A total of 80 city council buses were randomly selected. The arrival times of each bus at their last stop were compared to the published bus timetable to determine if they were late. Sixteen buses were observed to be late. Construct a 90% confidence interval for the true proportion for council buses that arrived late at their last stop.

let p be the proportion of buses that are late.

general form of CI estimate ± (critical value) × s.e. (estimate)

estimate $\hat{p} = \frac{16}{80} = 0.2$ $\quad$ $s.e.(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.2 \times 0.8}{80}} = 0.04472\ldots$

$0.90 = 1 - \alpha \Rightarrow \alpha = 0.1$ $\quad$ $z_{0.95} = 1.6449$

$$0.2 \pm 1.6449 \times 0.04472 \Rightarrow 0.2 \pm 0.0736$$

we are 90% confident the true proportion of late buses is in $(0.126, 0.274)$

Having received many complaints from upset passengers that buses were always late, the council decides to implement an electronic ticketing system. Three months after introducing the electronic ticketing system, another random sample of 80 city council buses was selected and 10 were observed to be late at their last stop. Construct a 99% confidence interval for the change in the proportion of buses that arrive late at their last stop.

Let p_{new} be the proportion of late buses with new system

p_{old} " " " with old system

$\hat{p}_{old} = \frac{16}{80} = 0.2$ $\quad$ $\hat{p}_{new} = \frac{10}{80} = 0.125$ $\quad$ $\hat{p}_{new} - \hat{p}_{old} = -0.075$

General form for CI estimate ± (critical value) × s.e. (estimate)

$$s.e.(\hat{p}_{new} - \hat{p}_{old}) = \sqrt{\frac{\hat{p}_{new}(1-\hat{p}_{new})}{n_{new}} + \frac{\hat{p}_{old}(1-\hat{p}_{old})}{n_{old}}} = \sqrt{\frac{0.125 \times 0.875}{80} + \frac{0.2 \times 0.8}{80}}$$

$$= 0.05803$$

99% CI $\quad$ $0.99 = 1 - \alpha$, $\alpha = 0.01$ $\quad$ $z_{1-\alpha/2} = 2.5758$

$$-0.075 \pm 2.5758 \times 0.05803$$

$$-0.075 \pm 0.1495$$

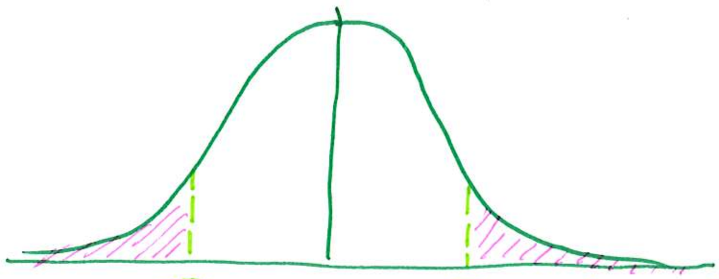
We are 99% that the true difference in the proportion of late buses (new − old) is between -0.224 and 0.075

- Two sided alternative ($H_1 : \theta \neq \theta_0$) The p-value is given by

$H_0 : \theta = \theta_0$

$$2 \min \left[\mathbb{P}(T(\mathbf{X}) > T(\mathbf{x})), \mathbb{P}(T(\mathbf{X}) < T(\mathbf{x}))\right],$$

where the probability is evaluated under the null hypothesis.

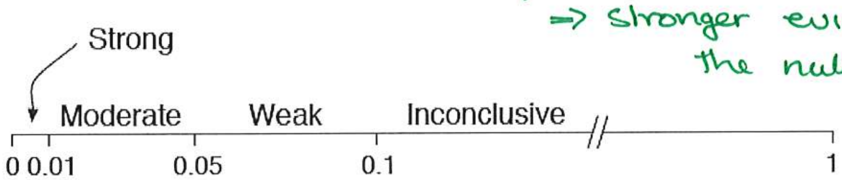


Like the test statistic, the p-value is a function data and so it also has a distribution. Under the null hypothesis

$$p - \text{value} \sim \mathsf{Uniform}(0, 1).$$

The strength of evidence against the null hypothesis provided by the p-value is summarised in the figure below.

Smaller p-values $\Rightarrow$ stronger evidence against the null hypothesis.



We must decide how small the p-value must be before we reject the null hypothesis. This cut-off point is called the **significance level** and is often denoted by α. The significance level determines the probability that we reject the null hypothesis when it is in fact true.

Question: Suppose you were to toss a coin that you believed was fair several times. How many consecutive heads would need to appear before you begin to doubt that it is really a fair coin?

It is common to use significance levels of 5% or 1%, though sometimes much smaller significance levels are needed.

