise depending on whether we can assume that $\sigma_x = \sigma_y$.

4.6.1 Case 1: The Two Sample t Test under equal variance assumption

Suppose we can assume that $\sigma_x = \sigma_y$. We can check this assumption by considering a normal-probability plot. We need to see if the slopes of the two probability plots are roughly same or not. (Recall that the slopes are standard deviations in a normal probability plot.)

We calculate the following to test if the two means are equal,

$$T = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{1}{m} + \frac{1}{n}}\sqrt{\frac{(n-1)s_x^2 + (m-1)s_y^2}{m+n-2}}},$$

follows the $t-$ distribution with $m + n - 2$ degrees of freedom.

In MINITAB

Stat→Basic Statistics→ 2-sample t

and check the box for equal variances. For the **latent.dat** we get $T = 3.47$ on 19 degrees of freedom. The 95% confidence interval for $\mu_x - \mu_y$ is (0.0167, 0.0673). Since this does not include zero we reject the hypothesis that the mean are equal at 5% level of significance.

4.6.2 Case 2: An approximate two Sample t-test when variances are unequal

If we cannot assume that $\sigma_x = \sigma_y$ we do not have an exact general solution. However, **provided that** either

(a) the sample sizes n and m are large, or

(b) the distributions of X and Y are approximately normal, (this may need to be checked using normal probability plots)

it can be shown that

$$T = \frac{\overline{x} - \overline{y} - (\mu_x - \mu_y)}{\sqrt{\frac{s_x^2}{n} + \frac{s_y^2}{m}}}$$

is an observation from a distribution which has (approximately) a t distribution with k degrees of freedom where

$$k = \frac{\left(\frac{s_x^2}{n} + \frac{s_y^2}{m}\right)^2}{\frac{(s_x^2/n)^2}{n-1} + \frac{(s_y^2/m)^2}{m-1}}.$$

Hence, using the t_k distribution, we can find c such that

$$P\left(-c \le \frac{\overline{x} - \overline{y} - (\mu_x - \mu_y)}{\sqrt{\frac{s_x^2}{n} + \frac{s_y^2}{m}}} \le c\right) \quad = \quad 0.95.$$

$$\Rightarrow \qquad P\left(\overline{x}-\overline{y}-c\sqrt{\frac{s_x^2}{n}+\frac{s_y^2}{m}} \le \mu_x-\mu_y \le \overline{x}-\overline{y}+c\sqrt{\frac{s_x^2}{n}+\frac{s_y^2}{m}}\right) \quad = \quad 0.95.$$

so the endpoints of a 95% confidence interval for the difference between the means, $\mu_x - \mu_y$ are

$$\overline{x}-\overline{y} \pm c\sqrt{\frac{s_x^2}{n}+\frac{s_y^2}{m}}$$

where c is evaluated using the t_k distribution, with k calculated as above.

Most commonly, we are interested in whether the distributions of X and Y are the same, or whether the data provide significant evidence that they differ. Hence a common hypothesis of interest is $\mu_x = \mu_y$, or equivalently $\mu_x - \mu_y = 0$. To test this hypothesis at the 5% level of significance, all that is required is to check whether or not zero falls in the confidence interval for $\mu_x - \mu_y$, or equivalently whether the test statistic T given above falls in the rejection region $|T| > c$. Again, a p-value for the test gives the largest significance level at which the hypothesis is not rejected, and small p-values indicate strong evidence against the hypothesis.

♡ **Example** 4.7. Consider the data in the file `latent.dat`, which are measurements of the latent heat of water using two methods. As measurements subject to error are often assumed to be normally distributed, and normal probability plots of the sample data produce straight lines for both samples, we assume that the population of measurements are approximately normal, for both methods.

A question of interest is whether there is a systematic difference between the measuring methods. We might conclude that there is a systematic difference if μ_x, the mean of all possible measurements made using method A was different from μ_y, the mean of all possible measurements made using method B. To determine this we test the hypothesis $\mu_x = \mu_y$.

Here, $\overline{x} = 80.021$, $s_x = 0.024$, $n = 13$, $\overline{y} = 79.979$, $s_y = 0.031$, $m = 8$, so $T = 3.25$ and $k = 12$ (rounded), and a 95% confidence interval for $\mu_x - \mu_y$ is $(0.0138, 0.0702)$, which does not include zero. Therefore, we reject the hypothesis that the means are equal at the 5% significance level.

The p-value for this test is 0.007 so there is highly significant evidence of a systematic difference between the measurement methods.

Recall that the 95% confidence interval for $\mu_x - \mu_y$ in under the equal variance assumption is (0.0167, 0.0673) and this is wider than the interval (0.0138, 0.0702). This is expected since we get tighter inferences under more assumptions (the equality of variances).