Example: Assume that the measurements from Cavendish's apparatus are a realisation of a simple random sample from $\mathsf{Normal}(\mu, \sigma^2)$. We wish to test whether or not Cavendish's apparatus gave unbiased measurements of the density of the earth, that is we are testing

$$H_0 : \mu = 5.517 g/cm^3 \quad \text{against} \quad H_1 : \mu \neq 5.517 g/cm^3.$$

Cavendish made 23 measurements of the earth's density, with $\bar{x} = 5.4835 g/cm^3$ and $s = 0.1904 g/cm^3$. The test statistic is

$$T(\mathbf{x}) = \frac{\bar{x} - 5.517}{s/\sqrt{n}} = \frac{5.4835 - 5.517}{0.1904/\sqrt{23}} = -0.8438$$

$$\text{test statistic} = \frac{\text{estimate} - \text{hypothesised}}{\text{s.e.(estimate)}}$$

MATLAB tcdf(-0.8438, 22)

R pt(-0.8438, 22)

Under H_0, $T(\mathbf{X})$ has a t_{n-1}-distribution. So in this case we need to compare our test statistic with the t_{22}-distribution to get the p-value.

$$
\begin{aligned}
&2\min\left[\mathbb{P}(T_{22} > -0.8438), \mathbb{P}(T_{22} < -0.8438)\right] \\
&= 2\mathbb{P}(T_{22} > 0.8438) \qquad \text{[as t-distribution is symmetric about zero]} \\
&= 2 \times (0.1, 0.25) \qquad \text{[from tables]} \\
&= (0.2, 0.5)
\end{aligned}
$$

$\mathbb{P}(T_{22} < -0.8438) = 0.2039$

p-value $= 0.4079$.

The p-value is between 0.2 and 0.5. This is inconclusive evidence against H_0. In other words, there is no evidence of bias in Cavendish's apparatus. At the 5% significance level, we retain the null hypothesis.

Connection to confidence intervals

In the previous chapter, we saw how to construct a confidence interval for the mean. Lets now construct a confidence interval for the mean density reading from Cavendish's apparatus.

95% CI for μ: estimate $\pm$ (critical value) $\times$ s.e.(estimate)

$0.95 = 1 - \alpha$, $\alpha = 0.05$ $\quad t_{0.975;22} = 2.0739$

$$\bar{x} \pm t_{0.975;22} \, s/\sqrt{23}$$

$$5.4835 \pm 2.0739 \times \frac{0.1904}{\sqrt{23}}$$

$$5.4835 \pm 0.0823 \ g/cm^3$$

$$(5.4012, 5.5658)$$

We can be 95% confident that the mean value of the density measurements made by his apparatus was between 5.401 and $5.566 g/cm^3$. Note that this interval contains the hypothesised true value of $5.517 g/cm^3$. Is it just a coincidence that 5.517 was accepted in our hypothesis test?

There is a nice duality between confidence intervals and hypothesis testing. In fact, confidence intervals can be defined as the "inverse" of hypothesis tests:

An alternative definition of a $(1-\alpha)100\%$ confidence interval for a parameter θ is

$$\{\theta \mid \theta \text{ is accepted at } \alpha \text{ significance level (two-sided test)}\}.$$

This is the set of all hypothesised parameter values that would be accepted in a two-sided hypothesis test at significance level α.

Type I and II errors

Whenever we make decisions, we run the risk of making errors. If we reject the null hypothesis when it is in fact true, we have made a **Type I error**. The probability of making a Type I error is precisely the significance level α that we choose for making decisions. For example, if we think a p-value less than $0.05 = 5\%$ is too rare to accept H_0, then we will accidentally reject H_0 precisely 5% of the time.

On the other hand, if we accept H_0 when it is false, then we make a **Type II error**. Related to the notion of Type II errors is the **power** of a statistical test. The power of a statistical test is the probability of detecting an effect when there is indeed an effect. If β is the probability of making a Type II error, then the power is given by $1 - \beta$.

We can think of these errors in terms of a court case:

- A Type I error is accidentally finding someone guilty when they are in fact innocent.
- A Type II error is accidentally finding someone innocent when they are in fact guilty.
- Power is the probability of finding a guilty person guilty.

To summarise, we the following probabilities for all four scenarios:

	Decision	
	Retain H_0	Reject H_0
H_0 is true	Correct $(1 - \alpha)$	Type I Error (α)
H_0 is false	Type II Error (β)	Correct $(1 - \beta)$

Comparing two means

Example: A real estate agency wants to compare the appraised values of studio apartments in Toowong and Dutton Park. The following results were obtained from random samples:

	Toowong	Dutton Park
Sample Size	25	30
Sample Mean	\$ 226 716	\$ 206 634
Sample Standard Deviation	\$ 32 338	\$ 13 464

Do the two regions have the same (population) mean value for studio apartments?

To address problems like this we follow the same argument that we used to construct the test of a single mean. Suppose we have a simple random sample $X_1, \dots, X_m$ from a $\mathsf{Normal}(\mu_X, \sigma^2)$ distribution and another simple random sample from $Y_1, \dots, Y_n$ from a $\mathsf{Normal}(\mu_Y, \sigma^2)$. We want to test the null hypothesis $H_0 : \mu_X - \mu_Y = d$, for some given value d against an alternative hypothesis H_1. The alternative hypothesis is usually one of the following forms:

typically $d=0$.

- One sided alternative: $H_1 : \ \mu_X - \mu_Y > d$.
- One sided alternative: $H_1 : \ \mu_X - \mu_Y < d$.
- Two sided alternative: $H_1 : \ \mu_X - \mu_Y \neq d$.

Example: For the real estate example, we formulate the null and alternative hypothesis as follows: Let μ_T be the mean appraised value of a studio apartment in Toowong and let μ_D be the mean appraised value of a studio apartment in Dutton Park.

$\mu_T - \mu_D = 0$ $H_0 : \mu_T = \mu_D$	$\mu_T - \mu_D \neq 0.$ $H_1 : \mu_T \neq \mu_D$

The test statistic for this hypothesis test is

d under H_0

$$T(\mathbf{X}, \mathbf{Y}) = \frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{S_p\sqrt{\frac{1}{m} + \frac{1}{n}}},$$

where S_p^2 is the sample pooled variance estimator

$$\begin{aligned} S_p^2 &= \frac{(m-1)S_X^2 + (n-1)S_Y^2}{n+m-2} \\ &= \frac{\sum_{i=1}^m (X_i - \bar{X})^2 + \sum_{j=1}^n (Y_j - \bar{Y})^2}{n+m-2}. \end{aligned}$$

As we saw in the previous chapter on confidence intervals, under H_0,

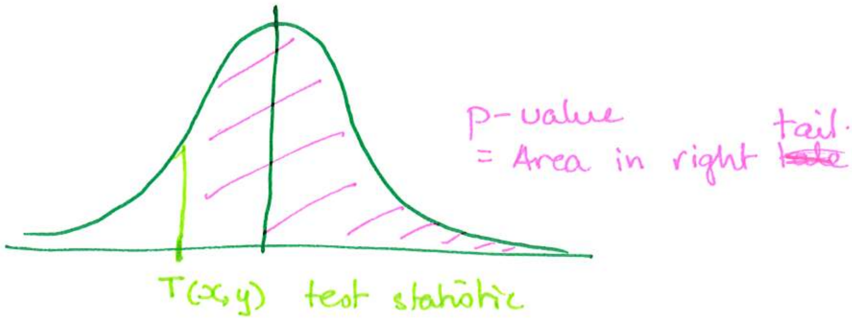
$$T(\mathbf{X}, \mathbf{Y}) \sim t_{n+m-2}.$$

When our test statistic computed from the sample data $T(\mathbf{x}, \mathbf{y})$ is 'large' in an appropriate sense, this will indicate evidence against the null hypothesis. The p-value is given by:

- One sided alternative ($H_1 : \mu_X - \mu_Y > d$) The p-value is given by

$$\mathbb{P}(T(\mathbf{X}, \mathbf{Y}) > T(\mathbf{x}, \mathbf{y})),$$

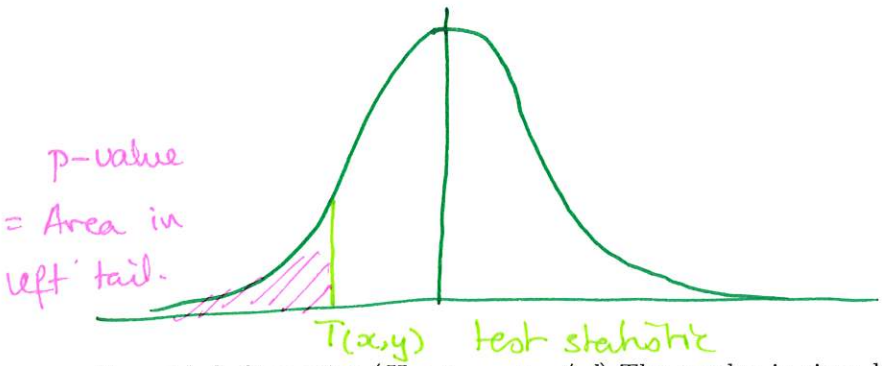
where the probability is evaluated under the null hypothesis.



- One sided alternative ($H_1 : \mu_X - \mu_Y < d$) The p-value is given by

$$\mathbb{P}(T(\mathbf{X}, \mathbf{Y}) < T(\mathbf{x}, \mathbf{y})),$$

where the probability is evaluated under the null hypothesis.



- Two sided alternative ($H_1 : \mu_X - \mu_Y \neq d$) The p-value is given by

$$2 \min \left[\mathbb{P}(T(\mathbf{X}, \mathbf{Y}) > T(\mathbf{x}, \mathbf{y})), \mathbb{P}(T(\mathbf{X}, \mathbf{Y}) < T(\mathbf{x}, \mathbf{y}))\right],$$

where the probability is evaluated under the null hypothesis.