In real life problems the choice between the two tests depends on which assumptions we can justify, i.e. can we assume that the variances are equal? Are the sample sizes large?

4.6.3 The paired t Test

The two sample t test is carried out under the assumption that the samples from the two distributions are **independent of one another**. In some situations this assumption is clearly violated. For example, consider the data in the file `labs.dat`.

♡ **Example** 4.8. In the USA, municipal wastewater treatment plants are required by law to monitor their discharges into rivers and streams on a regular basis. Concern about the reliability of data from one of these self-monitoring programs led to a study in which 11 volumes of effluent were divided and set to two laboratories for testing. One half of each volume was sent to the Wisconsin State Laboratory of Hygiene and one half was sent to a private commercial laboratory routinely used in the monitoring program. The data in the file `labs.dat` are measurements of biochemical oxygen demand (BOD – `c1`; commercial laboratory, `c3`; state laboratory) and suspended solids (SS – `c2`; commercial laboratory, `c4`; state laboratory) for each of the 11 volumes.

To investigate whether there is a systematic difference between the state and commercial laboratories we can test whether μ_x, the mean of the distribution of X, the BOD as measured by the commercial laboratory differs from μ_y, the mean of the distribution of Y, the BOD as measured by the state laboratory.

However, we have **not** observed **independent** samples from these two distributions, as the 11 volumes analysed by each of the two laboratories were not obtained as 22 independent volumes, but by splitting 11 larger volumes.

The observations of one sample are **paired** with the observations of the other sample. Therefore, we ought to expect x_1, the first measurement from the commercial laboratory, to be more closely related to y_1, the first measurement from the state laboratory, than to any other measurement, as these measurements were made on (essentially) the same volume of effluent.

In general, assume that we have n observations from each of the two populations, and that these observations have been collected in such a way that they are clearly paired, so that the samples are not independent. Denote the pairs of observations $(x_1, y_1), \ldots, (x_n, y_n)$.

In this situation, we consider the variable $D = X - Y$, the **difference between a pair of observations**. This distribution has mean μ_d and standard deviation σ_d. We rewrite our hypothesis of interest $\mu_x = \mu_y$, as $\mu_d = 0$, a hypothesis concerning the distribution of D.

A sample of differences $d_1, \ldots d_n$ from the distribution of D is calculated using the paired observations.

$$d_1 = x_1 - y_1 \qquad d_2 = x_2 - y_2 \quad \ldots \quad d_n = x_n - y_n.$$

We can test a hypothesis about the mean μ_d, of the distribution of D using a sample $d_1, \ldots d_n$, from that population, using the methods of §4.5.

♡ **Example** 4.9. Consider the data in `labs.dat`. Is there a systematic difference between the laboratories in the way in which they measure biochemical oxygen demand?

To examine this, we test the hypothesis that μ_d, the mean of the distribution of D, the difference between BOD measurements of a volume of water split and analysed by the two laboratories is zero ($\mu_d = 0$; no difference between laboratories).

We have a sample

$$-19 \quad -22 \quad -18 \quad -27 \quad -4 \quad -10 \quad -14 \quad 17 \quad 9 \quad 4 \quad -19$$

of differences (values of D). Here $\overline{d} = -9.36$, $s/\sqrt{n} = 4.26$, and $T = -2.20$. A 95% confidence interval for μ_d is $(-18.85, 0.12)$. This includes the hypothesised value, $\mu_d = 0$ so we do not reject the hypothesis at the 5% significance level.

The p-value is 5.2% so the evidence of a difference between the laboratories is very close to being significant, but not quite. Is there a systematic difference between the laboratories in the way in which they measure suspended solids?

To examine this, we test the hypothesis that μ_d, the mean of the distribution of D, the difference between SS measurements of a volume of water split and analysed by the two laboratories is zero ($\mu_d = 0$; no difference between laboratories).

We have a sample

$$12 \quad 10 \quad 42 \quad 15 \quad -1 \quad 11 \quad -4 \quad 60 \quad -2 \quad 10 \quad -7$$

from the population of differences. Here $\overline{d} = 13.27$, $s/\sqrt{n} = 6.17$, and $T = 2.15$. A 95% confidence interval for μ_d is $(-0.47, 27.02)$. This includes the hypothesised value, $\mu_d = 0$ so we do not reject the hypothesis at the 5% significance level.

The p-value is 5.7% so, again, the evidence of a difference between the laboratories is close to being significant, but not quite. In this example, it seems clear that collecting more data might lead one to conclude that there was a difference but, with the data available, we have not observed significant evidence of a difference.

In MINITAB

Stat→Basic Statistics→ Paired t